UNIVERSITY OF THE WITWATERSRAND

JOHANNESBURG

SCHOOL OF PHYSICS

PHYS1001/1006

LECTURE NOTES
SEMESTER 2

2018

Student		Class	

SECTION 3: INTRODUCTION TO WAVES

3.1 Wave Parameters
3.2 Types of Waves
3.3 Superposition of Waves

SECTION 4: ELECTRICITY AND MAGNETISM

4.1 Electrostatics
4.2 Current Electricity
4.3 Electromagnetism

SECTION 5: OPTICS

5.1 Geometrical Optics (Plane Interfaces)
5.2 Geometrical Optics (Curved Interfaces)
5.3 Physical Optics

SECTION 6: MODERN PHYSICS

6.1 Quantum Physics
6.2 Atomic Physics
6.3 Nuclear Physics

PHYSICS ID (PHYS1001/6) LECTURE NOTES

3. INTRODUCTION TO WAVES

3.1. Wave parameters

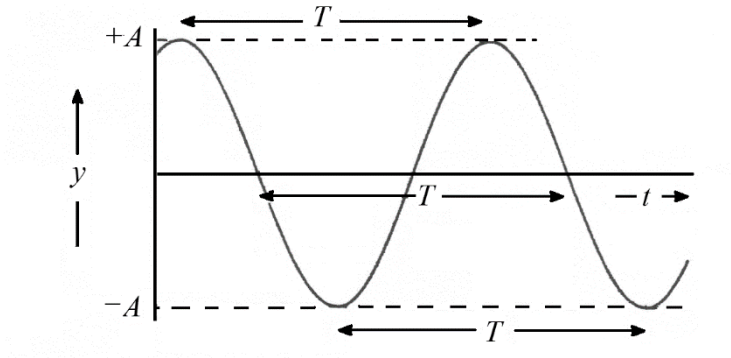
Consider a wave, assumed to be sinusoidal in shape, travelling across a water surface.

Crests high points on a wave

Troughs low points on a wave

Amplitude *A* distance from midpoint of wave to a crest or trough

If you stood at a fixed point in the water and measured the water level as a function of time, you would get a graph like that shown in the diagram below.



t is the time measured from some arbitrary starting time.

y is the height above the undisturbed level.

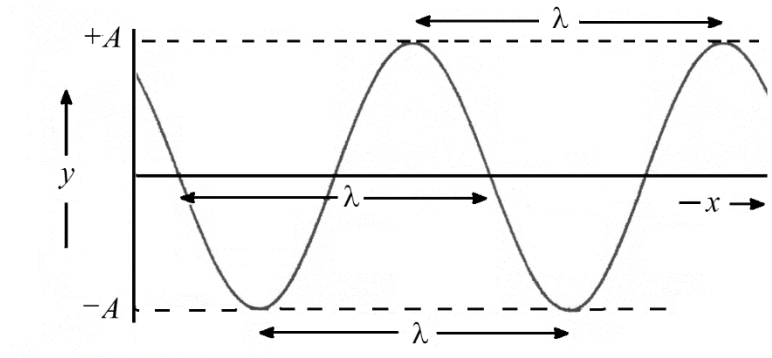
Frequency *f* number of crests (or complete cycles) that pass a given point per unit time; measured in $s^{-1} \equiv$ Hz (hertz)

Period *T* time for one complete cycle; this is the time for two successive corresponding points on the wave (for example, two successive crests) to pass a fixed point.

Frequency and period are related:

$$\boxed{f = \frac{1}{T}}$$

If you measure the water level at a fixed time as a function of position rather than time you would get a graph like that shown in the next diagram.



x is the distance from some arbitrary fixed point.

Note that the amplitude is the same as in the previous diagram.

Wavelength *λ* distance between any two successive identical points on the wave

Wave velocity *v* velocity at which the wave crests appear to move

Since the wave travels a distance λ in time T, the wave velocity is

$$v = \frac{\lambda}{T}$$

or

$$\boxed{v = f\lambda}$$

Some representative values are given in the table.

	v (m/s)	F	T	λ
FM radio	3×10^8	100 MHz	10 ns	3 m
TV	3×10^8	250 MHz	4 ns	1,2 m
Blue-green light	3×10^8	6×10^{14} Hz	$1{,}7 \times 10^{-15}$ s	500 nm
X rays	3×10^8	3×10^{18} Hz	$3{,}3 \times 10^{-19}$ s	0,1 nm
Sound (in air)	340	2 kHz	0,5 ms	0,17 m

3.2. Types of waves

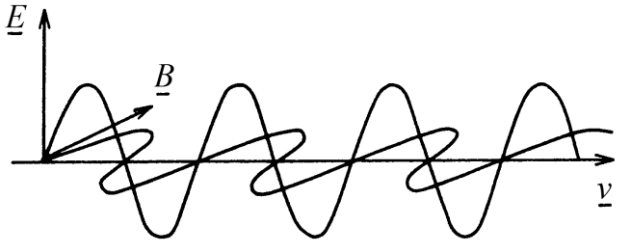
First we make a distinction between **travelling** and **non-travelling** waves. Examples of non-travelling waves are the oscillations of a pendulum and the vibrations of a mass on the end of a spring. In contrast to travelling waves, they do not carry energy from one point to another.

Travelling waves can be divided into **transverse** waves and **longitudinal** waves. We consider each in turn.

Transverse waves

Transverse waves are those in which the direction of vibration is at right angles to the direction of propagation. Examples are:

(i) Electromagnetic waves, e.g. light, heat, radio waves. These have an electric component ($\underset{\sim}{E}$) and a magnetic component ($\underset{\sim}{B}$) which oscillate at right angles to each other and to the velocity ($\underset{\sim}{v}$). Note that these waves do no require a medium through which to pass.



(ii) Water waves: the direction of propagation is horizontal, but the water molecules themselves oscillate (approximately) vertically up and down.

(iii) Waves on a rope.

Longitudinal waves

Longitudinal waves are those in which the direction of vibration is parallel to the direction of propagation. Examples are:

(i) Sound waves: molecules of the medium through which the sound is travelling vibrate back and forth parallel to the direction of propagation. Note that there is no bulk movement of the medium.

(ii) Pressure waves in blood caused by the heart's pumping action.

Earthquake waves have both transverse (S) and longitudinal (P) components. Liquids cannot transmit transverse (shear) waves through their bulk, and the absence of the transverse component of earthquake waves in regions of the earth's surface indicates the presence of a liquid core in the earth. The diameter of the liquid core can also be determined.

3.3. Superposition of waves

Two or more waves can travel simultaneously through some region, or arrive simultaneously at the same point in space. The resulting displacement is found by adding the displacements due to each wave.

The interaction of the waves is known as **interference** and it is of great importance in optics, sound and other fields. There are several cases of particular importance.

1. Two waves of the same frequency travelling the same direction.

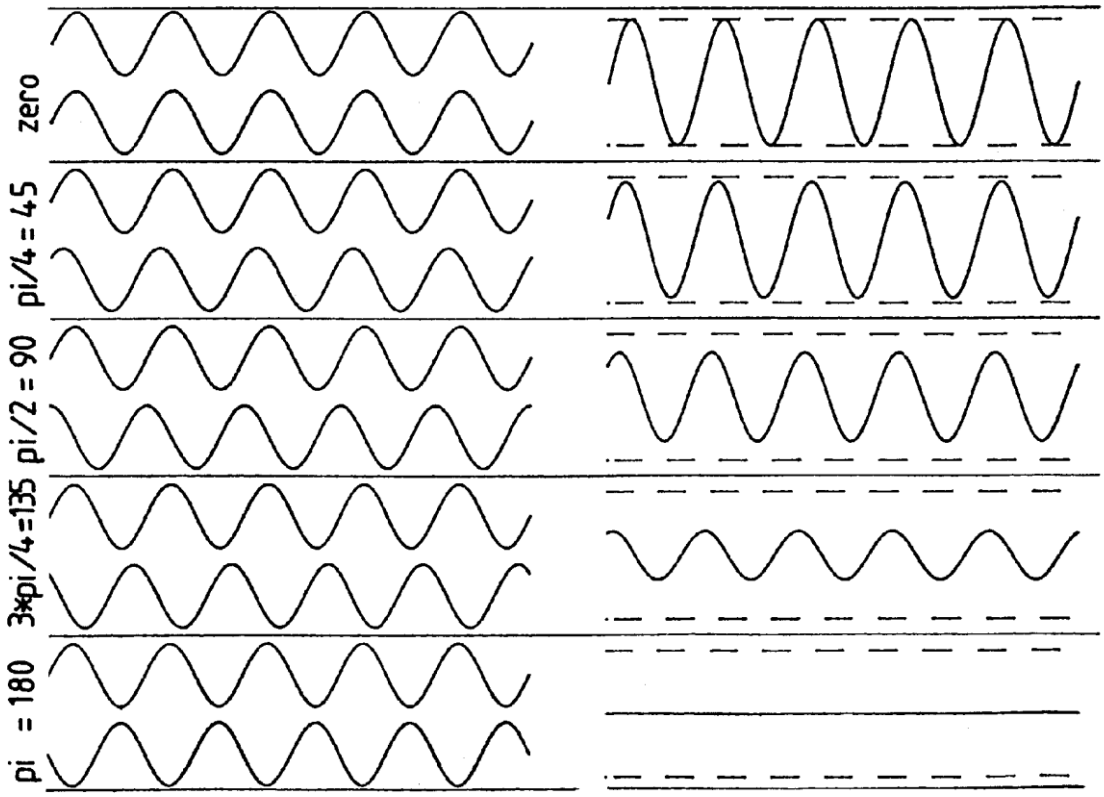
If the two waves are **exactly in phase** (uppermost illustration in the diagram below), the resultant has the same frequency (and wavelength) as the two waves. If the amplitudes of the two waves are A_1 and A_2 the resultant amplitude will be $A_1 + A_2$.

- This is called **constructive interference.**
- The same effect occurs if the two waves are a *whole number* of wavelengths out of phase.

If the two waves are **exactly out of phase** (lowest illustration), i.e. the two waves are $\lambda/2$ (or $T/2$ or π rad.) out of phase, then the resultant amplitude will be $|A_1 - A_2|$.

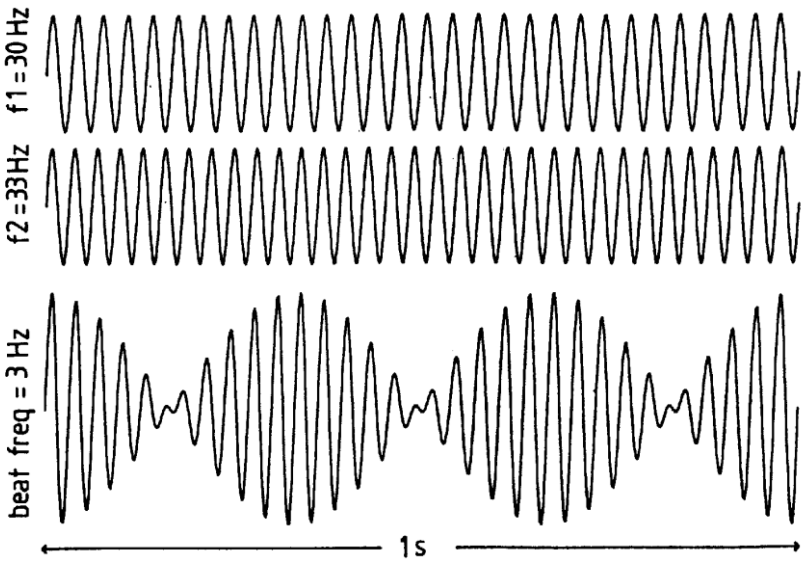
- This is called **destructive interference**.
- The same effect occurs if the waves are $3\lambda/2, 5\lambda/2, 7\lambda/2$ etc. out of phase.
- Complete destructive interference, as shown in the diagram, can occur only if the two amplitudes are equal.

If the phase difference lies between 0 and $\lambda/2$ (i.e. between 0 and π rad.), the resultant amplitude will lie between two extremes, i.e. between $A_1 + A_2$ and $|A_1 - A_2|$, see the middle diagrams.



2. Two waves of the same amplitude and nearly the same frequency travelling in the same direction.

The two waves move regularly in and out of phase, interfering constructively at one instant and destructively a short while later. The resultant amplitude varies sinusoidally. This is the phenomenon of **beats**.



The beat frequency (the frequency of the amplitude) is equal to the difference between the frequencies of the two waves, and the frequency of the resultant itself is equal to the mean of the two frequencies.

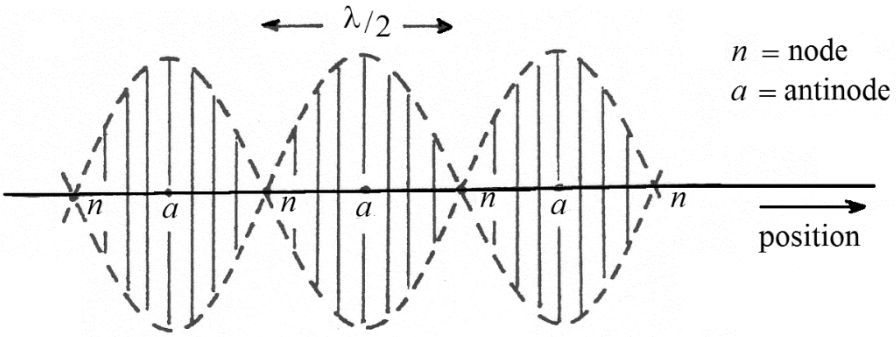
If the amplitudes of the two waves are not equal, the resultant amplitude can never be zero.

Most radio receivers make use this principle. The effect of beats is well known in music.

3. Two waves of the same frequency and amplitude travelling in opposite directions

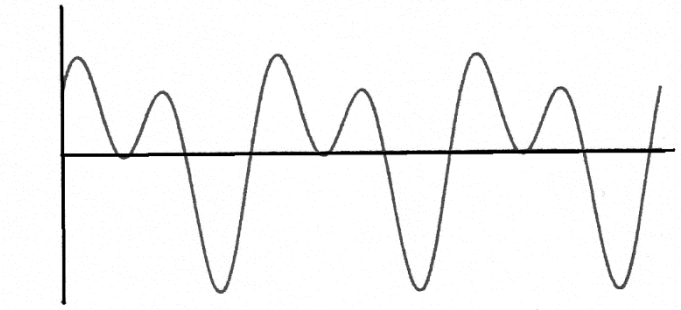
These can be set up in a tube by allowing a sound wave to be reflected from one end of the tube (the same is also true for a spring or string).

If the length of the tube or spring is an exact multiple of half wavelengths then **standing waves** are formed – these are not travelling waves. At certain points $\lambda/2$ apart the resultant amplitude is zero; these points are called **nodes**. Midway between the nodes are **antinodes** where the amplitude is a maximum.



The effects of superposition are very noticeable in speech and in music.

Combining two or more waves with different frequencies produces a periodic waveform which can be no longer described by a simple sine function, e.g.



PHYSICS ID [PHYS1001/1006] LECTURE NOTES

4. ELECTRICITY AND MAGNETISM

4.1. ELECTROSTATICS

4.1.1. Electric Charge

Experiments carried out in the eighteenth century determined that there are two kinds of electric charge; they were named by Benjamin Franklin (1706–1790) as **positive** and **negative** (the designation is entirely arbitrary). The experiments demonstrated that:

- Like charges repel each other and unlike charges attract each other.

Most materials are electrically neutral, containing equal amounts of positive and negative charge. An excess of either type of charge can, however, be created by rubbing certain materials together; for example:

- If ebonite and fur are rubbed together the ebonite becomes negatively charged and fur positively charged.
- If glass and silk are rubbed together the glass acquires a positive charge and the silk a negative charge.

In fact, almost any non-conducting material can be charged to some extent in this way.

When two materials are rubbed together, a few electrons (negatively charged particles) are transferred from one material to the other, thus making the latter negative. The material from which the electrons are removed is then left with an equal amount of positive charge. This is an illustration of a universal principle, the **conservation of electric charge**.

The SI **unit of charge, the coulomb** (C), is defined in terms of the unit of current, which will be defined later. Robert Millikan (1886–1953) discovered experimentally in 1909 that the charge on any object is always an integer multiple of the fundamental unit of charge, with magnitude $e = 1{,}60 \times 10^{-19}$ C. Other experiments determined that the electron has charge $-e$ and the proton charge $+e$. Atoms contain equal numbers of protons and electrons and are therefore electrically neutral.

Insulators and conductors

Materials are classified based on their ability to conduct electricity.

A **conductor** of electricity is a material through which charges can move readily, e.g. metals. Some electrons are not tightly bound to their atoms and may be made to move through the metal. If excess charge is placed on some small region of a conductor (by rubbing, for example, as discussed above), charges will readily redistribute themselves over the entire surface of the material.

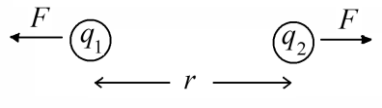
An **insulator** or **dielectric** is a material through which charges cannot move readily, e.g. rubber, glass, paper and most plastics; the atomic electrons in these materials are not free to move. When such a material is charged by rubbing, only the rubbed area becomes charged; there is no tendency for charges to move to other parts of the material.

Semiconductors are a third class of material with electrical properties somewhere between conductors and insulators. Examples are silicon and germanium. Semiconductors are used in the manufacture of most modern electronic devices.

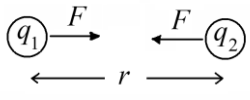
Coulomb's law

In the 1780s Charles Coulomb (1736–1806) carried out a series of experiments to investigate the electrostatic force between two static charged particles. He discovered that:

- The two particles each experience a force; these forces have the same magnitude but opposite directions (as required by Newton's third law).
- The forces act along the line joining the two charges.



If the two charges have the same sign, the forces are repulsive



If the two charges have opposite signs, the forces are attractive.

The **magnitude** of the force between two static charges q_1 and q_2 is directly proportional to the product of the magnitudes of the two charges $|q_1|$ and $|q_2|$ and inversely proportional to the square of the distance r between them:

$$F \propto \frac{q_1 q_2}{r^2} \quad \text{or} \quad F = k\frac{q_1 q_2}{r^2}$$

Note: in these and many subsequent expressions for the *magnitude* of a vector quantity, the *magnitudes of the charges* should be used. The direction of the vector should be determined separately.

In SI units the electrostatic constant has magnitude $k = 8{,}99 \times 10^9$ N.m^2.C^{-2}. It is convenient to introduce another constant through $k = 1/4\pi\varepsilon_0$ where ε_0 is the **permittivity of free space** with value $\varepsilon_0 = 8{,}85 \times 10^{-12}$ N^{-1}.m^{-2}.C^2. Then Coulomb's law becomes:

$$\boxed{F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}} \qquad \text{(Coulomb's law)}$$

Coulomb's law is strictly valid only for point charges.

- It can also be applied with little loss of accuracy to the force between any two charged objects, provided they are very small compared to their distance apart.
- We shall show later that it is also valid for the force between two charged spheres, provided the charge is distributed uniformly on each sphere.
- It should not be used to calculate the electrostatic force in any other situation.

Note the similarity between the expressions for the electric force between two charges and the gravitational force between two masses (Newton's law of gravitation); the latter force is proportional to the two masses and inversely proportional to the square of their distance apart. This leads to many analogies between electrostatics and gravitation.

4.1.2. Electric Fields

According to Coulomb's law, a charge Δq placed in the neighbourhood of another charge q experiences an electrostatic force. This is described by saying that an **electric field** exists in the region of space around the charge q, and that this field exerts an electric force on any charge Δq placed in the field (cf. the force exerted on a mass placed in a gravitational field). This concept was introduced by Michael Faraday (1791–1867).

The **strength** of the electric field at any point in the field is defined by

$$\boxed{\underset{\sim}{E} = \frac{\underset{\sim}{F}}{\Delta q}} \qquad \text{(Electric field defined)}$$

where $\underset{\sim}{F}$ is the force exerted on a small test charge Δq when placed at that point.

- It follows from this definition that the **direction** of the electric field is that of the force exerted on a **positive** charge placed in the field.
- A field exists around a charge q even if no test charge is placed in the field to detect or measure it.
- The magnitude of the field depends only on the charge distribution that creates it; it does not depend on the charge Δq used to measure it (which will always cancel out in any expression for the field – see the example below).
- The SI unit of E is N.C^{-1} (or more usually volts per metre, V.m^{-1}, see later).

Field due to a point charge

A small charge Δq placed a distance r from a point charge q experiences a force whose magnitude is given by Coulomb's law:

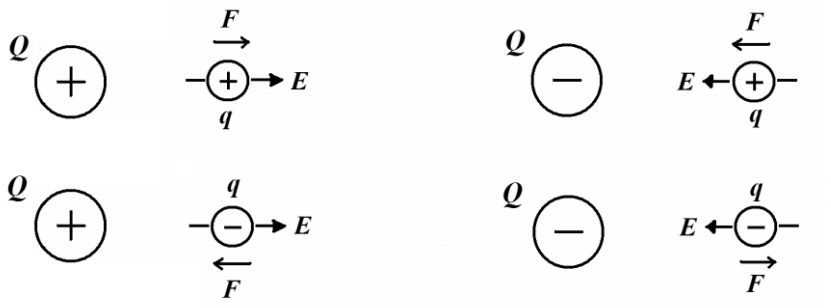
$$F = \frac{1}{4\pi\varepsilon_0}\frac{q\Delta q}{r^2}$$

Therefore, the magnitude of the electric field at a distance r from a point charge q is, from the definition $E = \frac{F}{\Delta q}$:

$$\boxed{E = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}} \qquad \text{(Field due to a point charge } q\text{)}$$

The field $\underset{\sim}{E}$ is directed radially away from the charge if q is positive and towards it if q is negative.

The diagrams below illustrate the direction of the electric field created by a charge Q and resulting electric force on a second charge q for four different cases; the field direction at the position of the second charge is shown.

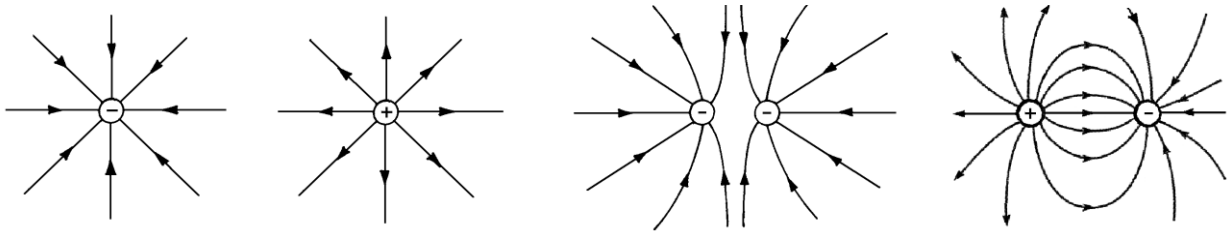


The **Principle of Superposition** states that the resultant field due to a number of point charges is the vector sum of the fields due to each.

Electric field lines

Field lines (or lines of force) were introduced by Faraday as an aid to visualising an electric field. They are imaginary lines showing the direction of the electric field at all points in space. In the absence of any other forces, a small positive test charge placed on a line will move along it in the direction of the field.

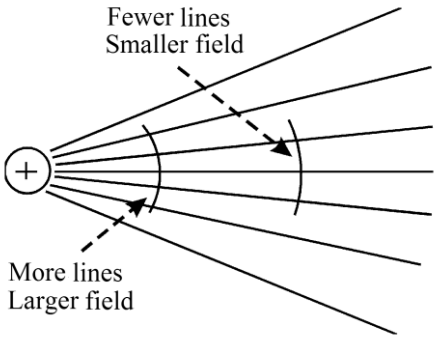
Some examples of field lines are shown below (the combination of equal positive and negative charges shown on the right is called an **electric dipole**); remember that field lines radiate in three dimensions.



- Field lines cannot cross one another since the field at a point can lie only in one direction.
- Field lines must start on a positive charge and end on a negative charge; they can also either start **or** end at infinity if there is an excess of one type of charge, as in three cases illustrated above.
- The density of lines in any region indicates the strength of the field in that region. In the two cases illustrated above on the left, the lines become further apart far from the charge, where the field is weaker.

Electrostatic flux

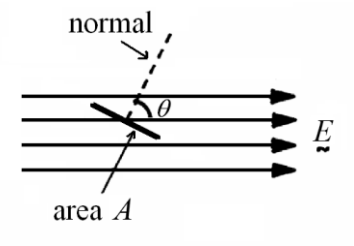
This is a measure of the number of field lines passing through a surface situated in an electric field.



The diagram shows two identical small surfaces placed perpendicular to the field lines radiating from a point charge (the surfaces are viewed side-on).

More lines pass through a surface placed where the field is large.

In fact, the number of lines passing through a surface perpendicular to the field lines is proportional both to the field strength E at the surface and to the area A of the surface. That is, $N \propto EA$: this product is called the **electric flux** through the surface.



If the surface is not perpendicular to the field, we must use the component of the field perpendicular to the surface

$$E_{\text{perp}} = E\cos\theta$$

where θ is the angle between the field and the normal to the surface.

Then the electrostatic flux through the surface is

$$\boxed{\phi_{\text{E}} = E_{\text{perp}}A} \qquad \text{(electric flux defined)}$$

- Note that this is **not the most general definition of electrostatic flux** – it is valid only if the field has the same value everywhere on the surface. This condition will be satisfied in all the examples considered in this course.
- The flux is a maximum when the field is perpendicular to the surface and zero when the field and surface are parallel.

Conductors in electrostatic equilibrium

A conductor is said to be in electrostatic equilibrium when there is no net motion of the free charges within it. An **isolated conductor**, one on which no **external** forces act (such as provided by a battery), has the following properties.

1. The electric field is zero everywhere inside the conductor.

   If the field were not zero, the free charges in the conductor would be subjected to an electrostatic force and would therefore move. The conductor would then not be in electrostatic equilibrium.

2. Any excess charge placed on an isolated conductor resides entirely on its surface.

   If excess charge is somehow placed inside a conductor, the repulsive Coulomb forces between the charges would push them towards the surface. It can be shown that with a force law such as Coulomb's Law, all the excess charge moves to the surface.

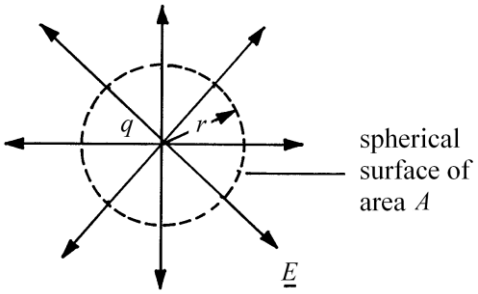
3. The electric field just outside a charged conductor is perpendicular to the conductor's surface.

   If this were not true, the field would have a component along the surface. This component would cause free charges on the surface of the conductor to move across the surface and the conductor would no longer be in electrostatic equilibrium.

Gauss's law

This law, due to Karl Friedrich Gauss (1777–1855), enables the electric field created by a distribution of charge to be calculated in a number of cases of sufficient symmetry.

To illustrate Gauss's law consider a point charge *q*, for which it was previously shown that the magnitude of the field is $E = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}$.



The diagram shows the field due to a positive point charge. For a negative charge the direction of the field must be reversed.

An imaginary sphere of radius r is drawn, centred on the point charge. The surface area of the sphere is $A = 4\pi r^2$.

- The magnitude of the field is the same everywhere on the surface of the sphere, since r has the same value everywhere on the surface.
- The field is normal to the surface everywhere, since the field lines go out radially. Hence $E_{\text{perp}} = E$.

Then the flux through the surface of the imaginary sphere is

$$\phi_{\text{E}} = E_{\text{perp}} A = E\,A = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}4\pi r^2 = \frac{q}{\varepsilon_0}$$

Thus the flux through a spherical surface centred on a point charge q is $\phi_E = q/\varepsilon_0$.

More generally, Gauss's law states that:

$$\boxed{\phi_{\text{E}} = \frac{Q}{\varepsilon_0}} \qquad \text{(Gauss's law)}$$

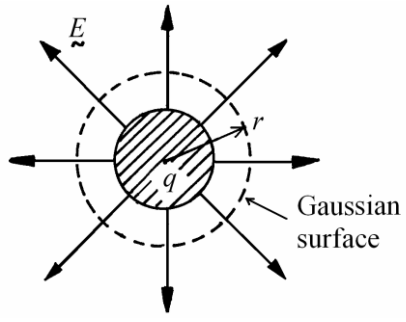
where ϕ_{E} is the flux through a closed surface and Q is the total charge enclosed by that surface.

- The law holds for any distribution of charge, not just for a point charge.
- The law is valid whatever the shape of the surface (the **Gaussian surface**) drawn around the charge distribution. The surface must however be a **closed** surface.
- The law is useful only if it is possible to draw a surface over which the flux can easily be calculated. In this course we consider only cases where a surface can be drawn that is either perpendicular to or parallel to the field.

The field due to a charged conducting sphere.

A spherical conductor of radius R carries a total charge q. The charge is assumed to be positive; otherwise the direction of the field in the diagram below must be reversed.

As discussed earlier, the field inside the sphere is zero, and just outside the sphere the field is perpendicular to the surface. Symmetry then requires that the field is radial everywhere outside the sphere.



To find the magnitude of the field outside the sphere, a spherical Gaussian surface of radius $r > R$ is drawn, concentric with the charged sphere.

- The field $\underset{\sim}{E}$ is normal to the surface everywhere and will by symmetry have the same magnitude E everywhere on the surface, so the flux through the surface is $\phi_E = E_{\text{perp}} A = E\,4\pi r^2$.
- The total charge enclosed by the surface is q.

By Gauss's law

$$\phi_E = E\,4\pi r^2 = \frac{1}{\varepsilon_0} q\,.$$

Therefore

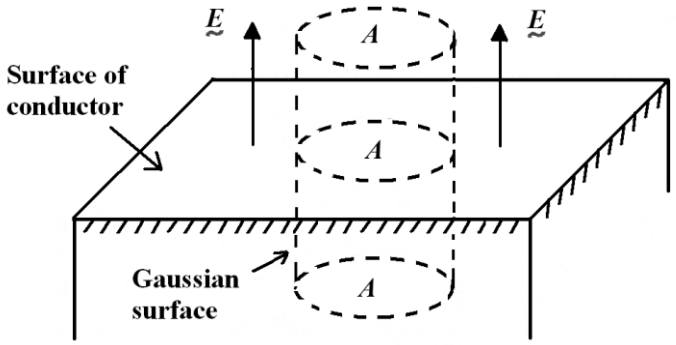
$$\boxed{E = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}} \qquad \text{(charged conducting sphere)}$$

The field is the same as if all of the charge were concentrated at the centre of the sphere (cf. gravitation: the gravitational field of the Earth can be found by assuming all the Earth's mass is concentrated at its centre).

The field due to a charged plane conducting surface.

Consider a plane conducting surface with excess charge uniformly distributed over its surface. The surface charge density (the charge per unit area of surface) is denoted σ.

The field **outside** the conductor is $\underset{\sim}{E}$. By symmetry, the field lines must be normal to the surface of the conductor (except near the edges). In the diagram below the excess charge is assumed to be positive; otherwise the direction of the field must be reversed.



A cylindrical Gaussian surface is drawn as shown, with base area A, with the base inside the conductor and the top above the surface, parallel to the surface.

- The field **inside** the conductor is zero, so there is no flux through the bottom of the Gaussian surface.
- The field is parallel to the curved side of the Gaussian surface and so the component of $\underset{\sim}{E}$ normal to the side is zero. Hence, there is no flux through the side.
- The only contribution to the flux is through the top of the Gaussian surface and so $\phi_E = EA$, since $\underset{\sim}{E}$ is perpendicular to A.

The only charge enclosed by the Gaussian surface is that on the section of the plane conducting surface inside the cylinder; this section has area A, so $Q = \sigma A$.

Gauss's law $\phi_E = Q/\varepsilon_0$ becomes $E\,A = \sigma A/\varepsilon_0$, giving

$$\boxed{E = \frac{\sigma}{\varepsilon_0}} \qquad \text{(plane conducting surface)}$$

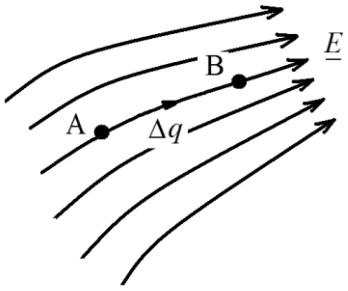
The field is normal to the surface of the conductor and is independent of the distance from the surface (this is strictly valid only if the distance from the surface is small compared with the distance from the edge of the conducting surface).

4.1.3. Electric Potential and Potential Energy

Electric potential energy

If we place a small positive test charge Δq in an electric field it will experience an electric force $\underset{\sim}{F}_{el} = \Delta q \underset{\sim}{E}$ where $\underset{\sim}{E}$ is the electric field at the point at which the charge is placed.

Consider for the moment the situation where no other forces act on the test charge.

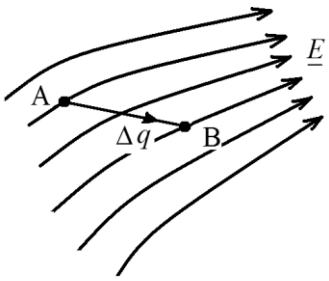


If the charge is released from rest at a point A in the field it will accelerate, moving along the field line that goes through its starting point.

As its kinetic energy increases, its **electric potential energy** must be decreasing so that its total energy is conserved.

- It follows that a charge will normally (if no other forces act on it) move towards a region of lower electrical potential energy, just as a massive body placed in the earth's gravitational field moves towards a region of lower gravitational potential energy.
- The **change in electric potential energy** that a charge in an electric field experiences is a result of the **work done on it by the electric force**, just as the gravitational potential energy of an object in a gravitational field is changed by the gravitational force.

Consider now the more general case, when other forces (such as gravity) may be acting on the charge in addition to the electric force.



If a charge Δq moves from any point A to another point B in an electric field, the change in its electric potential energy U is **defined** by

$$\boxed{\Delta U = U_{\mathrm{B}} - U_{\mathrm{A}} = -W_{\mathrm{A}\to\mathrm{B}}} \quad \text{(electric potential energy defined)}$$

where $W_{A\to B}$ is the work done by the electric field on the charge as it moves from A to B.

Note that the work done is independent of the path taken and depends only on the initial and final points of the motion (as is the case for the gravitational force).

For the **situation considered initially** (where the charge Δq is positive and the only force acting is the electric force):

- The charge is losing electric potential energy so that ΔU is negative.
- The work done $W_{A\to B}$ on the charge by the field is positive, since the electric force and the displacement of the charge are in the same direction.

Electric potential

A quantity more useful than electric potential energy is **electric potential**, which is defined as the **electric potential energy per unit charge**. Therefore the potential difference or **p.d.** between two points A and B in an electric field is:

$$\boxed{V_{\mathrm{B}} - V_{\mathrm{A}} = -\frac{W_{\mathrm{A}\to\mathrm{B}}}{\Delta q}} \qquad \text{(potential difference defined)}$$

where $W_{\mathrm{A}\to\mathrm{B}}$ is the work done by the field in taking a small test charge Δq from A to B.

- Note that potential difference is determined solely by the details of the charge distribution that creates the electric field; it does not depend on the magnitude of the test charge that is placed in the field in order to measure it.
- The unit of potential and potential difference is the **volt** V. From its definition we have $\mathrm{V} = \mathrm{J.C^{-1}}$. A p.d. of 1V exists between two points if +1 J of work is done against the field in taking +1 C of charge from the point at lower potential to the point at higher potential.
- Potential differences are sometimes referred to as **voltages**.

It follows from this definition that if the potential difference ΔV between two points in an electric field is known, the magnitude of the work that must be done to move a charge q between the points can be calculated from

$$\boxed{W = q\,\Delta V}$$

Only **differences in potential** or **changes in electric potential energy** have a physical significance. In order to define the potential **at a point** in a field, it is necessary to specify a **reference point** at which the potential is taken to be zero. In general this point can be chosen for convenience (as is the case for gravitational potential energy).

- The reference point is often taken to be infinity as is done for a point charge – see below.
- In an electric circuit, it could be a point connected to earth.

The general definition of electric potential is therefore:

> The potential at any point in an electric field is the work done (by us, against the electric force) per unit charge in bringing a positive test charge from the reference point to that point in the field.

Electron volt (eV)

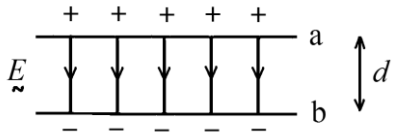
This is a unit of energy commonly used in atomic and sub-atomic physics. It is defined as the energy that an electron gains when accelerated through a potential difference of 1 V.

In general, when a charge q is accelerated through a potential difference ΔV, the work done is $W = q\Delta V$. If the particle is an electron of charge $q = 1{,}60 \times 10^{-19}$ C and if the potential difference is $\Delta V = 1$ V, then the work done is $W = q\Delta V = 1{,}60\times10^{-19}$ J. Hence

$$\boxed{1 \text{ eV} = 1{,}60\times10^{-19} \text{ J}}$$

Potential difference in a uniform electric field

A **uniform** electric field is one in which the magnitude and direction of the field are the same everywhere; the field lines are uniformly spaced and parallel. A simple device for creating a uniform electric field is shown in the diagram below.



It consists of two plane conductors, called plates, placed parallel to each other a distance d apart. The plates carry equal and opposite excess charges uniformly distributed over the inner surfaces.

The diagram shows a side view of the plates.

The field $\underset{\sim}{E}$ between the plates is:

- perpendicular to the plates, directed from the positively-charged plate to the negatively-charged plate;
- uniform except near the ends of the plates, with magnitude σ/ε_0 .

Work must be done on a small positive charge Δq to move it from the negative plate (b) to the positive plate (a), along a field line. The force that the field exerts on the charge has magnitude

$$F = E\,\Delta q \quad (\text{from } E = \frac{F}{\Delta q})$$

and is in the same direction as the field. The work done by the field is therefore

$$\begin{aligned} W_{\text{b}\to\text{a}} &= -Fd \qquad (\text{since } \cos\theta = -1) \\ &= -E\Delta q\, d \end{aligned}$$

The potential difference between points a and b, which by definition is the work done per unit charge $-W_{\text{b}\to\text{a}}/\Delta q$, is therefore

$$V_{\text{ab}} \equiv V_{\text{a}} - V_{\text{b}} = -\frac{W_{\text{b}\to\text{a}}}{\Delta q} = \frac{E\,\Delta q\, d}{\Delta q} = E\,d$$

The magnitude of the potential difference between any two points a and b a distance d apart in a uniform electric field of strength E is therefore given by

$$\boxed{E = \frac{V_{\text{ab}}}{d}} \qquad \text{(p.d. in a uniform field)}$$

- This equation leads to the unit V.m^{-1} for the electric field.

Potential due to a point charge

Calculation of the potential due to a point charge requires the specification of a reference point, at which the potential is defined to be zero; by convention this point is taken to be infinity.

- Infinity is often chosen as the reference point for measuring electric potential energy (and therefore electric potential) because the electrostatic force acting between two charges is zero when they are an infinite distance apart.
- The potential at some point in space due to a fixed point charge q is therefore the work that we must do, per unit charge, in bringing a small positive test charge Δq from infinity to that point.

Calculation of the work done is complicated by the fact that the electric force does not remain constant as the charge Δq is moved (from Coulomb's law it depends on the separation of the two charges), so that the simple definition of work done as the product of the force and displacement cannot be used.

However, using integral calculus it is easily shown that the work we must do to move the charge Δq from infinity to a point a distance r from the point charge q is:

$$\boxed{W = \frac{1}{4\pi\varepsilon_0}\frac{q\Delta q}{r}} \qquad \text{(P.E. of two point charges)}$$

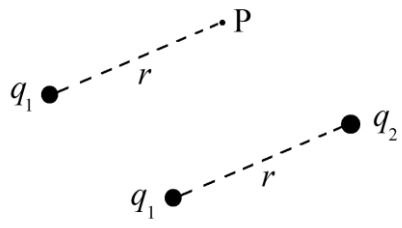
From the definition of electric potential energy, this is the potential energy of two point charges q and Δq a distance r apart. For two charges of the same sign the potential energy is positive; it is negative if the charges have different signs.

The potential at a distance r from a point charge q, by definition the work done per unit charge, is therefore:

$$\boxed{V = \frac{1}{4\pi\varepsilon_0}\frac{q}{r}} \qquad \text{(potential due to a point charge)}$$

- Since potential is a **scalar** quantity, to find the potential at a point due to several charges, simply calculate the algebraic sum of the potentials due to the individual charges.

The equations just derived are illustrated in the diagram below.



In the upper part of the diagram, the electric potential at point P a distance r from the point charge q_1 is

$$V = q_1/4\pi\varepsilon_0 r\ .$$

In the lower part of the diagram, the potential energy of two point charges a distance r apart is

$$U = q_1 q_2/4\pi\varepsilon_0 r\ .$$

Note that there are several very similar equations for point charges.

$$F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}, \quad E = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}, \quad V = \frac{1}{4\pi\varepsilon_0}\frac{q}{r}, \quad U = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r}$$

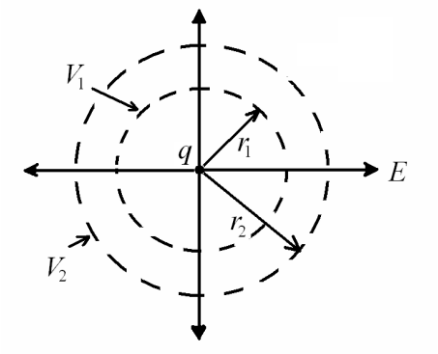
Do not confuse them, and note that they apply only to **point** charges (or spherical charge distributions).

They must **never** be used in connection with two charged parallel plates.

Equipotential surfaces

An **equipotential surface** is a surface over which the potential has the same value everywhere (this may or may not coincide with a physical surface).

As a **first example**, we consider a point charge q (assumed positive in the diagram below); the potential a distance r from a point charge q is $V = q/4\pi\varepsilon_0 r$.



It follows that the potential V_1 is the same everywhere on the surface of a sphere of radius r_1 centred on the charge.

On another sphere of radius r_2 there is a different constant potential V_2.

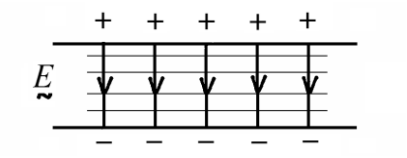
These surfaces are therefore equipotentials.

- No work is required to move a charge at constant speed on an equipotential surface, since the work done $W = q\Delta V$ is zero when $\Delta V = 0$.
- Equipotential surfaces are always perpendicular to the electric field, as in the example above.

  If a charge is moved on the equipotential surface, it follows from above that no work is done on the charge by the electric force. But the work done by a force can be zero only if the force is perpendicular to the motion. Hence the force, and therefore the field, must be perpendicular to the motion, i.e. to the surface.
- The surface and the whole volume of a conductor is an equipotential under **electrostatic** conditions.

  If it were not, there would be a potential difference between two points on or in the conductor and then free charges would flow. This would violate the electrostatic conditions.
- The surface (and volume) of an **insulator** is not necessarily an equipotential, since no charges are free to move even if a potential difference exists.

As a **second example**, consider the field between two parallel conducting plates that carry equal but opposite charges. As already discussed the electric field between the plates is uniform.



The diagram shows both the field lines and the surfaces of constant potential. These are planes perpendicular to the field and therefore parallel to the plates (which are themselves equipotentials).

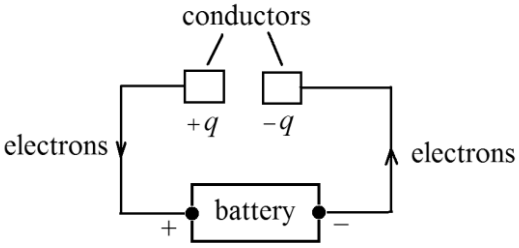
4.1.4. Capacitors and Capacitance

A capacitor is a device for storing charge (and therefore energy), normally consisting of two conductors placed near each other, but not in contact.

Capacitors are vital components of almost all electronic devices. Their ability to store and then release charge or energy is utilised directly in specialised devices such as camera flash units, defibrillators, high-powered pulsed lasers and emergency backup systems for computers.

Capacitance

A capacitor can be charged by connecting it to a battery (which is a device that maintains a constant potential difference between its terminals).



- Electrons are pulled from the conductor connected to the positive terminal, leaving it with a net positive charge, and transferred through the battery to the other conductor which then has an equal negative charge.
- This creates an electric field, and therefore potential difference, in the gap between the two conductors.
- The charge-transfer process stops when the potential difference between the conductors equals the p.d. provided by the battery.

If we measure the magnitude of the charge q on either conductor of a capacitor for different potential differences V between the conductors, we find that the ratio q/V is a constant for a given capacitor.

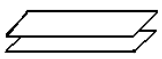
This ratio is defined as the **capacitance** C of the capacitor:

$$\boxed{C = \frac{q}{V}} \qquad \text{(capacitance defined)}$$

- The value of the capacitance depends on the size and shape of the conductors and their separation.
- The SI **unit of capacitance** is called the **farad** F. From the definition of capacitance, we have 1 F = 1 C.V^{-1}. Since the farad is a very large unit, the following are used: 1 μF = 10^{-6} F and 1 pF = 10^{-12} F.
- Capacitance is a measure of the amount of charge that a capacitor can store for a given potential difference between its two conductors.

The parallel-plate capacitor

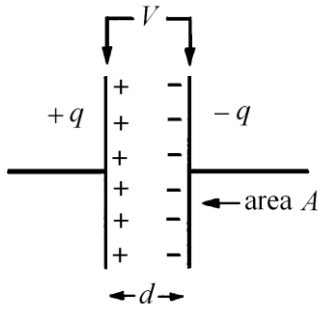
We describe here a parallel-plate capacitor, in which two plane conductors (the plates) are placed parallel to each other, a small distance apart.



The diagram on the left indicates both the physical appearance of a parallel plate capacitor and the pictorial representation of a capacitor in circuit diagrams.

This representation is used even if the capacitor is not of the parallel-plate construction.

Consider two parallel conducting plates, each of area A, a small distance d apart, which have been charged by connecting them to a battery.



The inner surfaces of the plates carry equal and opposite charges $+q$ and $-q$ and the p.d. between the plates is V (the charges accumulate on the inner surfaces because of the attractive forces between unlike charges).

The gap between the plates is filled with air, which is an insulator.

The field due to the charged conducting plates is $E = \sigma/\varepsilon_0$, where the surface charge density on each plates has magnitude $\sigma = q/A$. Therefore the field between the plates is $E = q/\varepsilon_0 A$.

The field and p.d. between the plates are related through $E = \frac{V}{d}$ so that $\frac{q}{\varepsilon_0 A} = \frac{V}{d}$, or $\frac{q}{V} = \frac{\varepsilon_0 A}{d}$, giving

$$\boxed{C = \frac{\varepsilon_0 A}{d}} \qquad \text{(parallel-plate capacitor)}$$

- One device that utilises directly the dependence of C on d is a computer keyboard. In some keyboards, pressing a key decreases the plate separation of a capacitor directly below the key. The charge-flow that results from the change in capacitance is detected and interpreted by the computer circuitry.
- Variable capacitors are also used in the tuning circuits of radios (the frequency to which the circuitry responds is determined by the capacitance in the circuit). One type of variable capacitor consists of two interwoven sets of metal plates, one fixed and the other movable; the capacitance depends on the size of the overlapping plate area.

For a capacitor with plates of area 1 cm^2 and plate separation 1 mm, the capacitance is 8,9 pF, which is very small. To increase the capacitance to about 1 nF we could:

(i) Increase the area of the plates by a factor of about 100. This has the obvious disadvantage of increasing the size of the capacitor considerably.

(ii) Decrease the plate separation to about 10 μm. However, for a fixed voltage this would, from $E = V/d$, increase the electric field between the plates, possibly leading to a breakdown of the air gap between the plates (which happens at a field of about 3 MV/m).

Another way to increase the capacitance, through the use of dielectrics, is described below.

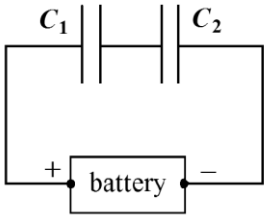
Capacitors in combination

Circuits normally contain a number of interconnected capacitors, to meet the specific needs of the device. Two simple ways to connect capacitors, in series or in parallel, are illustrated below.

- The **equivalent capacitance** of a network of capacitors is defined as the capacitance of a single capacitor that has exactly the same effect in the external circuit, in that it would, for the same potential difference, store the same amount of charge.

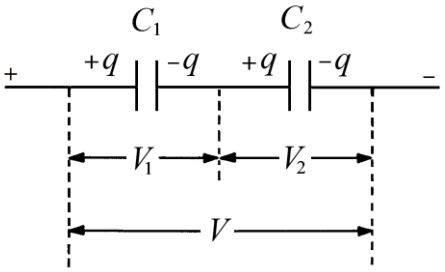
Series combination

We start with two uncharged capacitors of capacitances C_1 and C_2 connected in series.



A battery is now connected across the combination as shown in the diagram on the left. Because of differences in potential between the battery terminals and the capacitor plates, charges will now flow.

- Electrons are transferred from the left plate of the capacitor C_1 through the battery to the right plate of C_2, leaving the left plate of the capacitor C_1 positively charged and the right plate of C_2 negatively charged.
- The excess positive charge on the left plate of the capacitor C_1 attracts an equal amount of negative charge onto the right plate of capacitor C_1; this charge comes from the left plate of C_2. Since the two inner plates were initially uncharged, this leaves an excess positive charge of the same magnitude on the left plate of C_2.



- This process continues until the p.d. across the combination of capacitors equals the potential difference V provided by the battery. The charges on the plates are then as shown in the diagram above; each of the capacitors in series has the same charge on its plates.

The potential differences across C_1 and C_2 will in general be different:

$V_1 = \dfrac{q}{C_1}$ and $V_2 = \dfrac{q}{C_2}$,

The potential difference across the combination is

$$V = V_1 + V_2 = \frac{q}{C_1} + \frac{q}{C_2} = q\left(\frac{1}{C_1} + \frac{1}{C_2}\right).$$

Therefore

$$\frac{V}{q} = \frac{1}{C_1} + \frac{1}{C_2}$$

or

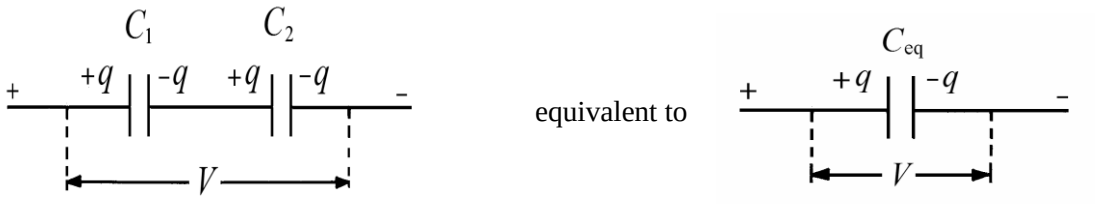
$$\boxed{\frac{1}{C_{\text{eq}}} = \frac{1}{C_1} + \frac{1}{C_2}} \qquad \text{(capacitors in series)}$$

where C_{eq} is the **equivalent capacitance** of the combination of two capacitors. For any number of capacitors C_1, C_2, C_3 ... in series

$$\frac{1}{C_{\text{eq}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots$$

Note that the combined capacitance is **smaller** than that of any individual capacitor.

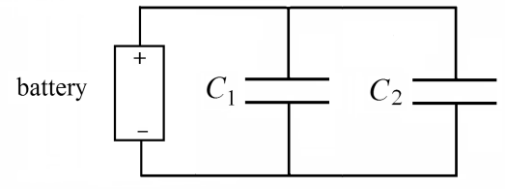
In summary, two capacitors of capacitances C_1 and C_2 can be replaced by a single capacitor of capacitance C_{eq}; for the same potential difference this will store the same amount of charge as the original capacitors.



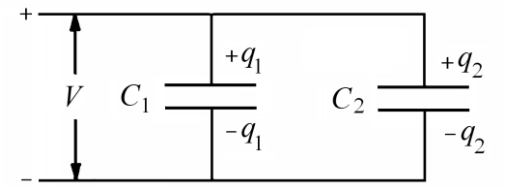
The charge on the equivalent capacitor is the same as the charge on each of the original two capacitors.

Parallel combination

Two capacitors that are initially uncharged are now connected in parallel with a battery. Because of differences in potential between the battery terminals and the capacitor plates, charges will flow.



- Electrons flow from the upper plates through the battery to the lower plates, leaving the upper plates positively charged and the lower plates negatively charged.



- This flow of charge stops when the p.d. across each capacitor equals the potential difference V provided by the battery. Note that for capacitors in parallel, the p.d. across each is the same, whereas the charges on each will in general be different:

$$q_1 = VC_1 \quad \text{and} \quad q_2 = VC_2$$

The total charge moved by the source is

$$q = q_1 + q_2 = VC_1 + VC_2 = V(C_1 + C_2)$$

Therefore

$$\frac{q}{V} = C_1 + C_2$$

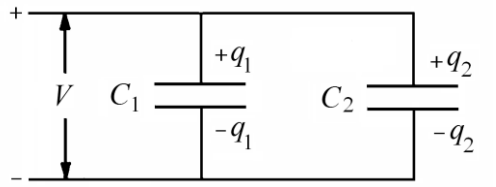
Or

$$\boxed{C_{eq} = C_1 + C_2} \qquad \text{(capacitors in parallel)}$$

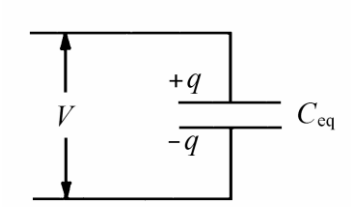
where C_{eq} is the equivalent capacitance. For any number of capacitors in parallel C_1, C_2, C_3....

$$C_{eq} = C_1 + C_2 + C_3 + \cdots$$

In summary, the two capacitors of capacitances C_1 and C_2 can be replaced by a single capacitor of capacitance C_{eq}; for the same potential difference this will store the same amount of charge as the original capacitors.



equivalent to



Note that the p.d. across the equivalent capacitor is equal to the p.d across each of the original two capacitors. The charge on the equivalent capacitor is the sum of the charges on the original two capacitors.

Energy stored in a charged capacitor

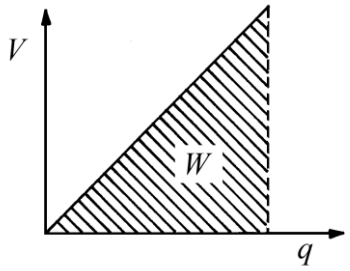
When a capacitor is charged, work must be done in transferring each small element of positive charge from the plate at lower potential to the plate at higher potential. The work done in transferring charge Δq when the p.d. between the plates is V is

$$\Delta W = V \Delta q$$

As more charge accumulates on the plates, the p.d. between the plates increases, and the amount of work that must be done to transfer the next element of charge therefore increases proportionately.

The total work done in charging the capacitor from zero p.d. to a final potential difference V can be found by considering a graph of V versus q. From $V = \frac{1}{C} q$, we see that plotting a graph of V against q results in a straight line of slope $1/C$ that passes through the origin.

The final charge q on the plates is moved against an *average* potential difference of $\bar{V} = V/2$, so that the total work done is $W = q\bar{V} = \frac{1}{2} qV$.



The total work done can also be found from the area under the curve: $W = \frac{1}{2} qV$ (recall the formula $W = \frac{1}{2} Fx$ for the work done in stretching a spring, which is also the area under a curve).

Using the relationship $q = VC$ two alternative formulas for the work done can be derived, yielding

$$\boxed{W = \tfrac{1}{2} qV = \frac{1}{2} \frac{q^2}{C} = \tfrac{1}{2} CV^2}$$ (energy stored in a capacitor)

The work done becomes energy stored in the capacitor; it can be released if the capacitor is discharged.

Dielectrics

We have so far assumed that the gap between the plates of a capacitor is filled with air. In practice there is usually a dielectric, i.e. an insulating medium, between the plates.

- Common dielectric materials include plastic, rubber and waxed paper.
- The main function of the dielectric is to increase the capacitance of the capacitor (how this happens is discussed later).

This effect of the dielectric is measured by the **dielectric constant** (or **relative permittivity**), defined by:

$$\boxed{\kappa = \frac{C}{C_0}} \qquad \text{(dielectric constant defined)}$$

where C_0 is the capacitance with **vacuum** between the plates, and C is the capacitance with the **dielectric** completely filling the gap between the plates.

- κ is a constant for a particular material.
- κ = 1 for vacuum, 1,00059 for air and is usually in the range 2–10 for common materials. However, κ can be much larger: for example for distilled water at room temperature κ = 80 and for barium strontium titanate $\kappa \sim 10000$.

Besides increasing the capacitance, the use of a dielectric has other benefits related to its **dielectric strength**; this is the field at which the dielectric breaks down, producing sparking between the plates and consequent loss of the stored charge. When a dielectric is present, the dielectric strength is typically much higher than for air.

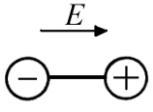
- Examples of dielectric strengths are 3 MV.m^{-1} for air, 16 MV.m^{-1} for paper and 60 MV.m^{-1} for Teflon.
- The dielectric insulates the plates from one another, allowing the capacitor to be charged to a higher voltage without breaking down; a higher voltage means more charge can be stored.
- Alternatively, because of the higher dielectric strength, the plates can be much closer together, thereby increasing the capacitance even further.

Microscopic description of dielectrics

To understand how the dielectric increases the capacitance of the capacitor, we consider what happens when a slab of dielectric material is placed in an applied electric field.

All charges in the dielectric will experience an electric force. Although there are no free charges in an insulator so that bulk movement of charges cannot take place, the charges in the molecules of the material will respond to the field. Positive charges will move slightly in the direction of the field and negative charges slightly in the opposite direction.

Consequently, the centres of gravity of the positive and negative charges in a molecule no longer coincide and the molecule behaves like an **electric dipole**.

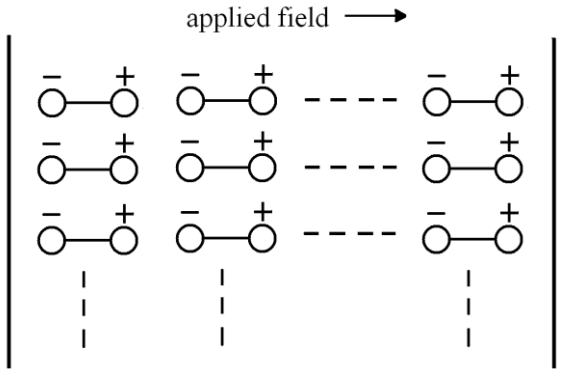


This is shown in the diagram on the left, where E indicates the direction of the applied field that causes the separation of charges.

This phenomenon is referred to as **induced polarization**.

- In some materials (such as water), molecules are naturally polarized but randomly orientated.

When a slab of dielectric is placed in the electric field between the charged plates of a capacitor, the dipoles that are created are aligned by the field.



In the diagram the applied field, created by the charges on the plates, points to the right.

The alignment of the dipoles effectively leads to an excess of negative charge on one face of the dielectric and an excess of positive charge on the other face.

This produces an **induced electric field** within the dielectric, in the opposite direction to the applied field but smaller in magnitude.

Therefore, when a slab of dielectric material is placed in the gap between the plates of a capacitor that has been charged by connecting it to a battery, the net electric field between the plates is reduced, the induced field partially cancelling the field due to the charges on the plates.

- The electric field between the plates is therefore **decreased** when the dielectric is inserted.

The magnitude of the potential difference between the plates and the charge on the plates may also change, depending on whether or not the battery remains connected to the capacitor when the dielectric is inserted.

If the capacitor is disconnected from the battery before the dielectric is inserted, the charge on each plate stays constant as no charge can arrive or leave. It follows that:

- Since $E = V/d$, V must **decrease**.
- Since $C = q/V$ and q cannot change, C **increases**.

However, if the capacitor is still connected directly to the battery when the dielectric is inserted, charges can move to or from the plates.

- Because the field between the plates has decreased, the p.d. between the plates **initially** decreases (according to $E = V/d$). The p.d. between the plates is then not equal to the p.d. provided by the battery to which it is connected, so charges must move.
- Negative poles on one face of the dielectric push more electrons off the positive plate of the capacitor, making it more positive. In the same way, the negative plate becomes more negative.
- This flow of charges from one plate to the other, through the battery, continues until the p.d. between the plates is again equal to the voltage supplied by the battery (this normally happens within a fraction of a second).
- The net result is that the charge stored on the capacitor has increased. From $C = q/V$, with V unchanged, this means that ***C* has increased**.

Note that in both cases the capacitance increases.

4.2. CURRENT ELECTRICITY

4.2.1. Electric Current

A major advance was made in the development of electricity in 1800 when Count Alessandro Volta (1745–1827) invented the battery, which converts chemical energy into electrical energy. A battery (and other devices invented subsequently) is able to maintain a constant potential difference between it terminals for long periods.

- If a potential difference V is established between two points in a body a distance d apart (by connecting the two points to the terminals of a battery, for example) an electric field will be created ($E = V/d$), and any charge q situated in the field will feel an electric force ($F = qE$).
- In a conductor there are charges which are free to move, so that if a potential difference exists between two points in a conductor the charges will move between the points, i.e. a **current** will be created.

The magnitude of the current is effectively the rate at which charge moves through the conductor. If an amount of charge q passes a given point in the conductor in time t, the current I is given by:

$$\boxed{I = \frac{q}{t}}$$

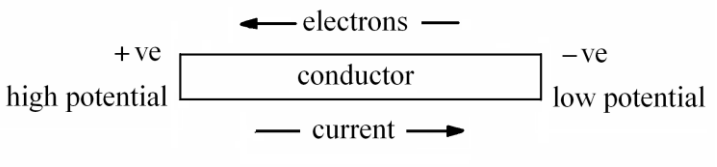
This equation defines the unit of charge, the coulomb. The **unit of current**, the ampere or amp (A), will be defined later in terms of the magnetic force between two current-carrying conductors.

The type of moving charges that form the current, called the **charge carriers**, depends on the conducting material.

- In a metal only electrons are free to move, and they will be attracted to the positive (i.e. high potential) end of a conductor when a potential difference is maintained across it.
- In solutions containing electrolytes (e.g. salt water) and in ionised gases (as in a fluorescent lamp, for example), ions of both sign are present and flow in opposite directions.
- In some semi-conductors movement of what are essentially positive charges occurs.

Negative charges flowing in one direction produce exactly the same effect as positive charges flowing in the opposite direction.

By convention, **positive current** flows from the high potential (positive) to the low potential (negative) end of a conductor.



4.2.2. Resistance and Resistivity

There is a **resistance** to the motion of charges through a conducting material.

In a metal, the moving electrons collide frequently with metal atoms (which are more or less fixed in the material); in copper for example a single electron may undergo as many as 10^{14} collisions each second. In each collision the electron loses its kinetic energy, which appears as heat, and is then accelerated again by the electric field.

The electrons move along a conductor with an average velocity, the **drift velocity**; because of the collisions this is very small (of the order of mm.s^{-1} or less), although their actual speed between collisions may be as large as $10^6\ \text{m.s}^{-1}$. It is the drift velocity that determines the magnitude of the current.

Resistance and Ohm's law

A systematic experimental investigation of the relationship between the current in a conductor and the potential difference between its ends was carried out by the German scientist, Georg Simon Ohm (1789–1854), who published his results in 1826.

Ohm found that for ordinary conductors, such as metals,

the current flowing in a conductor is proportional to the p.d. between the ends of the conductor

provided that the temperature of the conductor remains constant. This may be written

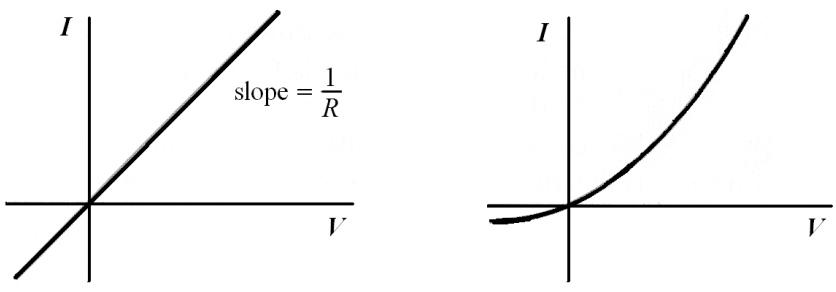
$$\boxed{V = I\,R} \qquad \text{(Ohm's law for a conductor)}$$

R is called the **resistance** of the conductor.

- The **unit of resistance** is V.A^{-1} which is called the **ohm** Ω.
- The **conductance**, $G = 1/R$ is used sometimes; it has unit $\Omega^{-1} \equiv S$ (siemens).

A large number of conductors (e.g. rectifiers, transistors, diodes) do not obey Ohm's law; they are referred to as **non-ohmic** conductors. Ohm's law also does not hold for the conduction of electricity through a gas.

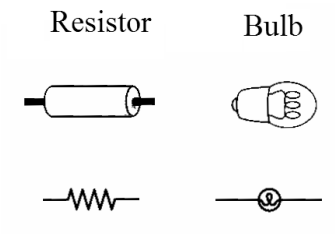
The resistance of a non-ohmic conductor can still be calculated from the formula $R = V/I$, but it is not constant and depends on the current flowing through the conductor. For example, rectifiers have resistances that depend on both the direction and magnitude of the current, being extremely large in one direction. This is illustrated in the diagram below which compares an ohmic device (i) with a semiconducting diode (ii).



(i) (ii)

A **resistor** is a conductor whose resistance obeys Ohm's law and whose function in an electrical circuit is to provide a specified resistance, which may range from an ohm or less to millions of ohms.

- A common type of resistor is the carbon-composition resistor, which ranges in size from about 1 Ω to several MΩ. It is constructed from a cylinder of carbon, in the form of graphite, mixed with non-conducting impurities to increase its resistance.
- Resistors for specialised use may be made by winding a fine wire around an insulating tube to get a long wire into a small space.



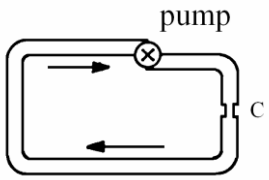
The diagram on the left shows the physical appearance of a resistor and indicates how it is represented pictorially in a circuit diagram.

The same symbol can also be used to represent any circuit element that provides a specific resistance in a circuit, although a separate symbol may be used for a light bulb.

A useful analogy

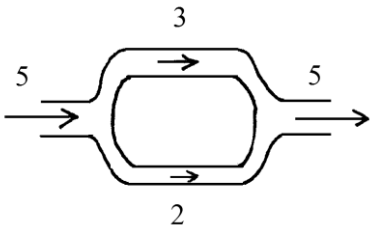
It is sometimes convenient to think of an electric circuit as being analogous to a pipe through which water is driven by a pump.

- A battery of specific **voltage** is analogous to a pump that delivers a specific **pressure**.
- The **current** is analogous to the water **volume flow rate**.
- Electrical resistance is analogous to the pipe's resistance to the flow of water.



If resistance to the flow of water is increased, e.g. by a constriction as at C, the flow rate in the whole pipe will decrease if the pressure remains constant. The pressure must be increased to maintain the same flow rate.

In the same way, if the resistance is increased in an electrical circuit, the voltage must be increased to maintain the same current; otherwise the current will decrease throughout the entire circuit.



If there are no branches in a pipe through which water is flowing, the flow rate will be the same at all points in the pipe (from the equation of continuity).

If, however, the pipe branches the sum of the flow rates in the branches must be equal to the flow rate in the main pipe.

In the same way, in a simple electrical circuit the current will be the same at all points in the circuit. If the circuit branches, the sum of the currents in the branches must equal the current in the main circuit. This follows from the conservation of charge and is often referred to as **Kirchhoff's junction rule**.

Resistivity

Ohm deduced from his experiments that the resistance of various samples of a particular material is proportional to the length and inversely proportional to the cross sectional area of the sample:

$$\boxed{R = \rho \frac{l}{A}} \qquad \text{(resistivity defined)}$$

where ρ is the **resistivity**, which is a constant (at constant temperature) for a particular material.

- From the equation $\rho = RA/l$, the **unit of resistivity** is Ω.m.
- The **conductivity** of a material is defined as $1/\rho$ and is measured in $\Omega^{-1}.m^{-1}$ or $S.m^{-1}$.
- As an example, the resistivity of copper at 20°C is $1{,}7 \times 10^{-8}$ Ω.m, to be compared with the resistivity of glass, an insulator, which is of the order 10^{12} Ω.m. Graphite, a semi-conducting material used in many resistors, has a resistivity of $3{,}5 \times 10^{-5}$ Ω.m

Effect of temperature on resistance and resistivity

The resistivity of a material depends somewhat on its temperature.

- The resistivity of metals increases with increasing temperature, since at higher temperatures the metal atoms have a larger amplitude of vibration and so impede the movement of electrons to a greater extent.
- Many metals (such as tin, lead and mercury), alloys and compounds become **superconducting** at very low temperatures, usually at only a few kelvin, where the resistivity suddenly drops to zero. This was discovered in 1911 by Heike Kamerlingh Onnes (1853–1926). Recently, some complicated compounds have been discovered that become superconducting at temperatures above 150 K.
- Semiconductors (used in diodes, transistors etc.) usually have a lower effective resistance at higher temperatures because increasing the temperature frees more charge carriers.

For **most metals** it is found that the resistivity increases approximately linearly with temperature (except just above absolute zero and at very high temperatures). So, provided the temperature range is not too large, the resistivity ρ at temperature θ is given by

$$\boxed{\rho = \rho_0 \left[1 + \alpha\left(\theta - \theta_0\right)\right]}$$

where ρ_0 is the resistivity at some reference temperature θ_0 (often taken as 20° C).

The proportionality constant α is called the **temperature coefficient of resistivity**, which is measured in $°C^{-1}$. (Do not confuse this quantity with the coefficient of linear expansion, for which the same symbol is used).

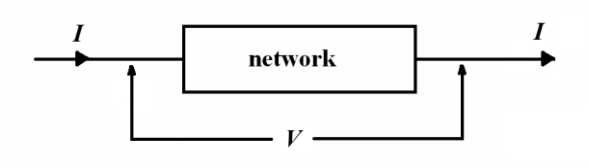
When materials are heated their resistivity changes, but so does the size of the specimen. For most materials the coefficient of resistivity is much larger than the coefficient of linear expansion (for example for aluminium, $4{,}3 \times 10^{-3}$ $°C^{-1}$ compared with $7{,}4 \times 10^{-5}$ $°C^{-1}$). This implies that changes in specimen size with temperature can be ignored compared with changes in resistivity, and then $R = \rho l/A$ gives

$$\boxed{R = R_0 \left[1 + \alpha\left(\theta - \theta_0\right)\right]}$$

Hence α can be regarded as the temperature coefficient of both resistance and resistivity.

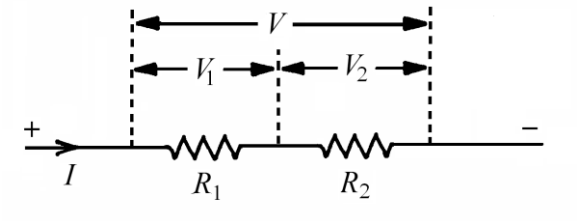
Resistors in Series and Parallel

The **equivalent resistance** of any network of resistances is defined by $R = V/I$ where V is the potential difference across the network and I is the current flowing into and out of the network. The network could be replaced by a single resistor of the equivalent resistance with no effect on the rest of the circuit.

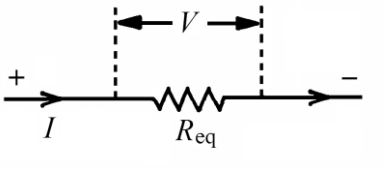


Series combination

Elements of a circuit are said to be connected in **series** if the same **current** flows through each, as in the diagram on the left. There can be no circuit branches between the resistors.



equivalent to



The current I is the same through both resistors, so that the potential difference across each resistor is:

$$V_1 = IR_1 \quad \text{and} \quad V_2 = IR_2$$

The potential difference across the combination is therefore:

$$V = V_1 + V_2 = IR_1 + IR_2 = I(R_1 + R_2).$$

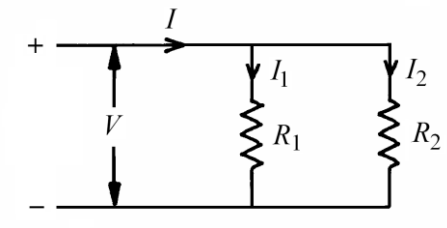
Hence the **equivalent resistance** is $R_{eq} = \dfrac{V}{I} = R_1 + R_2$.

It follows that for any number of resistances in series:

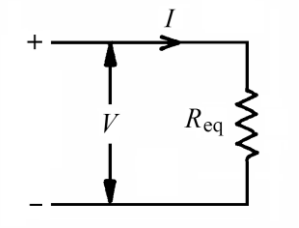
$$\boxed{R_{eq} = R_1 + R_2 + R_3 + \cdots}$$

Parallel combination

Elements of a circuit are said to be connected in **parallel** if the same **potential difference** exists across the elements.



equivalent to



The current flowing into the network splits, with I_1 going through one resistor and I_2 through the other. The potential difference V is the same across both resistors, so that the current through each is:

$$I_1 = \frac{V}{R_1} \quad \text{and} \quad I_2 = \frac{V}{R_2}$$

The total current entering the network is therefore:

$$I = I_1 + I_2 = \frac{V}{R_1} + \frac{V}{R_2} = V\left(\frac{1}{R_1} + \frac{1}{R_1}\right)$$

Hence $\frac{I}{V} = \frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2}$, where R is the **equivalent resistance**.

By extension, for any number of resistances in parallel:

$$\boxed{\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots}$$

Note that the combined resistance is always **smaller** than that of any individual resistor.

4.2.3. Sources of emf

A source is a device that provides a potential difference between its terminals, and can therefore do work in driving a current around a closed circuit. The source of energy used to do this depends on the type of source.

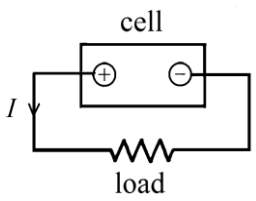
- A cell converts chemical energy into electrical energy. Note that strictly speaking a battery comprises several cells connected together in series to increase the potential difference provided, although single cells are often referred to as batteries.
- A generator in a power station converts mechanical energy into electrical energy in a turbine (although the ultimate source of energy may be the burning of gas, oil or coal or nuclear fission).
- A solar cell converts solar energy into electrical energy.

The simplest cells contain two plates or rods made of dissimilar metals, called electrodes. The electrodes are immersed in a solution called the electrolyte, which in a dry cell is absorbed in a powdery paste. The chemical energy results from complicated chemical reactions between the electrodes and the electrolyte. Each terminal of the cell is connected to one of the electrodes.

- The common dry cell used to power portable devices such as radios, torches, remote controls and so on normally provides a potential difference of 1,5 volts.
- A rechargeable nickel-cadmium cell has an output of 1,2 V and a lithium cell such as used in watches and some cameras as much as 3,0 V.
- A car battery of the lead-acid type comprises six 2-V cells connected together in series to provide an output of 12 V.

emf

When a source is connected in a circuit (as in the circuit diagram below), current flows from the positive to the negative terminal of the source **around the external circuit**.



For a continuous flow in the circuit, the current **inside the source** must flow from negative to positive (i.e. "uphill"). This is possible because there is another force acting within the source in addition to the electrostatic force that causes current flow in the external circuit. This force (with a chemical origin in a cell) is larger than the electrostatic force and acts in the opposite direction to it.

A source is an energy converter, converting chemical energy, for example, into electrical energy. The conversion is measured by the **emf**, $\mathcal{E}$, which may be defined as the **electrical energy delivered** by the source **per unit charge** passing through it, or equivalently, as the work done per unit charge in moving charge q through the source; i.e.

$$\boxed{\mathcal{E} = \frac{W}{q}} \qquad \text{(definition of emf)}$$

- emf is measured in volts, and is often referred to loosely as the "voltage" of the source.
- The emf is **not** a force, despite the origin of its name (electromotive force).

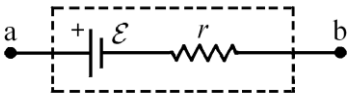
Internal resistance and terminal voltage

It follows from the definitions of emf and potential difference that:

The emf of an ideal cell is equal to the p.d. delivered by the cell.

However, for all cells, and other sources of emf, there is a resistance to the movement of charge within the cell. The cell is said to have an **internal resistance**, usually denoted r.

A real cell or battery is therefore represented as an emf $\mathcal{E}$ in series with its internal resistance r, as indicated in the diagram below.



The points a and b represent the two terminals of the battery.

- If no current is drawn from the battery the pd. between its terminals, the **terminal voltage**, is equal to the emf of the battery (this is often called the open-circuit voltage).
- However, if a current I flows from the battery there will be a drop in potential across the internal resistance, which is equal to Ir from Ohm's law.

Thus the terminal voltage is:

$$\boxed{V_{ab} = \mathcal{E} - Ir} \qquad \text{(p.d. across terminals of cell)}$$

assuming the current flows from b to a in the diagram above – see the later discussion. The potential difference measured across the terminals of a source thus depends on the current drawn from the source.

The internal resistance of a battery is usually quite small. For example, a torch battery may have an internal resistance of less than 0,1 Ω (which increases considerably as the battery ages) and a car battery has an even smaller internal resistance, typically about 0,005 Ω when in good condition.

4.2.4. Electrical Circuits

In this course we consider only simple circuits containing a limited number of circuit elements such as batteries, resistors and capacitors that have been discussed in previous sections.

These are assumed to be connected together using wires with **negligible resistance**; there is zero potential difference across a conductor with zero resistance, whatever the current (from $V = IR$), so these wires have no effect on the circuit.

Work and power in electric circuits

If an amount of charge q is taken through a potential different V in time t, the work is done on it is

$$\begin{aligned} W &= qV \quad \text{(from definition } V = W/q\text{)} \\ &= I\,tV \quad \text{(from } q = I\,t\text{)} \end{aligned}$$

The rate at which energy is delivered to a circuit element is therefore, from $P = W/t$,

$$\boxed{P = IV} \qquad \text{(power delivered to a circuit element)}$$

where I is the current passing through it and V is the potential difference across it.

From $P = W/t$ the unit of power is J.s^{-1}, which is defined as the watt (W).

What happens to the energy delivered to a circuit element depends on the nature of the device.

- In a light bulb, which contains a tiny wire element, it becomes heat and light energy.
- In devices such as electric heaters, toasters, kettles and hair dryers thermal energy is produced in the resistance wire, which is referred to as the heating element.
- In a motor it is converted to mechanical energy.
- In a loudspeaker it is turned into sound energy.

The formula $P = IV$ is valid for any electrical device. For an ohmic conductor, the potential difference, current and resistance are related by $V = IR$, and the expression for power can be written in two alternative forms:

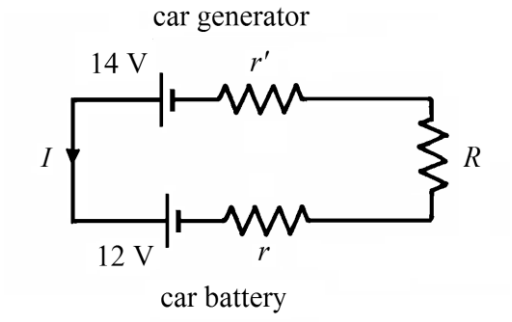
$$\boxed{P = I^2R = V^2/R} \qquad \text{(power loss in an ohmic conductor)}$$

The rate at which a source of emf converts energy (or supplies energy to the circuit) follows from the definition of emf: $\mathcal{E} = W/q$. When combined with $I = q/t$ this leads to $W = I\,t\mathcal{E}$

Therefore, the power delivered to the circuit by the source is, from $P = W/t$,

$$\boxed{P = I\mathcal{E}} \qquad \text{(power delivered by source)}$$

Usually the current in a circuit flows in the direction in which a source would normally drive it and the source supplies energy to the circuit. However, in circuits containing more than one source, the current may flow in the opposite direction to which a particular source would normally drive it. Then energy may be stored in the source (this happens, for example, if the chemical reaction within a cell is reversible).

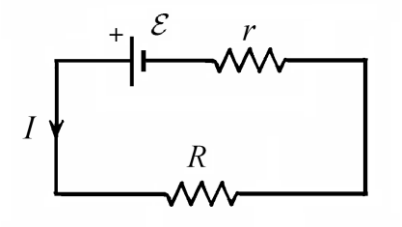


This is the situation shown in the diagram to the left, in which a car battery is being charged by the car's generator.

Note that ordinary torch cells cannot be recharged; then energy is not stored in the cell but is dissipated as heat.

The circuit equation

This is the basic equation for all circuit calculations. It is a consequence of the conservation of energy.



In the circuit to the left, a source has emf $\mathcal{E}$ and internal resistance r.

It is connected to an external resistor of resistance R and a current I flows around the circuit.

Power delivered to the circuit by the source = $I\mathcal{E}$

Power dissipated in the internal resistance (as heat) = I^2r

Power dissipated in the external resistance (as heat) = I^2R

From the conservation of energy, the rate at which energy is converted in the source must equal the rate at which it is dissipated in the internal and external resistances:

$$I\mathcal{E} = I^2r + I^2R$$

leading to

$$\boxed{\mathcal{E} = I(r + R)} \qquad \text{(circuit equation)}$$

- If the source is connected to more than one external resistor, R must be replaced by the equivalent resistance of the external circuit.
- Similarly, if the circuit contains more than a single source, the quantities $\mathcal{E}$ and r must be suitably modified.

Calculating potential differences

The circuit equation can be rewritten

$\mathcal{E} - Ir - IR = 0$.

Each term on the left-hand side is the potential difference across the corresponding circuit element.

- It follows that the sum of all the voltage changes in going round the complete circuit must be zero.
- This same principle can be applied to any closed current loop in a more complicated circuit and is often referred to as **Kirchhoff's loop rule**. Complicated circuits are not considered in this course.

To calculate the potential difference between two points in a circuit, we use the following:

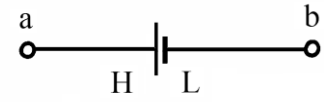
(a) Current always flows from high potential (H) to low potential (L) in the *external* circuit.

(b) The positive terminal of a source is always higher in potential by an amount $\mathcal{E}$ than the negative terminal (the internal resistance must be considered separately).

Four simple cases are considered in the diagrams below. In each case, we go from point a to point b to calculate $V_a - V_b \equiv V_{ab}$, the potential difference between points a and b:

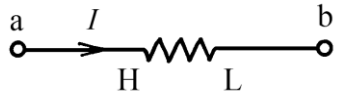
The potential increases by $\mathcal{E}$ when a source is crossed in the forward direction of the emf:

$$V_a + \mathcal{E} = V_b \text{ or } V_a - V_b = -\mathcal{E}$$



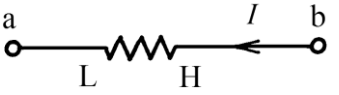
The potential decreases by $\mathcal{E}$ when a source is crossed in the backward direction of the emf:

$$V_a - \mathcal{E} = V_b \text{ or } V_a - V_b = +\mathcal{E}$$



The potential decreases by IR when a resistor is traversed in the direction of the current:

$$V_a - IR = V_b \text{ or } V_a - V_b = +IR$$



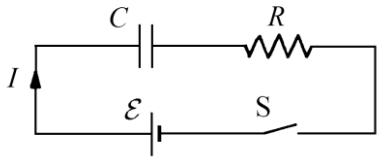
The potential increases by IR when a resistor is traversed in the opposite direction to the current:

$$V_a + IR = V_b \text{ or } V_a - V_b = -IR$$

RC circuits and charging a capacitor

In the circuits discussed so far currents and voltages have been constant. We now briefly consider circuits containing capacitors, in which the current initially varies with time.

Consider the following series circuit in which the capacitor is initially uncharged.



When switch S is closed, the cell immediately begins to charge the capacitor and a current passes through the resistor.

- Remember that charges cannot move directly across the gap between the capacitor's plates, so that no current flows **through** the capacitor. The current is created by charges moving from one plate to the other through the rest of the circuit, via the cell and resistor.

At any instant during the charging process the circuit equation is found by moving clockwise around the circuit from the switch:

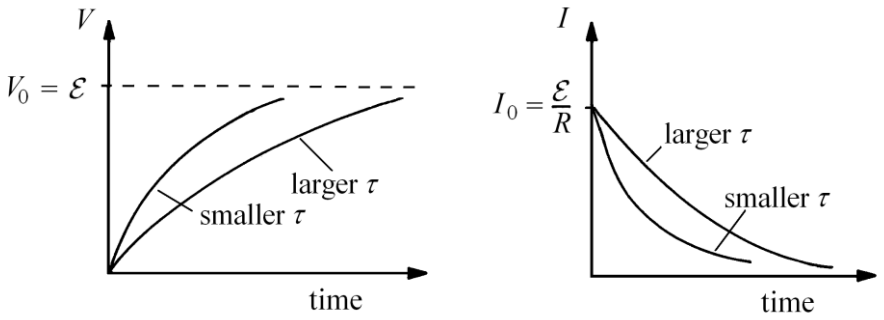
$$\mathcal{E} - \frac{Q}{C} - IR = 0$$

where I is the current at that instant and Q is the charge on the capacitor plates at the same instant.

Note that the potential difference across the capacitor appears with a negative sign in this equation – the left plate is connected to the positive terminal of the cell and is therefore at a higher potential than the other plate; there is therefore a decrease in potential as we move across the capacitor in the direction of the current.

- At the moment the switch is closed, $Q = 0$ so that from the circuit equation the initial current is $I_0 = \mathcal{E}/R$.
- The charging process continues until the potential difference across the capacitor equals the emf of the cell. At that time $I = 0$ and therefore $Q = C\mathcal{E}$.

Using integral calculus, it easy to discover how the current, the potential difference across the capacitor and the charge on the capacitor each change with time. The variation of the voltage and current with time is illustrated in the following diagrams.



Since $Q = CV$ with C constant, the charge Q on the capacitor increases at the same rate as V.

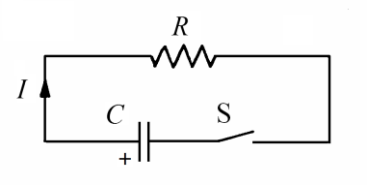
The rise and decay are **exponential** and involve the expression $\exp(-t/\tau)$ where

$$\boxed{\tau = RC} \qquad \text{(time constant)}$$

is called the **time constant** of the circuit. This is a measure of how quickly the capacitor becomes charged – after a time $t >> \tau$ all values are steady, having reached their final values. For example, the potential difference across the capacitor reaches more than 99% of its maximum value at $t = 5\tau$ and is thereafter effectively constant.

As an example, if R = 1,0 MΩ and C = 1 μF then τ = 1,0 s, and after about 5 seconds all values are effectively constant.

After charging, the capacitor can subsequently be discharged using a circuit such as that in the diagram below.



When the switch is closed, charge begins to flow from one plate of the capacitor to the other through the resistor, until the capacitor is fully discharged.

It is easily shown that the p.d. across the capacitor decreases with time from its initial value V_0 according to the equation

$$V = V_0 \exp(-t/\tau)$$

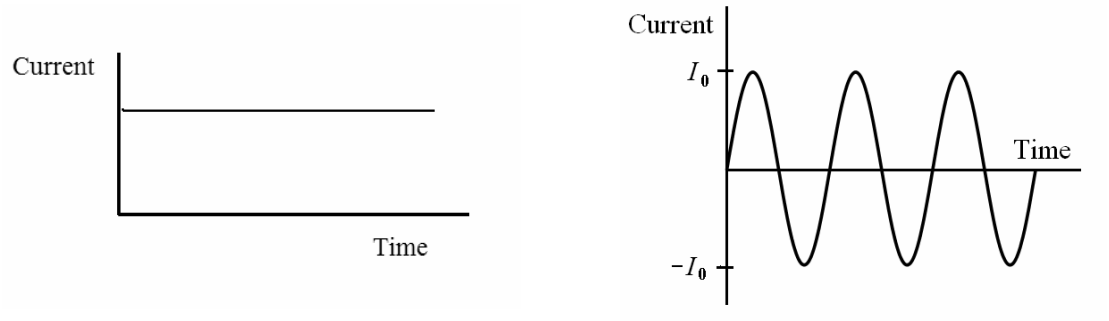
where again $\tau = RC$. The charge Q on the capacitor also decreases at the same rate (from $Q = CV$).

Examples of the use of RC circuits include camera flash units, heart pacemakers and the windshield wipers on a car. In the latter, the intermittent operation used in light rain is controlled by an RC circuit with an adjustable time constant (through selecting different values of R).

Alternating current

The current driven by a battery will flow through a circuit in one direction only, and its magnitude will be constant almost immediately the circuit is connected. This is referred to as **direct current** (DC).

On the other hand electric generators at power stations produce **alternating current** (AC), whose magnitude changes sinusoidally with time, reversing direction many times per second. The two currents are compared in the diagrams below.



The voltage produced by a generator can by written

$$V = V_0 \sin 2\pi ft$$

where V_0 is called the **peak voltage** and f is the **frequency**, i.e. the number of complete oscillations made per second. In South Africa and most other countries f = 50 Hz is used.

If a potential difference V exists across a resistance R in an AC circuit, the current through the resistor is, from Ohm's law,

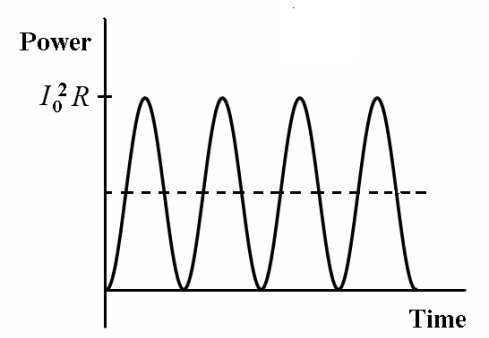
$$I = \frac{V}{R} = I_0 \sin 2\pi ft$$

where $I_0 = V_0/R$ is the **peak current**, as indicated in the previous diagram. Note that the *average* current is zero.

The power delivered to the resistance R at any instant is

$$P = I^2R = I_0^2 R \sin^2 2\pi ft$$

which oscillates between zero and $I_0^2 R$, as shown in the following diagram.



Since the average value of the sine function squared is ½, the *average* power delivered to the resistance is

$$\bar{P} = \frac{1}{2} I_0^2 R = \frac{1}{2} \frac{V_0^2}{R}$$

as indicated by the dashed line in the diagram.

It follows that, as far as power is concerned, the important quantities in an AC circuit are the mean values $\overline{I^2} = \frac{1}{2} I_0^2$ and $\overline{V^2} = \frac{1}{2} V_0^2$. The square roots of these quantities are called the root-mean-square (**rms**) values of the current and voltage:

$$I_{\text{rms}} = \frac{I_0}{\sqrt{2}} \quad \text{and} \quad V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$$

It is usually the rms voltage that is quoted for an AC supply; in South Africa V_{rms} is 230 volts.

4.2.5. Special circuits and devices

Ammeters and voltmeters

These are devices for measuring currents and potential differences in circuits; both are based on an instrument called a **galvanometer** that will be described later.

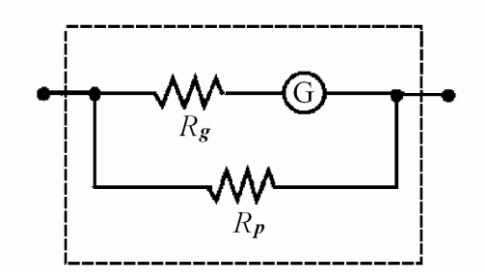
In the usual type of DC meter the deflection produced on a scale is proportional to the current through the meter; by Ohm's law it is also proportional to the voltage across the meter, since the meter resistance R_g is constant.

- The meter can therefore be used to measure either the current flowing through it or the potential difference across its terminals.

An **ammeter** is an instrument that measures the current through a circuit component. To ensure that the same current flows through the meter and the component, it must be connected in **series** with the component.

- The meter resistance must therefore be **low** (compared with other resistances in the circuit), so that it does not alter appreciably the current in the circuit.

The ammeter is constructed by connecting a resistor of low resistance (the **shunt resistor**) in parallel with the galvanometer, as shown in the following diagram.



The galvanometer G of internal resistance R_g is connected in parallel with a shunt resistance R_p to produce an ammeter, shown enclosed by the dashed line in the diagram.

The resistance R_p must be very small for an ideal ammeter.

The shunt resistor serves two purposes.

(i) The resistance of the ammeter in the circuit is equal to the equivalent resistance of the galvanometer and shunt resistor in parallel, which will be less then the resistance of the shunt and will therefore be very small, as required.

(ii) The galvanometer is a very sensitive instrument that gives a full-scale deflection for a very small current, usually of the order of a few mA or less. If a large current is to be measured

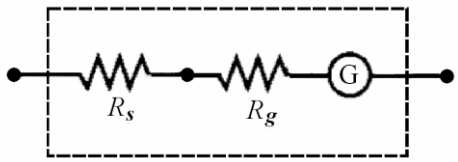
by the ammeter, some of the current in the main circuit must be made to bypass the galvanometer. By choosing a shunt resistor with the appropriate resistance, we can construct an ammeter of any desired range (as illustrated in the next lecture example).

A **voltmeter** is an instrument used to measure the p.d. between two points in a circuit; it must therefore be connected in **parallel** with a circuit element.

If the current drawn by the meter is not negligible, the current in the circuit will change when the meter is connected and the p.d. will be lower than it was before the meter was connected.

- The voltmeter must therefore have a very **high resistance**.

This is achieved by connecting a resistor of high resistance in series with the galvanometer.



This is illustrated in the diagram on the left. The dashed line encloses the entire voltmeter.

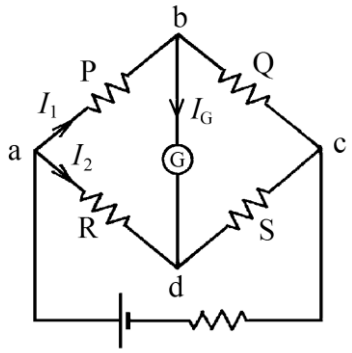
The resistance R_s must be large for an ideal meter.

This resistor limits the current through the galvanometer, and in addition its magnitude determines the range of voltages that the voltmeter can measure.

Wheatstone bridge

The Wheatstone bridge is a circuit that can be used for measuring accurately resistance (and also, suitably modified, capacitance).

In the circuit shown here, P, Q, R and S are resistors. One resistor has an unknown resistance that is to be measured, a second resistor has a resistance that can be varied in a controlled way, and the other two resistors have fixed, known resistances.



The bridge is said to be **balanced** when the current I_G through the galvanometer is zero.

Balance is achieved by adjusting the resistance of the variable resistor until the galvanometer reads zero.

At balance:

- The same current I_1 will flow in P and Q and the same current I_2 will flow in R and S, since $I_G = 0$.
- Points b and d are at the same potential (otherwise a current would flow between them).

Hence, using Ohm's law

$$V_{ab} = V_{ad} \quad \rightarrow \quad I_1 P = I_2 R$$

and

$$V_{bc} = V_{dc} \quad \rightarrow \quad I_1 Q = I_2 S \, .$$

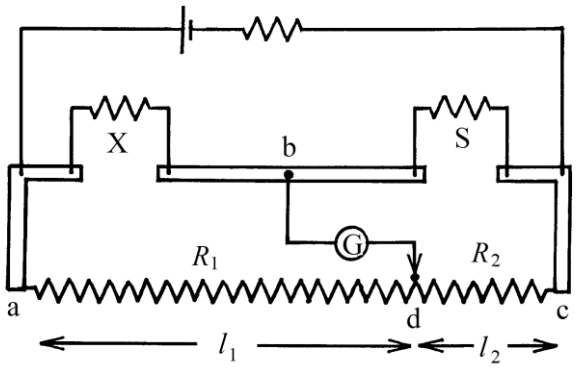
Dividing:

$$\boxed{\frac{P}{Q} = \frac{R}{S}}$$

Knowing the value of three resistances, the fourth can be determined.

Slidewire bridge

This is a practical implementation of the Wheatstone bridge circuit. In the circuit below, the points a, b, c and d refer to the same points with the corresponding labels in the Wheatstone bridge circuit.



S is a standard (known) resistance and *X* is the unknown resistance.

ac is a length of wire of uniform cross-section.

The resistors X and S are connected to points a, b and c by thick strips of copper with negligible resistance.

Balance is achieved by moving the sliding contact at d along the slidewire ac until the galvanometer reads zero. Then l_1 and l_2 are the lengths of the slidewire on either side of the balance point d.

The resistance of length l_1 is $R_1 = \rho \frac{l_1}{A}$ and the resistance of length l_2 is $R_2 = \rho \frac{l_2}{A}$, where *A* is the cross-sectional area of the wire and ρ its resistivity.

At balance

$$\frac{X}{S} = \frac{R_1}{R_2} = \frac{\rho l_1}{A} \frac{A}{\rho l_2} = \frac{l_1}{l_2} \, .$$

Hence

$$\boxed{X = S \frac{l_1}{l_2}}$$

Therefore *X* is found by measuring l_1 and l_2.

4.3. ELECTROMAGNETISM

4.3.1. Magnetic Forces and the Magnetic Field

It has been known for a very long time that certain materials have the property of attracting iron; these materials are said to be **magnetic**. Only iron and a few other materials such as cobalt and nickel show strong magnetic effects.

If the magnetic material is shaped into a bar or rod, most of the magnetism seems to be concentrated at the ends of the bar. When the bar is suspended at its centre, it will always come to rest with one particular end pointing roughly towards the earth's north pole.

The end of a magnet that points towards the north is called its **north** (seeking) **pole**. The other end is referred to as the **south** (seeking) **pole**.

If two magnets are brought close to one another, each is found to exert a force on the other. Experiments show that:

Like poles repel, unlike poles attract

Magnetic poles are in some ways similar to electric charges (there are two kinds, each exerts a force on the other), but there are important differences. For example:

- Isolated magnetic poles have not yet been detected although there are theoretical grounds for believing that they may exist. Note that cutting a bar magnet in half simply produces two smaller bar magnets, each with its own north and south poles.

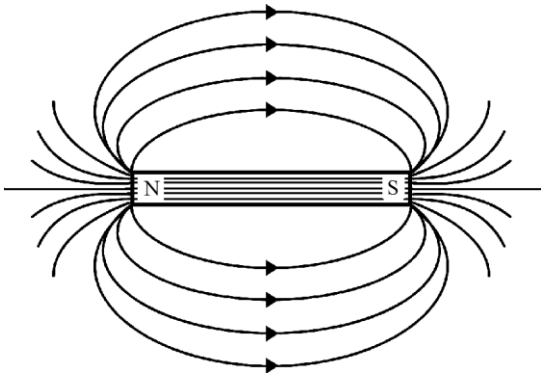
Magnetic fields

Earlier, we described the interaction between two charges by saying that one charge creates an electric field, and that this field exerts an electric force on the second charge.

In the same way a **magnetic field** is said to exist at a point if a force acts on a magnetic pole (e.g. one end of a compass needle) placed at that point. (A **compass** comprises a small bar magnet pivoted at its centre of gravity so that it can rotate freely in a horizontal plane.)

- The direction of the field at any location in space is the direction in which the north pole of a compass needle would point at that location.
- As is the case with the electric field, the magnetic field can be represented by **magnetic field lines**, the tangent to the lines indicating the direction of the field and the density of lines indicating the intensity of the field.
- Unlike the situation for electric fields, the magnetic field lines do not indicate the direction of the magnetic force (which is discussed below) and should therefore not be called lines of force.
- Again in contrast to the situation for electric field lines, magnetic field lines do not start and end on magnetic poles; they form continuous loops (which will pass through the poles).

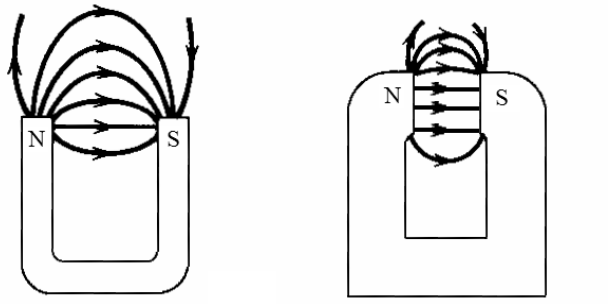
The diagram below shows the magnetic field created by a bar magnet.



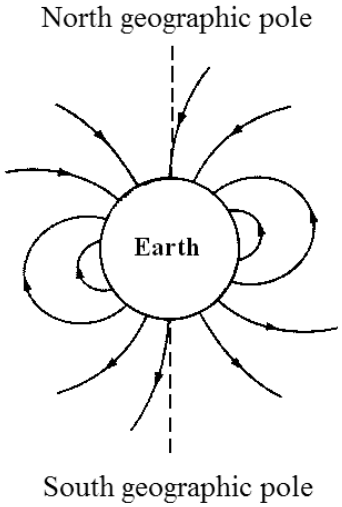
The field lines go from north to south outside the magnet, and from south to north inside the magnet to form closed loops.

The density of lines shows that the field is strongest near the poles of the magnet.

The following diagram shows the field created by a horseshoe magnet (the field inside the magnet is not shown). In the configuration on the right, the field between the poles is almost uniform.



The Earth creates a natural magnetic field, with field lines that roughly resemble those of a bar magnet.



Since the north pole of a compass needle points north, the magnetic pole that is situated near the Earth's geographic north pole is actually a south pole magnetically, although it is often referred to as the "north magnetic pole".

The magnetic poles do not coincide with the geographic poles, being about 1500 km apart at present, so that the deviation of magnetic north as indicated by a compass from true north may be considerable. This angle, the **magnetic declination**, varies from point to point on the Earth's surface.

The position of the poles is not fixed, the north pole moving at about 40 km/year over the past few years. In fact, it is believed that the field reverses its direction completely at irregular intervals, reversals being about 300,000 years apart on average (the last reversal was about 780,000 years ago).

As is obvious from the diagram, the Earth's magnetic field is not tangential to the Earth's surface in general. The angle that the field makes with the horizontal is called the **angle of dip**, and again it varies considerably with location on the surface of the Earth, being close to zero near the equator and almost 90° near the pole.

The Earth's field is believed to be caused by the electric currents in the liquid part of its core, and is not due to permanently magnetised material in the core.

Magnetic flux density

Experiments indicate that a charged particle moving through a magnetic field experiences a magnetic force; this is used to define the magnitude of the magnetic field. (The force on a magnetic pole cannot be used in the definition, since single magnetic poles do not exist.)

As previously discussed, the strength of an **electric** field is defined as $\underset{\sim}{E} = \underset{\sim}{F}/q$ where $\underset{\sim}{F}$ is the force on a stationary test charge q placed in the field. The direction of the force on a positive charge is parallel to the direction of the field, and a force is exerted whether the charge is moving or not.

In the case of a **magnetic** field:

- A force acts only on a **moving** charged particle.
- The force depends on both the magnitude and direction of the particle's **velocity**, as well as on the magnitude of its charge.
- The force always acts at **right angles** to the directions of both the velocity and the field.

The **flux density** of a magnetic field is defined in terms of the force on a charge moving perpendicular to the field

$$\boxed{B = \frac{F}{qv}} \quad \text{for} \quad \underset{\sim}{v} \perp \underset{\sim}{B} \qquad \text{(magnetic flux density defined)}$$

where $\underset{\sim}{F}$ is the force acting on a test charge q moving with velocity $\underset{\sim}{v}$ perpendicular to the magnetic field $\underset{\sim}{B}$.

Note that $\underset{\sim}{B}$ is called the magnetic flux density; the name 'magnetic field strength' is given to another quantity which we do not consider in this course.

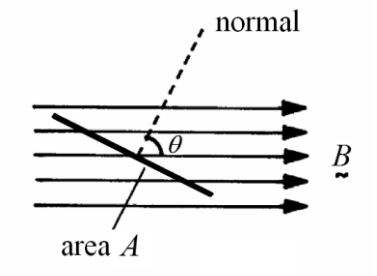
- The S.I. unit of magnetic flux density is $N.C^{-1}.m^{-1}.s = N.A^{-1}.m^{-1}$ which is defined as the **tesla** T.
- An older unit which is still frequently used is the **gauss** G: $1\ T = 10^4$ G.

Conventional permanent magnets can produce fields up to about 2 T; superconducting magnets have been constructed that create fields more than an order of magnitude larger. The field due to the Earth is about 0,5 G near the surface of the Earth.

Magnetic flux

Magnetic flux is defined in the same way as electrostatic flux. The flux ϕ_M through a small surface of area A perpendicular to a uniform magnetic field $\underset{\sim}{B}$ is $\phi_M = BA$.

The name *flux density* for B follows from the equation $B = \phi_M / A$.



If the normal to the surface makes an angle θ with $\underset{\sim}{B}$, as in the diagram to the left, then the magnetic flux is:

$$\boxed{\phi_M = BA\cos\theta} \qquad \text{(magnetic flux defined)}$$

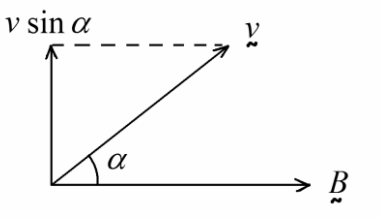
The unit of ϕ_M is the weber (Wb) which is equivalent to $T.m^2$.

Magnetic force on a moving charge

From the definition of flux density, the magnitude of the force on a charge moving perpendicular to a magnetic field is given by $F = qvB$.

More generally, if $\underset{\sim}{v}$ and $\underset{\sim}{B}$ make an angle α with each other, the component of the velocity perpendicular to the field must be used:

$$\boxed{F = qvB\sin\alpha}$$



- From experiment, the direction of $\underset{\sim}{F}$ is always perpendicular to the plane containing $\underset{\sim}{v}$ and $\underset{\sim}{B}$. In the diagram shown above $\underset{\sim}{F}$ is into the diagram for a positive charge and out of the diagram for a negative charge.

A single equation can be used to indicate both the magnitude and direction of the magnetic force:

$$\boxed{\underset{\sim}{F} = q(\underset{\sim}{v} \times \underset{\sim}{B})} \qquad \text{(force on a moving charge)}$$

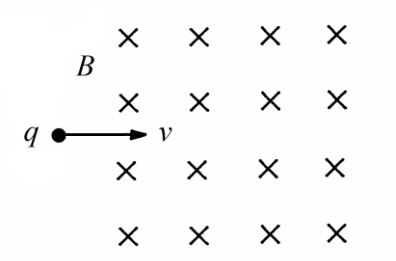
If the handle of a right-handed corkscrew is turned from the direction of $\underset{\sim}{v}$ to the direction of $\underset{\sim}{B}$ the corkscrew will move in the direction of $\underset{\sim}{F}$ for a positive charge q. The force on a negative charge is in the opposite direction.

- No force acts on a charge moving *parallel* to a magnetic field: if $\underset{\sim}{v}$ is parallel to $\underset{\sim}{B}$, then $\sin\alpha = 0$ and so $F = 0$.
- The force is a maximum for a charge moving *perpendicular* to the field (when $\alpha = 90°$).

Motion of a charge in a static, uniform magnetic field

Since the magnetic force on a charged particle is always perpendicular to the motion, no work is done on the particle by the magnetic force. It follows that the magnetic force cannot change the kinetic energy of the particle, nor its **speed**; the **direction** of the velocity will obviously be changed, with the particle being deflected sideways.

Consider a **positively**-charged particle entering a uniform magnetic field such that the direction of the particle's velocity is perpendicular to the field.



The symbol $\otimes$ represents a field pointing into the diagram (the symbol represents the tail of an arrow moving away from you).

Application of the corkscrew rule indicates that the direction of the magnetic force on a positive charge is **to its left** in this case.

This causes the particle's path to deviate to the left.

Therefore:

- The magnetic force on the particle is always perpendicular to its motion, to its left in this case.
- The magnitude of the force, $F = qvB$, is constant for a static, uniform field.

It follows that the path of the particle is circular, and it undergoes **uniform circular motion** (moving counterclockwise around the circle in the case illustrated in the diagram above). From Newton's second law:

$$F = qvB = m\frac{v^2}{r}$$

where m is the mass of the particle and r is the radius of the circular path. The radius of the path is therefore

$$r = \frac{mv}{qB}$$

- If the particle has a negative charge, the force is in the opposite direction and the motion is clockwise around a circular path for the situation described by the diagram above.
- If the initial direction of the particle is not perpendicular to the field, then the path followed by the particle is a helix around the magnetic field lines.

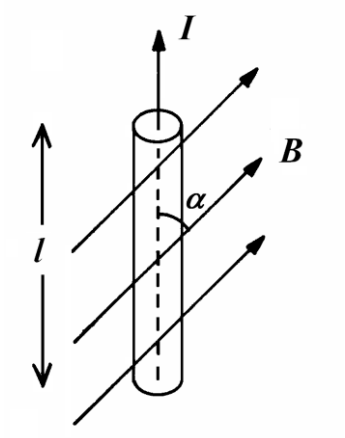
There are several applications of the effect described above:

- The mass spectrometer or spectrograph, a device to measure the mass (or charge to mass ratio) of a particle or ionised atom.
- Synchrotrons and cyclotrons, devices used to accelerate elementary particles to very large speeds for research purposes.

Force on current-carrying conductor in a magnetic field

If current flows in a conductor placed in a magnetic field, each charge carrier making up the current experiences a magnetic force; consequently, there is a force on any current-carrying conductor placed in a magnetic field. This effect was first demonstrated in 1820 by the Danish physicist Hans Christian Oersted (1777–1851).

Consider a length l of conductor that carries a current I and makes an angle α with a uniform external magnetic field of flux density B.



The charges that form the current are moving through the conductor with drift velocity v.

The total charge of the charge carriers in this length of conductor at any instant is Q.

All these charges will drift through the plane formed by the end of the conductor in a time t given by $t = l/v$.

The current in the conductor is therefore $I = Q/t$.

The average force on a single charge q moving with drift velocity v in the conductor is, from above,

$$F_q = qvB\sin\alpha .$$

The force on all the charge carriers in the conductor is therefore

$$F = QvB\sin\alpha$$
$$= (It)\left(\frac{l}{t}\right)B\sin\alpha$$

Therefore the force on the conductor is

$$\boxed{F = I\,l\,B\sin\alpha}$$

The force is at right angles to the plane containing l and B (this follows from the direction of the force on each charge carrier), so this can be written

$$\boxed{\underset{\sim}{F} = I(\underset{\sim}{l} \times \underset{\sim}{B})}$$

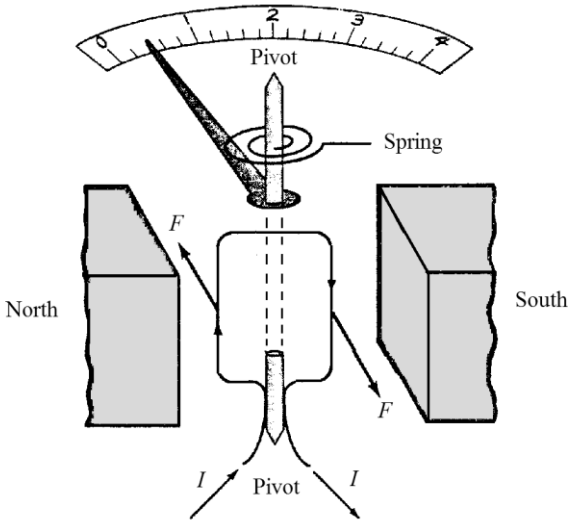
where $\underset{\sim}{l}$ is a vector of length l along the conductor in the direction of the current I.

- The direction of $\underset{\sim}{F}$ can be found from the corkscrew rule.
- If the current is in the direction of the field (or in the opposite direction), the magnetic force on the wire is zero.

The phenomenon described here is the principle of operation of the speakers found in most sound systems. It is also used in the design of electric motors and meters, where it produces a torque on a loop of wire.

The galvanometer

The force on a current-carrying conductor in a magnetic field can be used to rotate the conductor. In meters (and electric motors) a couple is exerted on a coil in the field. Electrical energy is thereby converted into mechanical energy, specifically kinetic energy of rotation.



The basic component of most meters, including voltmeters and ammeters, is the **galvanometer**, whose operation is illustrated in the diagram on the left.

A rectangular current loop is suspended in the uniform magnetic field between the poles of a permanent magnet.

In practice there will be many turns in the loop, rather than just one as shown here.

A current flows through the loop in the direction shown, and the function of the galvanometer is to measure this current.

- The force on the left side of the loop is into the diagram and the force on the right side of the loop is in the opposite direction, so that a couple is created that tries to rotate the loop about the central pivot.
- The magnitude of the forces and hence the couple is proportional to the magnitude of the current.

- This rotation is opposed by a spring that exerts a moment roughly proportional to the angle through which it is turned (Hooke's law).
- A pointer attached to the pivot moves across a scale in response to the rotation.
- Since the angle of rotation is proportional to the current, the deflection on the scale is also proportional to the current.

4.3.2. Ampere's Law and its Applications

In 1819 Hans Christian Oersted (1777–1851) discovered (accidentally, during a lecture demonstration at the University of Copenhagen) that a force is exerted on a magnet placed near a wire carrying an electric current, indicating the presence of a magnetic field around such a wire.

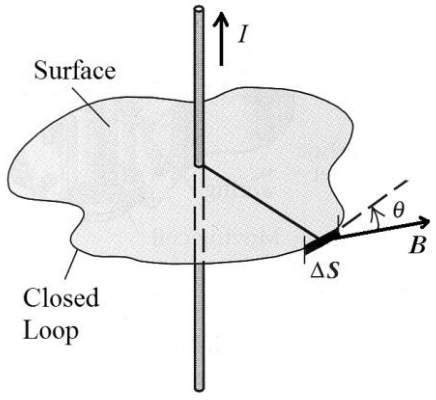
This was a most important discovery, since it showed for the first time that there is a connection between electricity and magnetism.

Ampere's law

Gauss's law is used in electrostatics to calculate electric fields. A similar type of calculation, due to the French scientist André-Marie Ampère (1775–1836), may be used in electromagnetism to find the magnetic flux density B in situations with sufficient symmetry.

To find the magnetic flux density due to a collection of conductors carrying steady currents, we proceed as follows.

- An arbitrary **closed loop** is drawn around the current-carrying conductors.



- The closed loop is divided into small elements of length Δs, and for each element we calculate the product $B_{\tan}\Delta s$ of Δs and the component of $\underset{\sim}{B}$ parallel or tangential to $\Delta \underset{\sim}{s}$. These terms are then summed around the entire loop.
- The total current I_{total} flowing through the surface bounded by the loop is calculated, taking into account the directions of the currents.

Ampere's law states that:

$$\boxed{\sum (B_{\tan}\ \Delta s) = \mu_0\, I_{\text{total}}} \qquad \text{(Ampere's law)}$$

The proportionality constant $\mu_0 = 4\pi \times 10^{-7}$ H.m^{-1} (H = henry) is analogous to ε_0, and is called the **permeability** of free space.

- The law is only useful for calculating magnetic fields due to highly symmetric current configurations. In this course we shall consider only cases in which we can draw a loop with $\underset{\sim}{B}$ parallel or perpendicular to the loop.

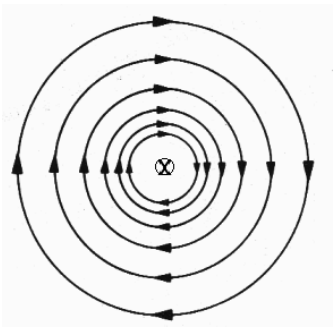
The uses of Gauss's and Ampere's laws to calculate electric and magnetic fields are contrasted in the table below.

	Gauss's law	Ampere's law
(i)	Draw a closed surface around the charges.	Draw a closed loop around the conductors (currents).
(ii)	Calculate the component of the electric field normal to the surface, E_{perp}.	Calculate the component of the magnetic field tangential to the loop, B_{tan}.
(iii)	Calculate the sum of the products $E_{perp}\Delta A$ for all elements ΔA of the surface.	Calculate the sum of the products $B_{tan}\Delta s$ for all elements Δs of the loop.
(iv)	Use the total charge enclosed by the surface.	Use the total current enclosed by the loop.

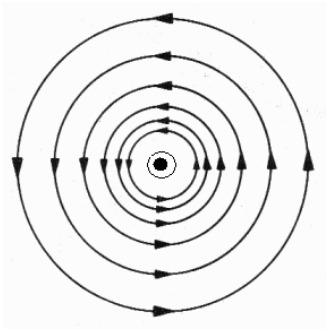
Flux density due to a long, straight conductor

Experiments such as those carried out by Oersted show that the magnetic field lines for a long, straight current-carrying conductor form a set of circles concentric with the conductor.

- By symmetry, the magnitude of $\underset{\sim}{B}$ must be the same everywhere on a circle centred on the conductor, so the field lines are also lines of constant B.
- The direction of $\underset{\sim}{B}$ is given by the corkscrew rule (or right-hand rule).
- The flux density decreases with distance from the conductor, so that the field lines become more widely spaced.

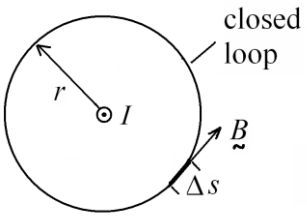


Current into diagram ⊗



Current out of diagram ⊙

To calculate the flux density a distance r from a long, straight conductor carrying a current I, we draw a closed circular loop of radius r around it.



$\underset{\sim}{B}$ is everywhere tangential to the loop and so $B_{\tan} = B$ for each element Δs.

The total current enclosed by the loop is I.

Therefore

$$\sum(B_{\tan}\,\Delta s) = B\sum\Delta s \quad (B \text{ constant on the circle})$$
$$= B\ 2\pi r \quad \left(\sum\Delta s = \text{circumference}\right)$$

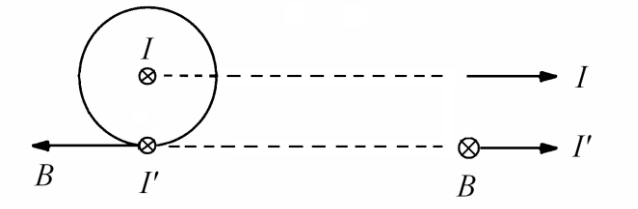
So, from Ampere's law, $B2\pi r = \mu_0 I$, giving

$$\boxed{B = \frac{\mu_0}{2\pi}\frac{I}{r}} \qquad \text{(long straight conductor)}$$

Force between parallel conductors

If two conductors both carrying currents are placed near each other, each is in the magnetic field due to the other and so each will experience a magnetic force due to the presence of the other. In other words, there will be a magnetic force between them. We consider only the case where the two conductors are parallel.

Two long, straight conductors are placed parallel to each other a distance r apart, and they carry currents I and I' in the same direction.



View from behind View from right side

In the diagram on the left, the two currents flow into the diagram, as indicated by the symbol ⊗. The circle of radius r centred on the upper conductor indicates the magnetic field B created by the current I, and the arrow labelled B shows the direction of this field at the position of the lower conductor.

In the diagram on the right, the same situation is viewed from a different angle. The two currents flow to the right. The direction of the field due to the upper conductor at the position of the lower conductor is indicated by the symbol ⊗ and is labelled B.

The flux density a distance r from upper wire is

$$B = \frac{\mu_0}{2\pi}\frac{I}{r}.$$

The field $\underset{\sim}{B}$ and lower conductor are perpendicular, giving $\alpha = 90°$. So the force on length l of lower conductor is

$$F = I'lB\sin\alpha = I'l\frac{\mu_0}{2\pi}\frac{I}{r}.$$

The force per unit length on the lower conductor is therefore:

$$\boxed{\frac{F}{l} = \frac{\mu_0}{2\pi}\frac{I\,I'}{r}} \qquad \text{(force between two currents)}$$

From Newton III, the force per unit length on the upper conductor has the same magnitude but opposite direction. The force on the lower conductor is upwards, that on the upper conductor is downwards.

- Thus, if the currents are parallel the conductors attract one another

If the two currents flow in opposite directions to each other, both forces are reversed in direction.

- If the currents are anti-parallel the conductors repel one another.

The ampere

The equation just derived is used in the definition of the ampere. Since $\mu_0 = 4\pi \times 10^{-7}$ H.m^{-1}, we have

$$\mu_0/2\pi = 2\times10^{-7}\,\text{H.m}^{-1}.$$

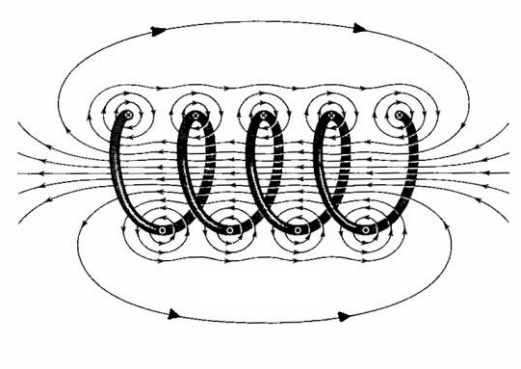
The definition of the **ampere** follows:

> If two long, parallel conductors, 1 m apart and carrying the same currents, each experiences a magnetic force of 2×10^{-7} N per metre length, the current flowing in each is defined to be 1 A.

All other electrical units are derived from this.

Flux density within a long solenoid

A solenoid is created by taking a long, straight wire (covered by an insulating layer) and bending it into a coil of many closely spaced loops.



The diagram shows the field due to a few loosely coiled loops of current-carrying wire.

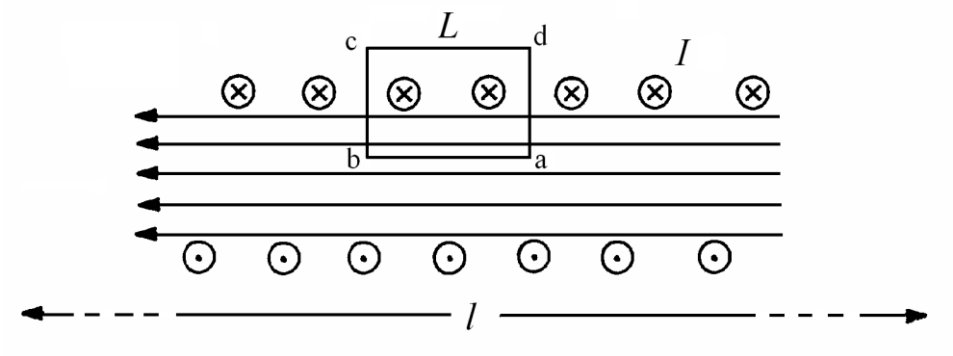
In a solenoid there is an extremely large number of current loops that are not widely spaced but closely-packed and overlapping.

If the solenoid is long enough compared with its diameter:

- The magnetic field lines inside the solenoid will be parallel to the axis, uniformly spaced and close together; the flux density *B* will be constant along these lines except near the ends of the solenoid.
- The flux density outside the solenoid will be very small in comparison with the field within the solenoid, and can usually be ignored (the same number of lines inside the solenoid is spread over all space outside).

The field due to the solenoid approximates that of a bar magnet; the end of the solenoid where the field lines emerge (on the left in the diagram above) acts as a north pole and the other end acts as a south pole. For this reason a solenoid is sometimes referred to as an electromagnet – it acts as a magnet only when it carries a current.

The diagram below shows a length-wise cut through a long solenoid, through which a current I flows (into the page for the turns at the top, and out of the page for the turns at the bottom). The loops are again shown widely spaced for clarity.



To calculate the flux density within the solenoid, we draw a rectangular closed loop abcd far from the ends of a solenoid, partly inside and partly outside the solenoid, as shown in the diagram. We apply Ampere's law to the closed loop abcd.

- Outside the solenoid $B = 0$, so that $B_{\tan} = 0$.
- Inside the solenoid $\underset{\sim}{B}$ is perpendicular to the lines ad and bc, so $B_{\tan} = 0$ at the sides of the loop.
- The only contribution is from the line ab, where $B_{\tan} = B$ since $\underset{\sim}{B}$ is parallel to ab. Then

$$\sum (B_{\tan}\,\Delta s) = B\sum \Delta s = BL$$

  where L is the length of the line ab
- The total length of the solenoid is l and it has N turns. Thus the number of turns per unit length is N/l, and in the length L there are $L(N/l)$ turns. The same current I flows through each turn; therefore the total current enclosed by the loop abcd is $LI(N/l)$.

From Ampere's law:

$$BL = \mu_0 LI\,\frac{N}{l},$$

giving

$$\boxed{B = \mu_0 \frac{N}{l} I} \qquad \text{(field inside a long solenoid)}$$

This equation does not hold near the ends of the solenoid.

Solenoids are used to generate magnetic fields in many practical devices, particularly where a uniform field is required:

- In TV sets horizontal and vertical electromagnets are used to control the direction of the electron beams that hits the screen;
- in the control of electric door locks in cars and in most types of doorbells;

- in the speakers of TV sets, tape players and so on;
- a solenoid of diameter about 1 m is used to create the magnetic field for magnetic imaging devices – the patient is placed inside the solenoid.

4.3.3. Electromagnetic Induction

Experiments carried out by Oersted, Ampere and others in 1819–1821 indicated two ways in which electricity and magnetism are related.

(i) An electric current produces a magnetic field.

(ii) A magnetic field exerts a force on an electric current and on a single moving charge.

Scientists then began to investigate if the reverse of the first effect was possible, i.e. whether a magnetic field could produce an electric current. That this was possible was discovered almost simultaneously in 1831 by the American scientist Joseph Henry (1797–1878) and the English scientist Michael Faraday (1791–1867).

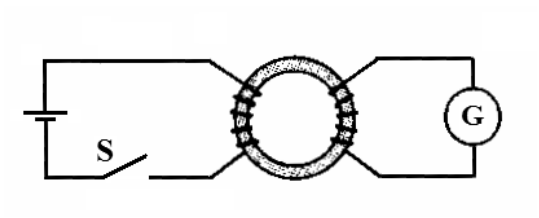
What Faraday and Henry discovered was that a **changing** magnetic field can induce an electrical current in a loop of wire; no current is produced by a steady magnetic field.

Modern devices that use electromagnetic induction for their operation include:

- Electric generators, such as the AC generators found in all power stations and also the more portable DC generators and alternators in cars.
- Transformers (discussed further below), tape and video recorders, computer disc drives (both removable discs and hard drives), and some types of microphones.
- The pick-up mechanism of an electric guitar, which converts the vibrations of a magnetised metal string into electrical oscillations that are sent to an amplifier.

Induced emf

The diagram below illustrates one of the experiments carried out by Faraday.



The *primary circuit* on the left contains a coil connected to a switch and a battery.

The coil is wrapped around an iron ring to intensify the magnetic field produced when a current flows in the primary coil. A magnetic field is created everywhere inside the ring.

The *secondary circuit*, on the right, contains only a coil wrapped around the same iron ring and a galvanometer to detect any current in the circuit.

Faraday discovered that:

- The galvanometer deflects strongly the instant the switch is closed.
- There is no deflection when the switch remains closed.
- The galvanometer deflects strongly in the opposite direction the instant the switch is opened again.

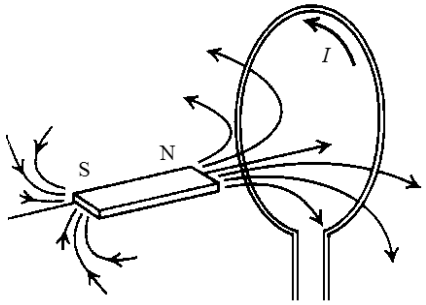
In other words, a steady current in the primary circuit produces no current in the secondary circuit. But when the current in the primary starts flowing or stops flowing, a current is briefly induced in the secondary circuit.

When the current in the primary circuit starts to flow or stops flowing, the magnetic field in the primary coil changes and therefore the magnetic flux through the secondary coil also changes.

Faraday concluded that although a steady magnetic field produces no effect,

- a changing magnetic field produces an induced current in the secondary circuit;
- the presence of the induced current implies the existence of an induced emf in the secondary coil.

The next diagram shows a further experiment conducted by Faraday to investigate electromagnetic induction.



A wire loop is connected to a galvanometer.

If a bar magnet is moved **towards** the loop, a current is induced in the loop in the direction shown.

If the magnet is held steady, no current flows.

If the magnet is moved **away from** the loop, a current flows through the loop in the opposite direction.

Exactly the same effect is observed if the magnet is held stationary and the coil is moved towards it or away from it. Again, it is the fact that the magnetic flux through the coil is changing that produces the induced current in the coil.

Faraday's law of electromagnetic induction and Lenz's law

From a careful study of these and similar experiments, Faraday concluded that the emf induced in a coil depends on the **rate** at which the **magnetic flux** through the coil is **changing**.

If a circuit contains N tightly wound loops and the magnetic flux through the surface bounded by each loop changes by an amount $\Delta\phi_M$ in a time interval Δt, the average emf induced during this time is given by

$$\boxed{\mathcal{E} = -N\frac{\Delta\phi_M}{\Delta t}} \qquad \text{(Faraday's law)}$$

By convention, a minus sign is included in Faraday's law to indicate the polarity of the induced emf, which can be found from Lenz's law as discussed below.

The magnetic flux through a loop is defined as $\phi_M = BA\cos\theta$ where θ is the angle between the field and the normal to the loop. So, from Faraday's law, an emf is produced if any of the factors B, A and θ changes with time.

In applying the laws of Faraday and Lenz, remember the following:

- A changing magnetic field will produce an **induced emf** in a coil.

- This creates an **induced current** in the coil. The direction of the induced current is determined by the polarity of the induced emf.
- The induced current produces an **induced magnetic field** (which should not be confused with the changing magnetic field that created it). The direction of the induced field and induced current are related by the right-hand screw rule.

The **polarity of the induced emf** and, consequently the direction of the induced current and magnetic field, are given by **Lenz's law**. This is a consequence of energy conservation, and was proposed by the Russian scientist Heinrich Lenz (1804–1865) soon after Faraday formulated his law.

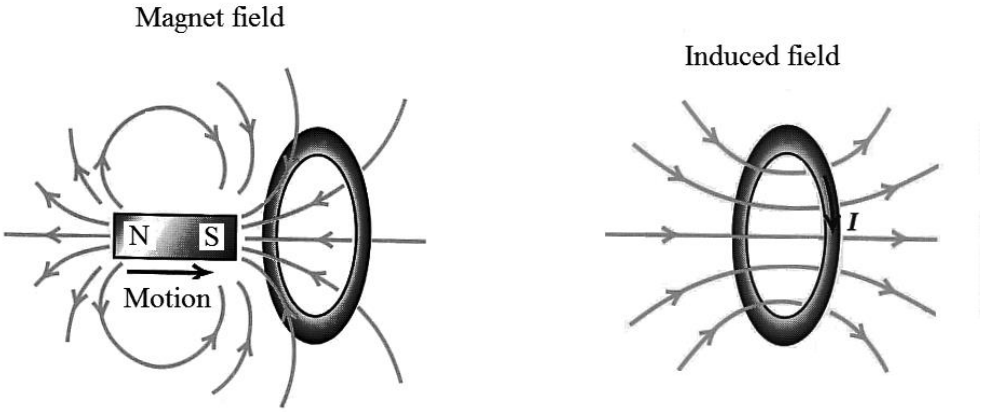
> The current caused by the induced emf travels in the direction that creates a magnetic field with flux opposing the change in the original flux through the circuit.

In other words, the induced current attempts to maintain the original flux through the circuit.

- If the flux is increasing, the induced current will produce a magnetic field which opposes the increasing applied field (i.e. the induced field is in the opposite direction to the applied field).
- If the flux is decreasing the field produced by the induced current reinforces the decreasing applied field (i.e. the induced field is in the same direction as the applied field).

Lenz's law can be used to find the polarity of the induced emf even if no current flows (because the circuit is not closed) by finding the direction it would flow if the circuit were complete.

Lenz's law can be illustrated by its application to the experiment of Faraday described above.



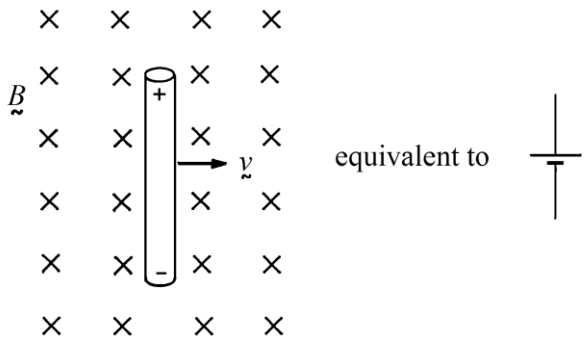
- As the south pole of the magnet gets closer to the coil, the flux through the coil increases, inducing a current in the coil.
- By Lenz's law, the induced current must produce a field that opposes the increase in the flux. Since the field due to the magnet is **increasing**, this requires that the induced field through the loop is in the **opposite** direction to the magnet's field (i.e. from left to right in the diagram).
- Application of the corkscrew rule then says that in order to produce such a field, the current in the coil must be clockwise, as shown in the diagram.

Motional emf

We describe in this section a particular example of electromagnetic induction in which a **motional emf** is produced; this is an emf induced in a conductor that is moving through a magnetic field.

To see the origin of the emf, consider a straight conductor of length l moving with constant velocity $\underset{\sim}{v}$ through a uniform magnetic field $\underset{\sim}{B}$.

The field is into the diagram below, and the motion is perpendicular to the field.



- A magnetic force $F_M = qvB$ acts on free electrons in the conductor; these move downwards and accumulate at the lower end of the conductor. The upper end of the conductor becomes positively charged and the lower end negatively charged.
- As a result of this charge separation an electric field is set up in the conductor, directed from positive to negative, i.e. downwards in the diagram above. Electrons in the conductor will therefore experience an electrical force $F_E = qE$ which is in the opposite direction to the magnetic force. As the separation of charges continues the magnitude of the electrical field will increase until the electrical and magnetic forces on an individual charge balance.
- If the ends of the rod were connected to an external circuit (at least partly outside the field), a current would flow in the external circuit from the positively charged end of the rod to the negative end. The system therefore acts as a source of emf, and we say that a **motional emf** is being induced in the moving conductor. The polarity of the induced emf is as shown on the right of the diagram.

We can derive an expression for the magnitude of the induced emf as follows. As indicated, the flow of charge in the conductor stops when the magnetic and electric forces on a charge are equal, $qE = qvB$, i.e. when the magnitude of the induced electric field in the conductor is

$$E = vB$$

Because the electric field in the conductor is uniform, the magnitude of the field is related to the p.d. across the ends of the conductor by

$$\Delta V = El = vBl$$

where l is the length of the conductor. But the potential difference ΔV across the terminals of a source when no current is drawn is just the emf of the source, so that

$$\boxed{\mathcal{E} = Blv} \qquad \text{(motional emf)}$$

- If the conductor is not moving at right angles to $\underset{\sim}{B}$, we must take the component of $\underset{\sim}{v}$ perpendicular to $\underset{\sim}{B}$ (or vice versa).

We have found the magnitude of the motional emf without using Faraday's law. This was possible because in the situation described here the physical cause of the induced emf is clear. Use of Faraday's law would lead to the same formula for the emf.

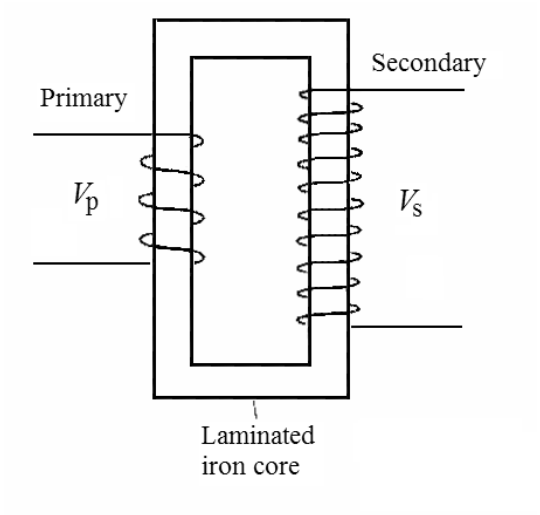
The polarity of the induced emf as deduced above is also consistent with the prediction of Lenz's law.

The transformer

A transformer is a device for increasing or decreasing AC voltage. Transformers are found

- in TV sets to give the high voltage needed for the picture tube;
- in devices for connecting to the mains portable radios, electric razors etc as well as rechargeable devices such as cell phones (these devices may also contain special circuitry to convert AC to DC);
- in power stations, to convert the output of the station to very high voltages.

The transformer in its simplest form consists of two coils of wire, called the primary and the secondary, linked by a soft-iron laminated core.



Transformers are designed so that nearly all the magnetic flux produced by a current in the primary also passes through the secondary, via the iron core.

An AC voltage V_P is applied to the primary. The changing magnetic field that this produces induces an AC voltage V_S of the same frequency in the secondary coil.

Using Faraday's law, it can be shown that these voltages are related by

$$\boxed{\frac{V_S}{V_P} = \frac{N_S}{N_P}} \qquad \text{(transformer equation)}$$

where N_S and N_P are the number of turns in the coils in the secondary and primary circuits.

- For a step-up transformer, $N_S > N_P$.
- For a step-down transformer, $N_S < N_P$.

However the conservation of energy requires that the power output cannot exceed the power input. Energy losses due to resistance in the coils and losses in the iron core are usually extremely small (typically only a few percent) and will be ignored.

From $P = IV$, we have $V_P I_P = V_S I_S$ or

$$\frac{I_S}{I_P} = \frac{V_P}{V_S} = \frac{N_P}{N_S}$$

So for example, if we increase the voltage by a factor of 10, the current will decrease by at least the same factor.

The role of transformers in the transmission of electricity is extremely important.

- The power stations that produce our electricity are usually situated far from the consumers of the electricity, so that electricity needs to be transmitted over long distances. This requires transmission at very high voltages (and therefore very low currents) in order to reduce power losses (which depend on I^2R). Transmission at 132 kV is the standard in some countries, although voltages that are much larger are also used. Power is generated at the power station at voltages much lower than this, so that step-up transformers are an integral part of all power stations.
- Low voltages are more practical in the home because circuits are more easily insulated against breakdown (and low voltages are intrinsically safer). The power supplied to homes is therefore AC at 230 V, so the voltage must be reduced using step-down transformers at local sub-stations.
- As already indicated, most domestic appliances operate at even lower voltages, and therefore have step-down transformers built into their design.

ELECTRICITY AND MAGNETISM

LECTURE EXAMPLES

University of the Witwatersrand, Johannesburg

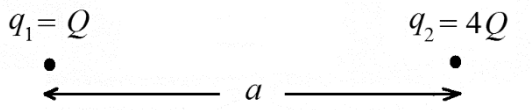
School of Physics

PHYS1001/6 (Physics I D)

Lecture Examples 4.1 – [2018] **Electrostatics**

Question 1

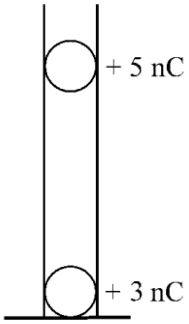
Two charges, $q_1 = Q$ and $q_2 = 4Q$, are separated by a distance a. A third charge is placed on the line joining q_1 and q_2.



Where must this third charge be placed so that the net force exerted on it is zero? [a/3 from q]

Question 2

A sphere of mass 10^{-3} kg which carries a charge of +5,0 nC is released from rest from a small distance above a fixed charge of +3,0 nC. Calculate the equilibrium separation of the charges.



[3,67 mm]

Question 3

Water drops of mass $4{,}0 \times 10^{-14}$ kg, each carrying 2000 excess electrons, are sprayed parallel to the surface of a horizontal metal plate.

(i) What surface charge density must the plate carry if the path of the drops is to remain horizontal? Hint: consider the forces that act on the drop; ignore the upthrust due to the air.

(ii) If the charge density on the plate were doubled, what would be the upward acceleration of the drops?

[$-1{,}11\times10^{-8}$ C.m^{-2}; 10 m.s^{-2}]

Question 4

A water droplet of radius $r = 0{,}050$ mm carries a charge q such that the electric field E' at its surface is $6{,}0 \times 10^4$ V.m^{-1}. If it is placed between two parallel metal plates a distance $d = 10$ mm apart, what potential difference V must be applied to the plates to keep the drop from falling?

Hint: solve this problem by carrying out the following steps, working with symbols rather than numbers until the last step.

(i) Find an expression for the charge q on the drop in terms of the field E' on its surface.

(ii) Hence determine an expression for the electric force on the charge when placed between the plates, in terms of the V and d.

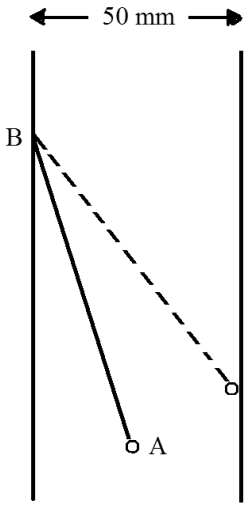
(iii) Now calculate the value of V by considering all the forces acting on the droplet.

Density of water = 1000 kg.m^{-3}.

[3,14 kV]

Question 5

A sphere of mass 0,10 g carrying a charge of +20 nC is suspended from one of a pair of vertical parallel plates 50 mm apart by an insulating thread AB which is 120 mm long.

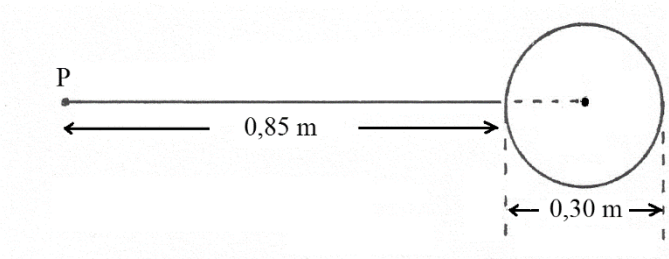


If the potential difference between the plates is gradually increased, calculate the value when the sphere just touches the negative plate. Hint: draw a free-body diagram for the sphere when it is just touching the plate, and then use the fact that the sphere is then in equilibrium. [1150 V]

Question 6

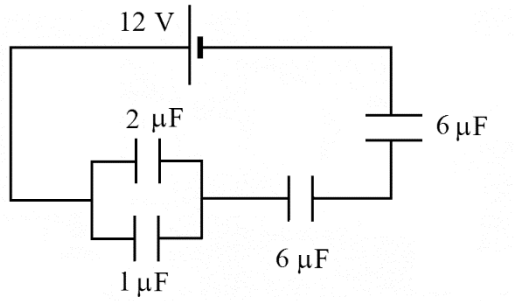
A square ABCD has sides 0,20 m long, and has charges +10 nC at A and −10 nC at C. Calculate the magnitudes of the electric field and potential at D. [3,18 kV.m^{-1}; 0 V]

Question 7



The copper sphere shown above carries a charge of +5,0 μC. How much work must be done to bring an additional charge of +10,0 nC from point P to the surface of the sphere? Hint: first write down expressions for the potentials at P and on the surface of the charge, and then use the fact that the work done is the product of the charge and the potential difference. [2,55 mJ]

Question 8



Calculate the charge on the 2 μC capacitor in the circuit above. Hint: solve this problem by carrying out the following steps.

(i) Calculate the equivalent capacitance of the circuit. [1,5 μF]

(ii) What is the charge on the equivalent capacitor? [18 μC]

(iii) Find the p.d. across the capacitors in parallel and hence calculate the charge on the 2 μC capacitor. [6 V; 12 μC]

Question 9

A 2 μF capacitor is charged by a 12 V battery. It is then disconnected from the battery and its terminals are connected to those of a 6 μF capacitor which is initially uncharged.

(i) What was the charge on the 2 μF capacitor before it was disconnected from the battery? [24 μC]

(ii) Calculate the final charge on each capacitor and the potential difference across each capacitor. [6 μC; 18 μC; 3 V]

(iii) Calculate the final energy stored in each capacitor. Compare the total energy stored in the capacitors with the energy originally stored in the 2 μF capacitor and account for any difference. [9 μJ; 27 μJ; 144 μJ]

Question 10

A parallel-plate capacitor is charged in air. It is then electrically isolated and lowered into a liquid dielectric. As a result:

(a) both the capacitance and potential difference between the plates decrease;

(b) both the capacitance and potential difference between the plates increase;

(c) the capacitance increases and the potential difference between the plates decreases;

(d) both the capacitance and the charge on the plates decrease;

(e) both the capacitance and the charge on the plates increase.

Which statement is correct? [(c)]

Question 11

Two identical parallel-plate capacitors of 2 nF each with air between their plates are connected in parallel and charged with a 10 V battery. The battery is then disconnected and the space between the plates of one of the capacitors is filled with a material having a dielectric constant $\kappa = 4$. Calculate:

(i) the initial charge on each capacitor, before the battery is disconnected, [20 nC]

(ii) the new charge on the capacitor filled with the dielectric, [32 nC]

(iii) the new voltage across each capacitor, [4 V]

(iv) the energy of the system before and after the dielectric is inserted. [200 nJ; 80 nJ]

University of the Witwatersrand, Johannesburg

School of Physics

PHYS1001/6 (Physics I D)

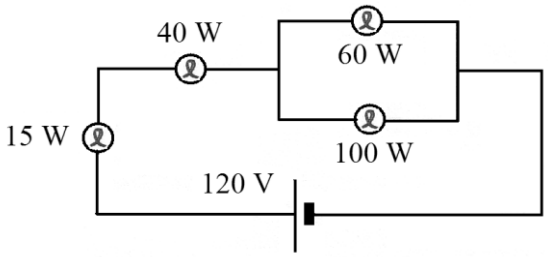
Lecture Example 4.2 – [2018] **Current Electricity**

Question 1

A wire of resistance *R* is drawn out so that its length is twice the original length. If the resistivity and density of the wire do not change in the process, what is the new resistance of the wire? [4R]

Question 2

Light bulbs with the wattage ratings shown are connected as in the diagram below. The wattage rating is the power that would be dissipated if each bulb were connected on its own across the 120-V supply.



Calculate:

(i) the current delivered to the circuit by the battery,

(ii) the power dissipated in each bulb, and

(iii) the total power delivered by the battery.

Hint: first calculate the resistance of each light bulb, and then calculate the equivalent resistance of all the bulbs. Assume that the resistance of the bulbs is constant for any voltage.

[85,1 mA; 6,95 W; 2,61 W; 0,244 W; 0,407 W; 10,2 W]

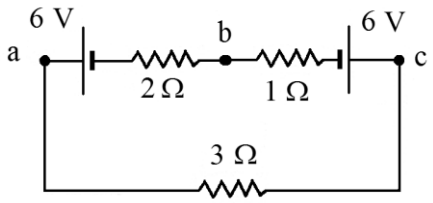
Question 3

A direct current is passed through a human limb in such a way that layers of skin, fat and muscle form three resistances in series. Calculate the relative rate of heat dissipation in each layer. [500: 1,5: 1]

The thickness of each layer and its resistivity is given in the table below. Assume that each layer has the same cross sectional area.

	Thickness (mm)	Resistivity (Ω.m)
Skin	0,01	10^6
Fat	2	15
muscle	10	2

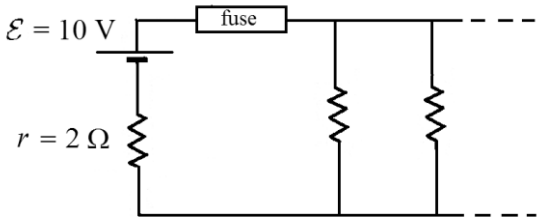
Question 4



For the circuit above calculate (i) V_{ab}, (ii) V_{bc} and (iii) V_{ac}. Hint: what is the current through the circuit? [+6 V; –6 V; 0 V]

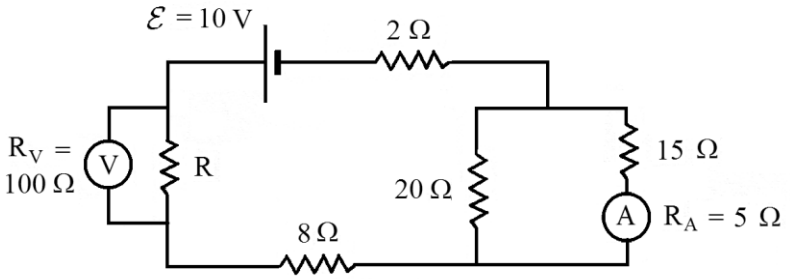
Question 5

A number of identical lamps rated 20 W, 10 V are connected in parallel with a source of emf 10 V and internal resistance 2 Ω. Assume that the resistance of the lamps is independent of temperature.



How many of them can be connected without blowing a fuse rated at 3 A connected in series with the source? Note: the fuse will blow when the current through it exceeds its rating. What does this tell you about the equivalent resistance of the lamps connected in parallel? [3]

Question 6



Calculate:

(i) the value of *R* if the reading on the ammeter is 50 mA, and

(ii) the reading on the voltmeter. [400 Ω; 8 V]

Hint; for part (i), first calculate the equivalent resistance of the parallel arrangement that includes the ammeter. You should then be able to use the circuit equation to find the equivalent resistance of the voltmeter and the resistance *R*, which are connected in parallel.

Question 7

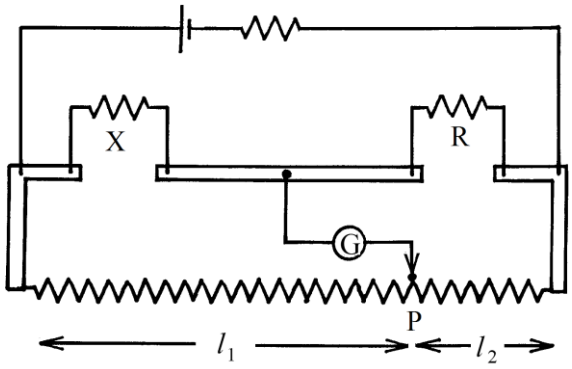
The resistance of a resistor is measured using a voltmeter and an ammeter. When the voltmeter is connected directly across the resistor, the readings obtained are 50 V and 0,55 A. When the voltmeter is connected across both the ammeter and the resistor, the readings are 54,3 V and 0,54 A. The resistance of the voltmeter is 1000 Ω.

Calculate the resistances of the resistor and ammeter. [100 Ω; 0,556 Ω]

Hint: draw the segment of the circuit containing the voltmeter, ammeter and resistor for each of the two cases. From the first you should be able to determine the resistance of the resistor.

Question 8

When a length of nichrome wire of resistance X is placed in the bridge circuit shown, the balance point is at P. The wire is now replaced by another piece of nichrome wire of the same length but twice the diameter.



Which one of the following statements is true? [d]

(a) The balance point will change because the current in the bridge circuit changes.

(b) The balance point will be the same because the bridge measures resistivity.

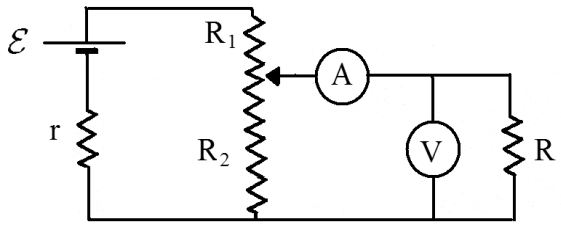
(c) The balance point will be the same because the length of nichrome wire is the same.

(d) The balance point will move to the left.

(e) The balance point will move to the right.

Question 9

In the circuit below $R_1 = 10\ \Omega$, $R_2 = 20\ \Omega$, the resistance of the voltmeter is 100 Ω, the resistance of the ammeter is 5 Ω and r, the internal resistance of the source, is 2 Ω.



(i) When the voltmeter reads 5 V the ammeter reads 0,25 A. What is the true resistance of the resistor R? [25 Ω]

(ii) What is the emf of the source and what is the current drawn from it?
Hint: from your answer to the first part you should be able to determine the p.d. across R_2 and hence the current through R_2. [13,0 V: 0,563 A]

University of the Witwatersrand, Johannesburg

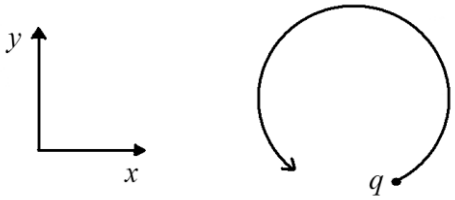
School of Physics

PHYS1001/6 (Physics I D)

Lecture Example 4.3 – [2018] **Electromagnetism**

Question 1

A particle of charge q enters a magnetic field **B** and moves counterclockwise in a circle in the plane of the paper.



Which one of the following statements is true? [e]

(a) q is positive and **B** is in the $+x$ direction.

(b) q is negative and **B** is in the $+y$ direction.

(c) q is positive and **B** is out of the plane of the paper.

(d) q is negative and **B** is into the plane of the paper.

(e) q is negative and **B** is out of the plane of the paper.

Question 2

A particle with mass and charge numerically equal to those of the electron enters a cloud chamber within which there is a constant magnetic field of flux density $9{,}0 \times 10^{-4}$ T. The particle enters at right angles to the field, and travels in a circular path of radius 0,20 m in a counter-clockwise direction (viewed in the direction of the field).

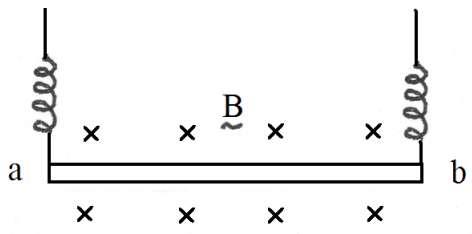
(i) Is the particle an electron or a positron (the antiparticle of the electron, with the same mass but positive charge)? Explain. [positron]

(ii) Calculate the speed of the particle. [$3{,}16 \times 10^{7}$ m·s^{-1}]

(iii) Through what potential difference must the particle have been accelerated to acquire this speed? Hint: use the conservation of energy. [2,85 kV]

Question 3

A wire ab of length 0,50 m and mass 0,010 kg is suspended by a pair of flexible leads in a magnetic field of flux density 0,40 T as shown in the figure.



What are the magnitude and direction of the current required to remove the tension in the supporting leads? Hint: consider the forces acting on the wire. [0,50 A from *a* to *b*]

Question 4

Two long, parallel vertical wires 0,30 m apart are placed east-west of each other. The current in the easterly one is 30 A and that in the other is 20 A. Both currents flow upwards.

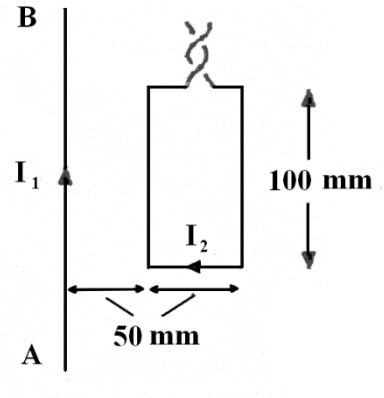
The earth's magnetic field is horizontal, directed north, and has flux density 20 μT.

Calculate the resultant force per metre on each wire.

Hint: to find the resultant force on a wire you need to find the resultant field at the position of the wire; this is the resultant of the field due to the other wire and the earth's field (make sure you take into account the direction of each). [$1{,}0 \times 10^{-3}$ N; 0 N]

Question 5

The long straight wire AB carries a current $I_1 = 20\,\text{A}$. A current $I_2 = 10\,\text{A}$ flows through a rectangular loop whose long edges are parallel to the wire AB.



Calculate the magnitude and direction of the force exerted on the loop by the magnetic field due to the current in AB. Note: the forces on the sections of the loop perpendicular to AB cancel.

[40 μN towards AB]

Question 6

Electrons are travelling in a circular path of radius 1,0 mm at the centre of a long solenoid in which a current of 1,0 A is flowing. The solenoid is 1,0 m long and has a total of 10^4 turns.

(i) What is the magnitude of the magnetic field inside the solenoid? [12,6 mT]

(ii) What electric field, and in what direction relative to the magnetic field due to the solenoid, must be applied to make the electrons travel in a straight line?

Hint: if an electron travels in a straight line, what can you deduce about the electric and magnetic forces acting on the electron? [27,8 $\text{kV}\cdot\text{m}^{-1}$ perpendicular to B]

Question 7

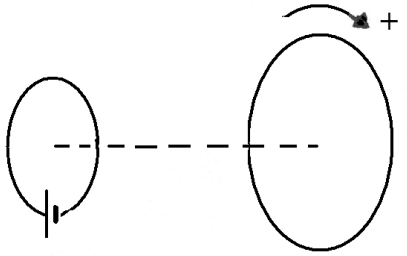
A wire loop with radius 0,10 m and resistance 2,0 Ω is placed inside a long solenoid with the plane of the loop perpendicular to the axis of the solenoid. The solenoid has a radius of 0,15 m and has 2500 turns per metre.

A current is switched on in the solenoid and reaches its maximum value in 0,010 s. If a current of 4,0 mA is induced in the loop, what is the maximum value of the current in the solenoid?

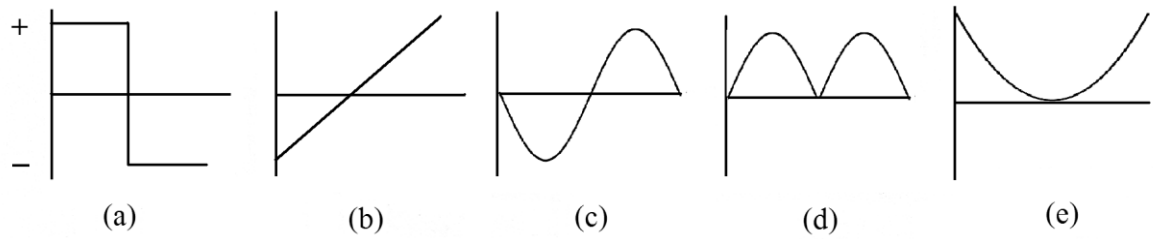
Hint: the easiest way to solve this problem is to work backwards from the induced current. You should be able to calculate the emf induced in the loop and then, using Faraday's law, the maximum field produced by the solenoid. [0,811 A]

Question 8

The smaller loop in the diagram below is nearer to you, and initially far from the larger loop. It moves away from you towards the larger loop, passes through it and continues on the far side.



Which graph below best shows the variation in the current induced in the larger loop? Assume clockwise to be positive when viewed from your position. [c]



Question 9

A skate-boarder decides to speed himself up with an electric motor, driven by the emf induced as one axle of the skate-board moves through the earth's magnetic field.

The horizontal axle is 0,10 m long, and it moves at 5,0 $m.s^{-1}$ through a field of flux density 30 μT inclined at 30° to the horizontal.

What power will be produced if the total resistance of the circuit is 10 Ω? Do you think the plan is feasible? Hint: first calculate the emf induced in the skate-board. [5,6 pW]

PHYS 1001/1006 TUTORIALS
YEAR 2018
3rd BLOCK

TUTORIALS TO PREPARE

WEEK: 16 July:	**No Tutorials**	
WEEK: 23 July:	**Electrostatics:**	**1, 2, 3**
WEEK: 30 July:	**Electrostatics:**	**5, 6, 8**
WEEK: 06 August:	**Current Electricity:**	**1, 2, 3**
WEEK: 13 August:	**Current Electricity:**	**4, 5, 6**
WEEK: 20 August:	**Electromagnetism:**	**1, 2, 3, 4**
WEEK: 27 August:	**Electromagnetism:**	**5, 6, 7**

- All students are expected to prepare solutions to these questions listed above prior to attending the tutorial session. Tutors will do a quick check. Preparation of tutorials will be recorded in the class registers as Not Prepared (NP), Partially Prepared (PP) and Well Prepared (WP).

- Students should form mini-groups of 3 and discussion within the groups should commence immediately on arrival at the tutorial venue. This discussion should be for a minimum of 20 minutes. Tutors will float around groups and assist students and will only give feedback on questions to the entire class that he/she might find as problematic. Note tutors are expected to give an overview of the solution and not complete solutions. Any problems not addressed will be carried over to the next week for discussion.

- A tutorial test will then be given at the end of each session (approximately 10 minutes). Tutors are expected to give feedback on the tutorial test at the beginning of the next session.

PHYS1001/1006 Course Co-ordinator

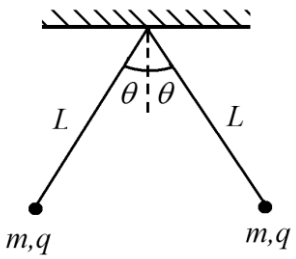
University of the Witwatersrand, Johannesburg

School of Physics

PHYS1001/6 (Physics I D)

Tutorial 4.1 – [2018] **Electrostatics**

Question 1



Two conducting spheres, each of mass m = 1,0 g, carry equal charges q.

They are suspended from the same point by light strings each of length L = 100 mm.

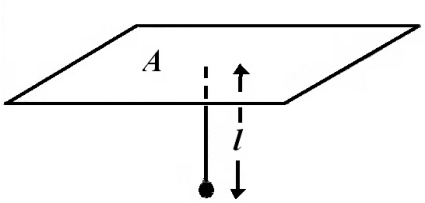
If the angle between the strings is 60°, what is the magnitude of the charge q on each sphere?

[80 nC]

Question 2

A small sphere of mass M carrying m electrons is suspended by a thread of length l from a plate of area A carrying n electrons uniformly distributed over its lower surface.

Derive an expression for the tension in the thread.



Note that for this question 3 and 4 the gravitational force on the electron is sufficiently small compared with the electric force that it can be ignored

Question 3

Two parallel conducting plates are 50 mm apart. An electron is fired with an initial velocity of 10^3 m.s^{-1} from the positive plate directly towards the negative plate.

What must the charge density on the plates be if the electron's distance of closest approach to the negative plate is 5 mm?

Charge of electron $-e = -1{,}60 \times 10^{-19}$ C; mass of electron $m = 9{,}11 \times 10^{-31}$ kg.

[5.59×10^{-16} C.m^{-2}]

Question 4

An electron is released from rest from the negative of two oppositely-charged parallel plates, and strikes the positive plate, 20 mm away, after 20 ns.

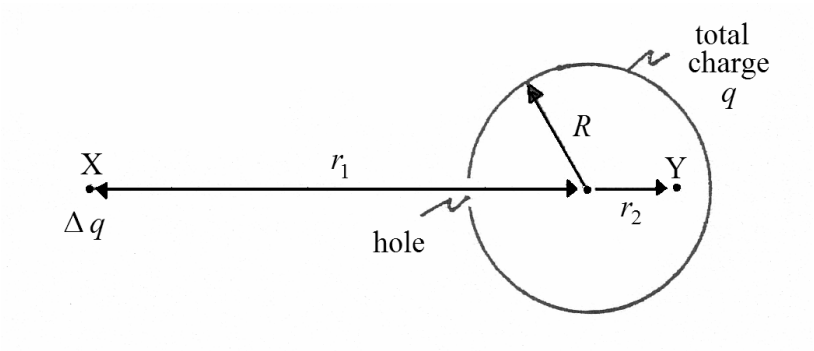
Calculate:

(i) the velocity of the electron when it strikes the positive plate,
(ii) the electric field strength between the plates, and
(iii) the p.d. between the plates.

[(i) 2×10^6 m.s^{-1} (ii) 569 V.m^{-1}, (iii) 11.4 V]

Question 5

The diagram represents a hollow metal sphere of radius R carrying a total positive charge q. A small charged body with a positive charge Δq is taken from point X to point Y through a very small hole in the sphere.



Derive an expression for the work that must be done against the electric field.

Question 6

A parallel-plate capacitor is charged by connecting a battery across it. The electric field between the plates is E and the charge on the capacitor plates is q.

Determine what happens to the magnitudes of E and q as the separation between the plates is decreased while the capacitor remains connected to the battery. Do they increase, decrease or stay the same?

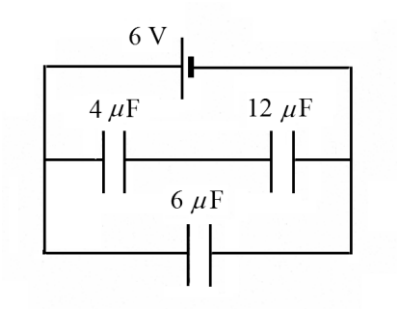
Question 7

A 2 μF capacitor is charged to a potential difference of 200 V and then isolated. When it is connected in parallel with a second capacitor which is initially uncharged, the common potential difference becomes 40 V.

What is the capacitance of the second capacitor?

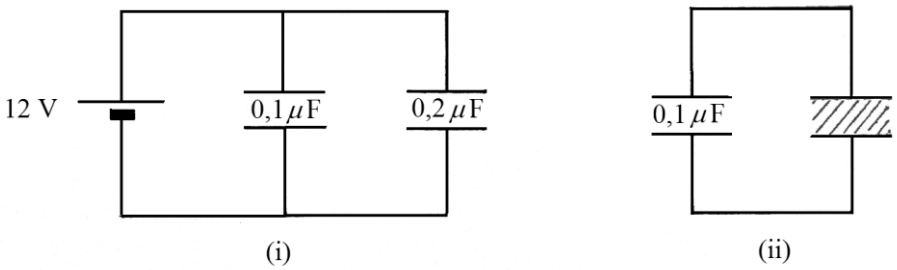
Question 8

A battery is used to supply a potential difference of 6 V to the circuit shown below. Calculate the energy stored in the 4 μF capacitor.



Note the symbol used in circuit diagrams to indicate a battery; the longer line represents the positive terminal.

Question 9



(i) Two capacitors of capacitances 0,1 μF and a 0,2 μF are connected to a 12 V battery as shown in the diagram on the left. Calculate the energy stored in the 0,2 μF capacitor.

(ii) The battery is disconnected and then a dielectric of dielectric constant $\kappa = 4$ is inserted into the 0,2 μF capacitor, as shown in the diagram on the right. Calculate the energy now stored in this capacitor.
Explain whether you would do work or have work done on you in inserting the dielectric.

[(i) 14.4 μJ, (ii) 6.4 μJ, work done on you]

University of the Witwatersrand, Johannesburg

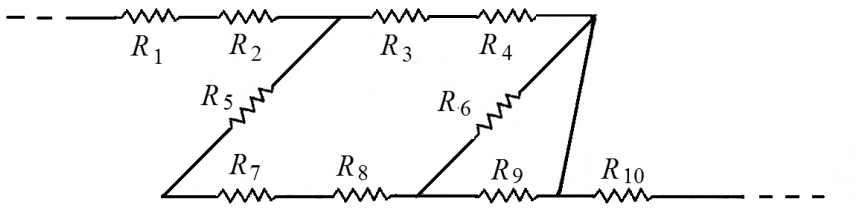
School of Physics

PHYS1001/6 (Physics I D)

Tutorial 4.2 – [2018] **Current Electricity**

Question 1

The diagram below shows part of a circuit. Which pairs of resistors are in series and which pairs are in parallel?



Question 2

A set of bulbs are each rated at 20 V, 10 W. (This means that if the p.d. across a bulb is 20 V, it consumes 10 W.)

(i) If the minimum voltage at which they will glow is 10 V, what maximum number can be connected in series across a source of emf 250 V and internal resistance 400 Ω and still glow?

(ii) What maximum number can be connected in parallel across a source of emf 20 V and internal resistance 2 Ω without blowing a 2 A fuse in series with the source?

Assume the resistance of the bulbs does not change with temperature.

[(i) 15, (ii) 5]

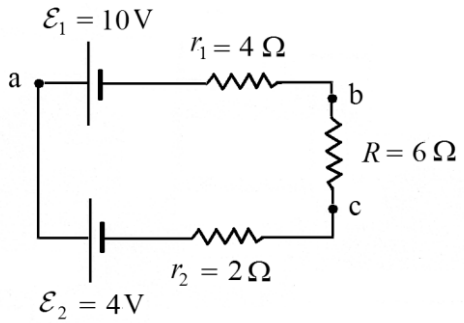
Question 3

For the circuit below, calculate:

(i) the potential differences V_{ab}, V_{ac} and V_{bc};

(ii) the power delivered by or stored in each cell, and the power dissipated in each resistor;

(iii) the total energy delivered to the circuit in 1 minute.

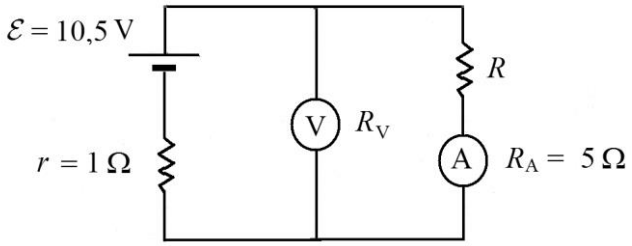


[(i) +8 V, +5 V, -3V; (ii) 5 W (delivered by 10 V cell), 2 W (stored in 4 V cell), 1 W (4 Ω), 0.5 W (2 Ω), 1.5 W (6 Ω); (iii) 300 J

Question 4

An unknown resistance is to be measured using a cell, voltmeter and ammeter, connected as shown below. The two meters are **not** ideal, since the voltmeter resistance is not large and the ammeter resistance is not small.

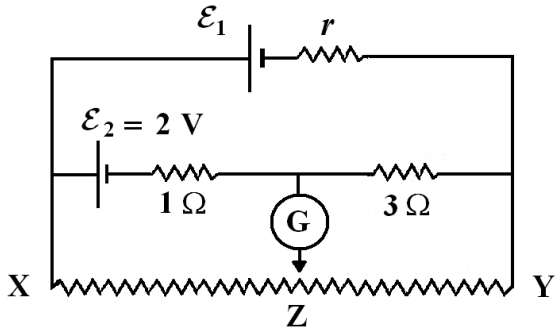
If the voltmeter reads 10 V when the ammeter reads 0,4 A, calculate the values of the resistances R and R_V. Compare the computed value of R with the ratio of the readings on the voltmeter and ammeter.



[20 Ω , 100 Ω]

Question 5

In the circuit below XY is a slidewire of total resistance 5 Ω. The emf $\mathcal{E}_1$ of the battery is greater than 2 V, and the internal resistance of the battery of emf $\mathcal{E}_2$ is zero. With the sliding contact Z at the centre of the slidewire no current flows through the galvanometer.



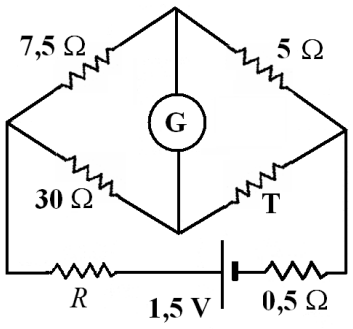
Calculate the currents through the 1 Ω and 3 Ω resistances and through the slidewire under these conditions.

[1 A , 1.2 A]

Question 6

The circuit shown on the right can be used to measure temperature by measuring the resistance of a thermistor T.

What resistance *R* must be included in the circuit if the current through the thermistor at balance is to be 10 mA?



[19.5 Ω]

University of the Witwatersrand, Johannesburg

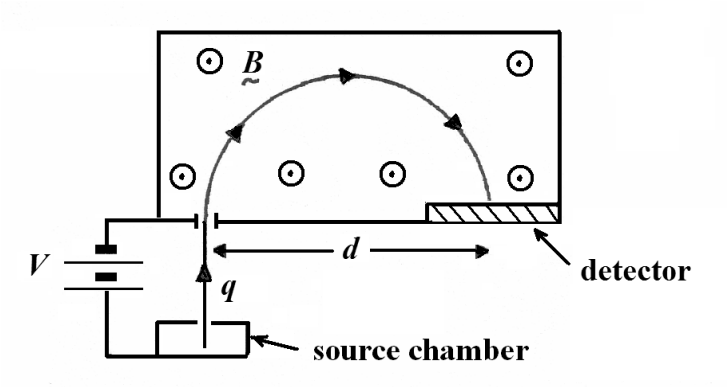
School of Physics

PHYS1001/6 (Physics I D)

Tutorial 4.3 – [2018] **Electromagnetism**

Question 1

The diagram below shows a simple mass spectrometer, used to measure the charge to mass ratio of ions.



Ions of mass *m* and charge *q* are produced with negligible initial velocity in the source chamber.

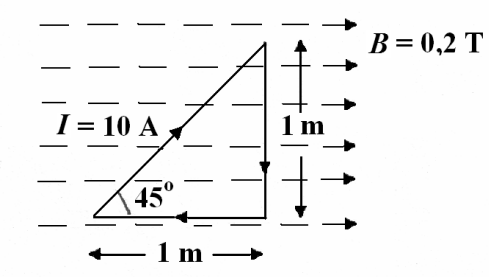
The ions are accelerated through a potential difference *V* and then enter a chamber within which there is a uniform magnetic field *B* directed out of the diagram.

There they move in a semicircle, striking a detector at a distance *d* from the entry point.

Prove that the charge to mass ratio is given by $\frac{q}{m} = \frac{8V}{B^2 d^2}$.

Question 2

The wire loop in the diagram carries a current of 10 A in a magnetic field B = 0,2 T.

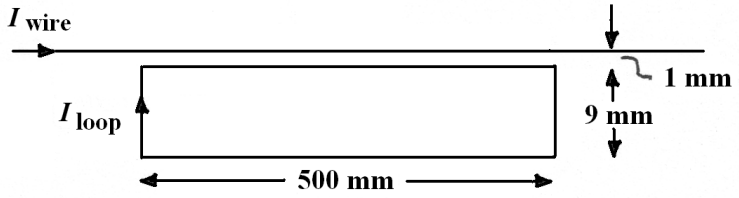


Determine the magnitude and direction of the net force on the loop.

[0, irrelevant]

Question 3

A rectangular wire loop rests on a horizontal surface and has dimensions 500 mm by 9 mm. It carries a current of 10 A. Parallel to the longer side of the loop and attached to the same horizontal surface is a conductor carrying a current of 30 A.



Calculate the force exerted on the loop by the long conductor if the conductor is 1 mm from the loop.

Note that forces exist on the sides of the loop of length 9 mm. However the forces on the left side and right side are equal in magnitude but have opposite directions (since the current flows in opposite directions in the two sides); they do not therefore contribute to the total force.

[0.027 N towards the conductor]

Question 4

You are asked to make a solenoid that is to produce a magnetic field of 2,5 mT at the mid-point of its axis. The solenoid is to be made by winding insulated copper wire round a cardboard cylinder with an outer radius of 20 mm and length of 200 mm.

(i) If a current of 2,0 A is to flow, how many turns must there be on the solenoid?

(ii) What length of copper wire will you need?

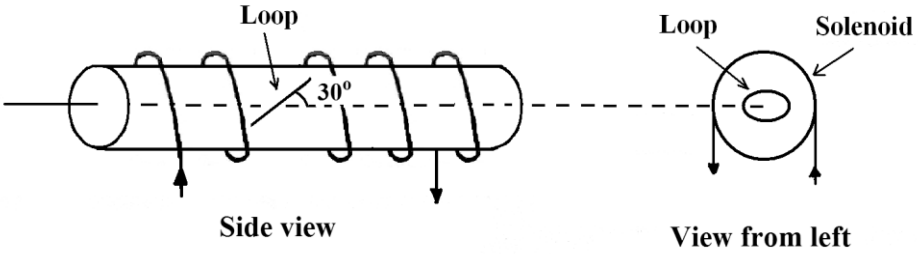
(iii) You have a battery of emf 6,0 V and internal resistance 1,0 Ω with which to drive the current through the solenoid. What must the diameter of the wire be?

The resistivity of copper is $1{,}72 \times 10^{-8}$ Ω.m.

[(i) 199 (ii) 25.0 m (iii) 0.523 mm]

Question 5

A solenoid is 0,5 m long and is wound with 10 layers of wire, each containing 1000 turns. At the centre of the solenoid is a single loop of wire of area 10^{-6} m^2. The plane of the loop makes an angle of 30° with the axis of the solenoid, as shown schematically below.



The current in the solenoid is reduced linearly from 2 A to zero in 2 ms.

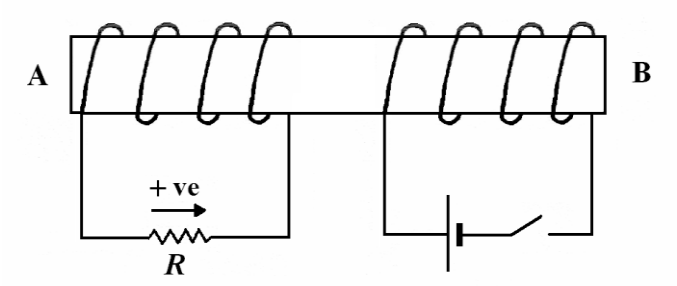
(i) What is the magnitude of the emf induced in the loop?

(ii) Explain whether the direction of the induced current is clockwise or counterclockwise as you look down the axis of the solenoid from the left.

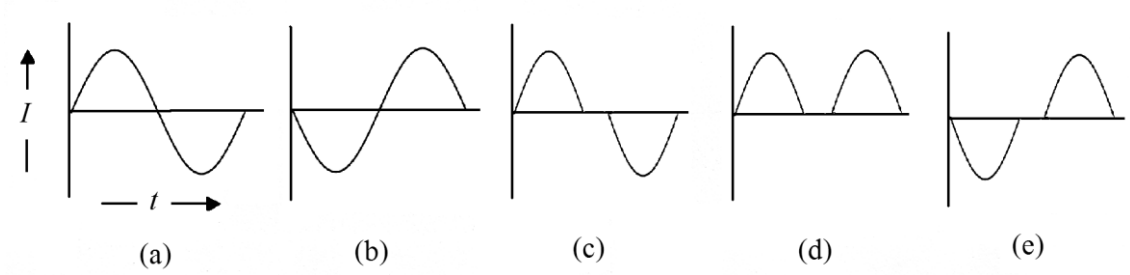
[(i) 12.6 μV (ii) counterclockwise]

Question 6

Two coils A and B are connected as shown in the figure and placed close to each other. The switch is initially open. The switch is then closed and left closed for a while before it is opened again.



Which of the following graphs best represents the flow of the induced current through the resistor *R* as a function of time? The positive direction of the induced current is indicated in the figure on the left.



[(c)]

Question 7

A step-up transformer connected to a 204 V line is to supply 18 kV for a neon sign. To reduce shock hazard, a fuse is to be inserted in the primary circuit. The fuse is to blow when the current in the secondary circuit exceeds 10 mA. Calculate,

(i) the turns ration of the transformer

(ii) power which must be supplied to the transformer when the secondary current is 10 mA

(iii) current rating that the fuse in the primary circuit should have.

[(i) 75:1 (ii) 180 W (iii) 0.75 A]

PHYSICS ID (PHYS1001/1006) LECTURE NOTES

5. OPTICS

5.1. GEOMETRICAL OPTICS (PLANE INTERFACES)

5.1.1. Nature and Propagation of Light

Light sometimes seems to behave like a wave and sometimes like a stream of particles, depending on the type of experiment used to investigate it.

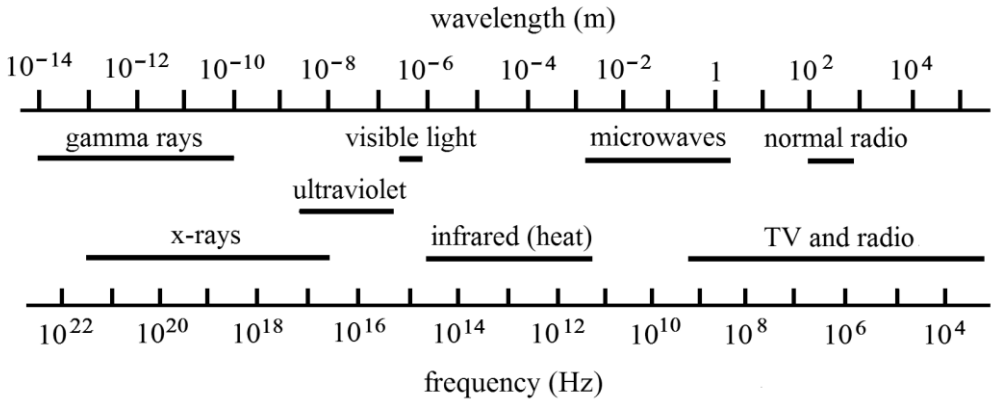
The wavelike properties are important in physical optics, but the nature of light is unimportant in geometrical optics, where we are concerned with the propagation of light and its directional properties in travelling through media.

Note that light appears to travel in straight lines, e.g. we cannot see around corners and light sources cast shadows of an object. Light can however be deviated by:

Gravity	
Reflection and refraction	- studied in geometrical optics
Diffraction	- studied in physical optics

The electromagnetic spectrum

Visible light forms only a very small part of the whole spectrum of electromagnetic waves, as illustrated in the diagram below.



The divisions between regions of the spectrum are not sharp and they overlap to some extent. The name given to the radiation sometimes depends on its origin (as, for example, with γ rays which emanate from the nucleus and X rays whose origin is the atom).

Our eyes are sensitive to electromagnetic radiation with wavelengths from about 400 nm (violet) to about 700 nm (red). Visible light may be divided into regions:

	Representative wavelengths	Approximate limits
Violet	410	400 - 424
Blue	470	424 - 491
Green	520	491 - 575
Yellow	580	575 - 585
Orange	600	585 - 647
Red	650	647 - 700

The eye is not equally sensitive to all wavelengths, being greatest for wavelengths around 560 nm.

The speed of light

In a vacuum all electromagnetic waves (including light) travel with the same velocity, which we take as $c = 3{,}00 \times 10^8$ m.s^{-1}. In a medium the velocity of light is **reduced** to a value c_n, where

$$\boxed{c_n = \frac{c}{n}}$$

and n is known as the **refractive index** of the medium ($n > 1$ always).

- The frequency of light is determined by its source and remains constant when it passes from one medium to another.

The wavelength λ of light is related to its speed and frequency; in a vacuum

$$c = f\lambda$$

Hence $c/n = f\lambda/n$, so that in a medium of refractive index n

$$\boxed{c_n = f\lambda_n}$$

where $\lambda_n = \lambda/n$ is the wavelength of the radiation in the medium.

- The wavelength of light therefore decreases by a factor n when it passes from a vacuum into a medium of refractive index n.

Some useful values of the refractive index are given in the table below.

Medium	Refractive index
Air (at 0°C, 1 atm)	1.00029
Glass (zinc crown)	1.52
Water	1.33
Diamond	2.42

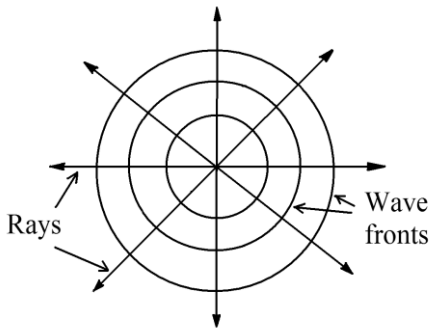
The actual value of c_n (and hence n) for a particular medium depends slightly on the wavelength (and therefore also on the colour) of the radiation, as discussed later.

- For air $n_{air} = 1{,}00029$, which we can approximate as $n_{air} = 1$.

Wavefronts and Rays

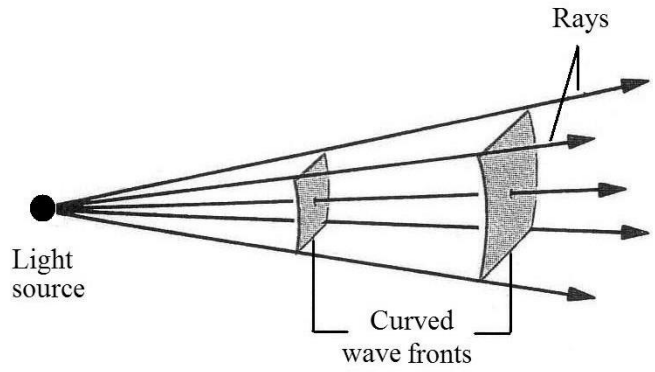
To illustrate and explain phenomena in geometrical optics, the **ray approximation** is used. Here we introduce the concepts of wavefronts and rays; we will investigate the connection between light rays and the wave nature of light when we discuss Huygens' principle in a later section.

- A **wavefront** is a surface passing through the points of a wave that have the same phase and amplitude.
- A **ray** is an imaginary line drawn perpendicular to the wavefront; it indicates the direction of travel of the wave.



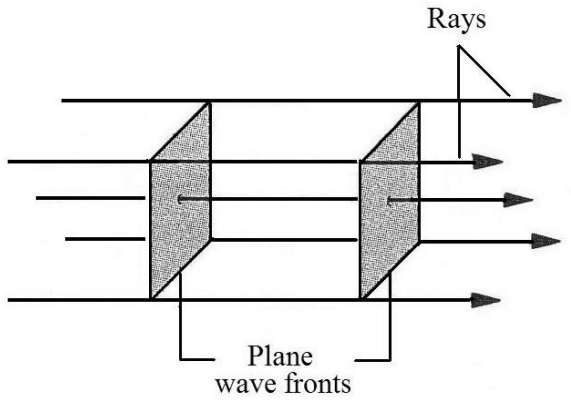
A stone dropped into a pool of water produces circular waves which propagate radially outwards.

The wave fronts are one wavelength apart.



Similarly, near a point source of light the wavefronts are spherical and the rays are directed radially outwards in all directions.

Only portions of the wave fronts are shown in the diagram on the left.



The curvature of the wave front decreases with distance from the light source.

Far from the source, the wave fronts can be regarded as planar and the rays are effectively parallel.

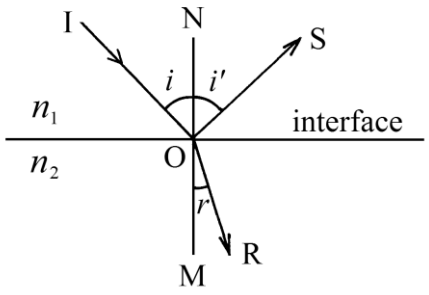
- It is important to realise that a **source does not emit rays** – it emits waves in the form of wave packets (or photons), as discussed later.
- Rays are construction lines drawn to show the direction in which light travels.

5.1.2. Reflection and Refraction at a Plane Interface

Light travels in a straight line in a homogeneous medium until it encounters a boundary between two different materials.

When light strikes the interface between two media of refractive indices n_1 and n_2, some of it is reflected and some (for a transparent medium, most) is refracted into the second medium.

This is illustrated in the following diagram; note the convention that all angles are measured from the normal, and not from the interface.



IO incident ray

NOM normal to the interface

OS reflected ray

OR refracted ray

i angle of incidence

i′ angle of reflection

r angle of refraction

The reflected and refracted rays at a **smooth** interface obey the following relationships:

(i) The incident ray, the reflected ray, the refracted ray and the normal to the interface at the point of incidence all lie in the same plane.

(ii) The angle of incidence *i* is equal to the angle of reflection *i′*. This is the law of reflection.

(iii) The angle of refraction and the angle of incidence are related by:

$$\boxed{n_1 \sin i = n_2 \sin r} \qquad \text{(Snell's law)}$$

This is Snell's law, which states that the ratio of the sine of the angle of incidence to the sine of the angle of refraction is a constant for two particular media. It is generally accepted that this relationship was discovered experimentally in 1621 by Willebrørd Snell (1591-1626).

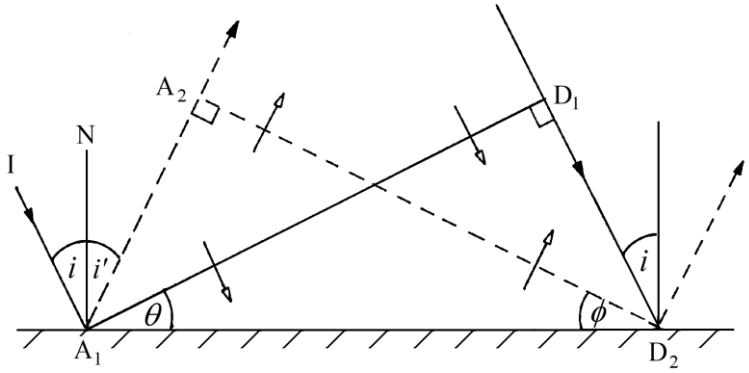
Note that:

- In Snell's law, each sine term is multiplied by the refractive index of the medium in which the angle lies.
- The path of a light ray through a refracting surface is reversible (this follows from interchanging the angles in Snell's law).
- It follows from Snell's law that if $n_2 > n_1$, then $r < i$. Therefore when a ray goes from a medium of lower refractive index to one of higher refractive index (as in the diagram above), it is deviated towards the normal.
- When going in the opposite direction a ray is deviated **away from** the normal.

Proof of the law of reflection

The law of reflection follows from the fact that the incident and reflected waves are travelling in the same medium and therefore travel at the same speed; consequently they travel equal distances in any time interval.

In the diagram below, A_1D_1 represents the wave front at some instant and A_2D_2 represents the same wave front at some later instant.

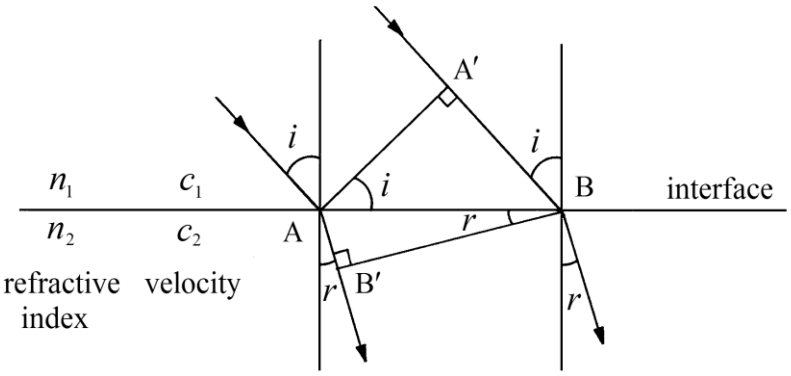


$D_1D_2 = A_1A_2$	Two rays in the same medium travel the same distance in the same time interval.
Angle $A_2 = D_1 = 90°$	The rays are perpendicular to wavefronts.
$\theta = \phi$	The triangles $A_1D_2D_1$ and $A_1D_2A_2$ are congruent, since the hypotenuse A_1D_2 is common to the two triangles and $A_1A_2 = D_1D_2$.
$i = \theta$	Angle $NA_1D_1 + i = 90°$ (IA_1 is perpendicular to A_1D_1) and angle $NA_1D_1 + \theta = 90°$ (NA_1 is perpendicular to A_1D_2).
$i' = \phi$	Angle $A_2A_1D_2 + i' = 90°$ (NA_1 is perpendicular to A_1D_2) and angle $A_2A_1D_2 + \phi = 90°$ (in triangle $A_1D_1D_2$).

Therefore, since i and i' are each equal to angles which are equal to each other, we obtain $i = i'$.

Proof of Snell's law

Consider a parallel beam of light incident on a plane interface separating media of refractive indices n_1 and n_2. In the diagram below, all angles marked i are equal (as shown above) and similarly all angles marked r are equal.



At some instant the wavefront is at AA′; after time t it has reached B′B. In this time one ray travels from A to B′ at speed c_2 whilst the other ray travels from A′ to B at speed c_1. Therefore:

$$A'B = c_1 t \quad \text{and} \quad AB' = c_2 t$$

leading to

$$\frac{A'B}{AB'} = \frac{c_1}{c_2} = \frac{n_2}{n_1}$$

where the last equality follows from the definition of the refractive index: $c_1 = c/n_1$ and $c_2 = c/n_2$.

But from the diagram

$$\mathrm{A'B} = \mathrm{AB}\sin i \quad \text{and} \quad \mathrm{AB'} = \mathrm{AB}\sin r$$

leading to

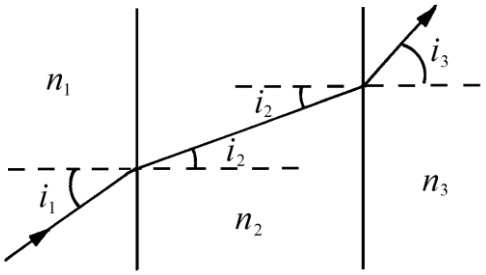
$$\frac{\mathrm{A'B}}{\mathrm{AB'}} = \frac{\sin i}{\sin r}$$

Combining equations we obtain Snell's law:

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1}.$$

Refraction of light by parallel layers

In the diagram below, light is travelling from medium 1 to medium 3 via medium 2; the light is refracted at both interfaces, which are parallel to each other.

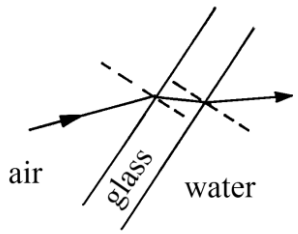


At interface 1/2: $n_1 \sin i_1 = n_2 \sin i_2$

At interface 2/3: $n_2 \sin i_2 = n_3 \sin i_3$

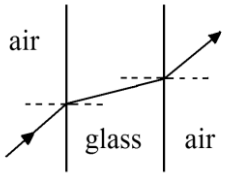
Combining: $n_1 \sin i_1 = n_3 \sin i_3$

Therefore, if the intervening layer is parallel-sided, the angle i_3 at which the light emerges depends only on the refractive indices of the first and last media and on the angle of incidence i_1. The intervening layer has no effect on the direction of the emerging ray.



For example, this can be regarded as an air/water interface.

Note that if $n_1 = n_3$, then if follows from the equation $n_1 \sin i_1 = n_3 \sin i_3$ that $\sin i_1 = \sin i_3$; therefore $i_1 = i_3$.

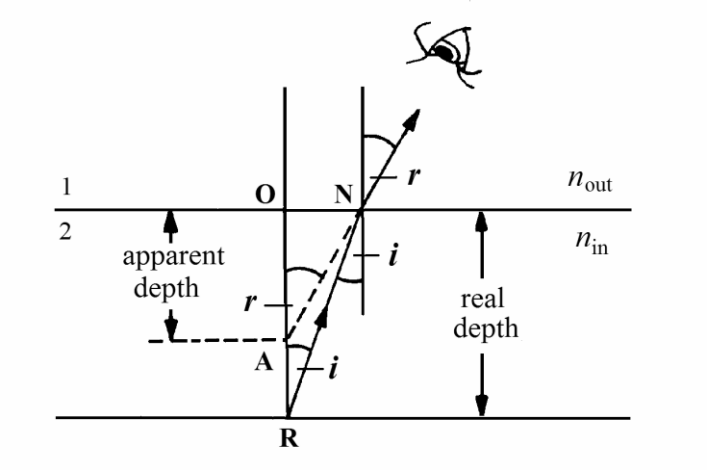


So, if the refractive index of the medium in which the ray emerges is the same as that in which it is incident, the **emerging ray will be parallel to the incident ray but laterally displaced**.

Real and apparent depth

An object within a medium, when viewed by an observer outside the medium, appears to be either further away or closer than it really is; this is a result of refraction of the emerging light rays.

If the observer is in the optically less dense medium (i.e. smaller refractive index) the object seems closer, as in the diagram below. This would be appropriate for an object at the bottom of a swimming pool being viewed from above the surface of the water.



An object is situated at point R. To the eye the rays appear to come from point A, nearer to the interface than R, because they are refracted away from the normal on going from medium 2 to medium 1 which has a smaller refractive index. This follows from Snell's law:

$$\frac{\sin r}{\sin i} = \frac{n_{\text{in}}}{n_{\text{out}}}$$

where n_{in} is the refractive index of the medium in which the object is situated and n_{out} is the refractive index of the outside medium, in which the observer is situated.

We now assume that the object in the diagram is viewed from a point **nearly vertically** above it, so that the rays are almost normal to the interface; then i and r are small and

$$\sin i \simeq \tan i \text{ and } \sin r \simeq \tan r .$$

Snell's law now becomes

$$\frac{n_{\text{in}}}{n_{\text{out}}} = \frac{\tan r}{\tan i} = \frac{\text{ON}}{\text{OA}} \cdot \frac{\text{OR}}{\text{ON}} = \frac{\text{OR}}{\text{OA}},$$

or

$$\boxed{\frac{n_{\text{in}}}{n_{\text{out}}} = \frac{\text{real depth}}{\text{apparent depth}}}$$

Note that this equation holds for **nearly-normal incidence only**.

If medium 1 is air, the equation becomes

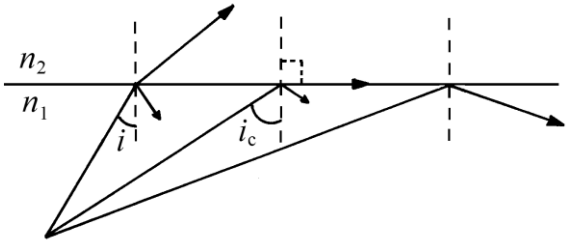
$$n_{\text{in}} = \frac{\text{real depth}}{\text{apparent depth}}$$

This provides a method for measuring the refractive index of a transparent material.

Total internal reflection

It follows from Snell's law that when light travels from an optically dense to a less dense medium (i.e. smaller refractive index) it is refracted away from the normal and that the angle of refraction increases as the angle of incidence is increased.

At some angle of incidence i_c the angle of refraction becomes 90°. If i is increased beyond i_c the beam is **totally reflected** and **no light is transmitted** into the optically less dense medium. The reflected ray still obeys the normal laws of reflection.



This phenomenon, illustrated in the diagram to the left, is known as **total internal reflection** and i_c is called the **critical angle**.

At the critical angle, $n_1 \sin i_c = n_2 \sin 90° = n_2$. Hence the critical angle is given by

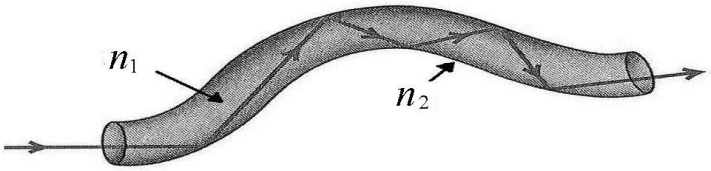
$$\boxed{\sin i_c = \frac{n_2}{n_1} \quad \text{for } n_1 > n_2} \qquad \text{(total internal reflection)}$$

For a glass/air interface with n_{glass} = 1,52 the critical angle is 41,0°. Therefore, when i > 41,0° the light will be totally reflected.

Note that total internal reflection can only occur when light goes from a medium of higher refractive index to one of lower refractive index.

Total internal reflection is employed in the construction of the periscope used in submarines.

Fibre optics is based on total internal reflection. A fibre consists of a thin flexible transparent core of glass or transparent plastic, surrounded by cladding, which is a material of **lower** refractive index than the core.



If $n_2 < n_1$ and the angle of incidence is large enough, light will be totally reflected each time it strikes the interface, and will emerge from the end of the fibre with little loss. This will happen even if the fibre is bent. In practice, a large number of fine fibres is used (for flexibility and good resolution), forming a fibre bundle.

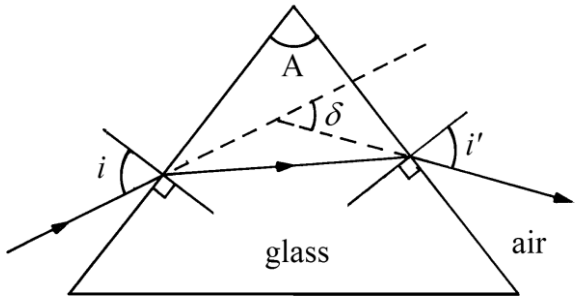
Optical fibres of this kind are used:

- in medicine, to examine the internal organs. Two fibre-optic cables are used, one to transmit light into the body and illuminate the organ, the other to transmit the images back.
- in the communications industry, to carry high-speed Internet traffic, radio and television signals, and telephone calls.

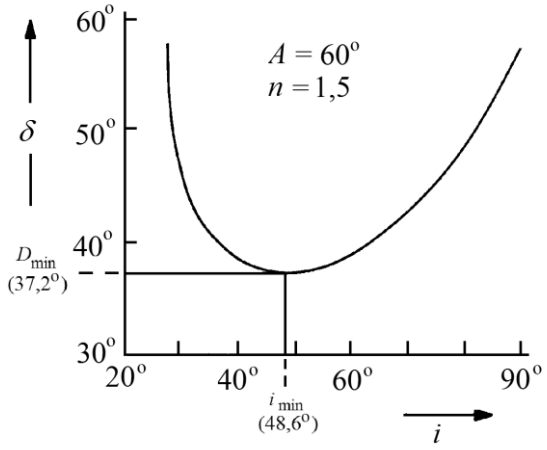
5.1.3. The Prism

Light is deviated when it passes through a prism.

- The **angle of deviation** δ is the angle between the incident and emergent rays.
- The **refracting angle** A is the angle between the refracting faces.



The angle δ depends on A, n (the refractive index of the prism material) and i (the angle of incidence). For a given prism, A and n are fixed and plotting δ versus i produces a curve similar to that in the next diagram.



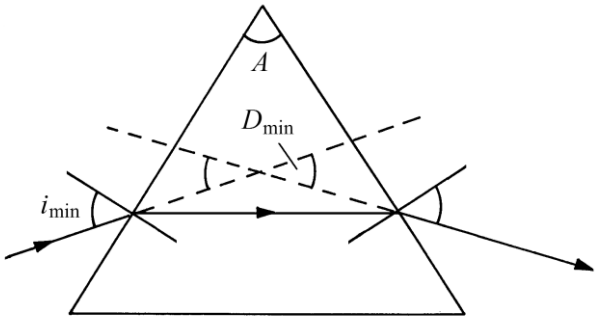
Note that the curve is not symmetrical.

In general two values of the angle of incidence produce the same deviation.

This follows from the fact that the path of any ray is reversible; in the diagram above the angles of incidence i and i' both produce the same deviation δ.

There is a particular value of the angle of incidence, i_{min}, for which the deviation has its minimum value, D_{min}. This is called the **angle of minimum deviation**, where the angles i_{min} and D_{min} depend on A and n.

At minimum deviation the ray travels symmetrically through the prism, i.e. it makes equal angles with the two refracting faces, which we prove as follows.



The prism in the diagram above is set at the position of minimum deviation.

Imagine that the direction of the ray through the prism is reversed so that light is incident from the right. From the diagram, the angle of deviation of the reversed ray is also D_{min} (vertically opposite angles are equal).

Since only one angle of incidence can produce this angle of deviation (see the previous graph), the angle of incidence for the reversed ray must also be i_{min}.

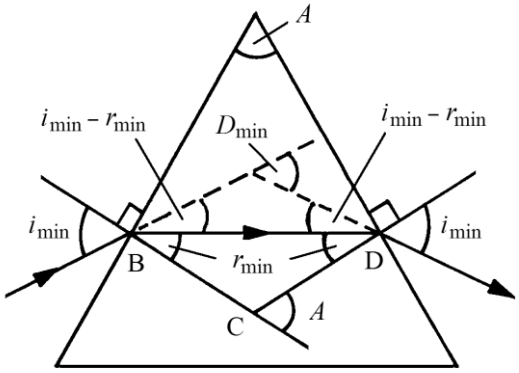
Thus this ray must travel symmetrically through the prism, as shown in the diagram.

Refractive Index of a Prism

Measurement of the angle D_{min} provides a very accurate method for determining the refractive index of a prism.

- The method is accurate because the curve of δ versus i has a flat minimum and so even if the angle of incidence is not exactly i_{min} the deviation will be very close to D_{min}.

The prism in the diagram below is set at the position of minimum deviation.



In the quadrilateral ABCD there are right angles at B and D so that within the quadrilateral $A + C = 180°$. But the angle within the quadrilateral at C and the angle exterior to the quadrilateral at C must also sum to 180°. Therefore the exterior angle at C must equal the refracting angle A, as shown.

For any triangle, the exterior angle equals the sum of the interior opposite angles. Hence

$$A = 2r_{min} \quad \text{and} \quad D_{min} = 2\left(i_{min} - r_{min}\right)$$

and so

$$r_{min} = \frac{A}{2} \quad \text{and} \quad i_{min} = r_{min} + \frac{D_{min}}{2} = \frac{A + D_{min}}{2}.$$

From Snell's law at B:

$$n_{air} \sin i_{min} = n_{prism} \sin r_{min}$$

giving

$$\boxed{n_{prism} = \frac{\sin\left(\frac{A + D_{min}}{2}\right)}{\sin\frac{A}{2}}}$$

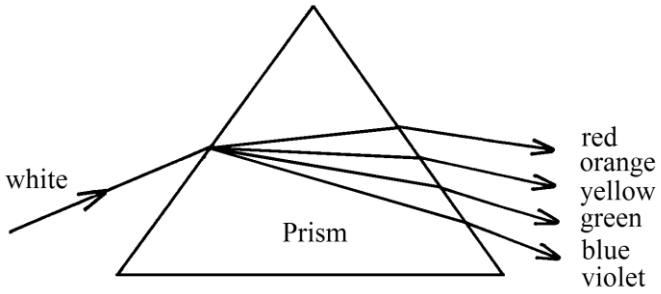
Chromatic dispersion

Careful measurement shows that the refractive index (and therefore the speed of light) in anything but a vacuum depends somewhat on the wavelength of the light, i.e. its colour. This property is called **dispersion** and the medium is said to be **dispersive**.

As an example, the refractive index of light travelling in flint glass decreases smoothly from about 1,66 for a wavelength of 400 nm to about 1,62 for light of wavelength 700 nm.

It follows that a prism will deviate different colours through different angles, forming a **spectrum**.

- If the light incident on the prism contains all visible wavelengths (white light), a continuous spectrum is produced.



We assume that the refractive index of the prism material *decreases* with wavelength, which is the situation for most materials.

Then, from Snell's law, smaller wavelengths (e.g. violet light) will be deviated by the prism more than larger wavelengths (e.g. red light).

The angle between the emerging red and the violet rays is a measure of the dispersive power of the prism – the larger the dispersive power the greater the separation. Diamond owes its brilliance partly to its large dispersive power.

- If the incident light contains only certain wavelengths (such as light from a mercury vapour lamp) then the spectrum will contain a bright coloured line for each of the wavelengths that are present.

  As we shall discuss later, the wavelengths emitted by a particular element are characteristic of the element and can be used to uniquely identify it.

  This is the basis of the **prism spectrometer**, which is commonly used to study the wavelengths emitted by a light source. It is employed in biology and chemistry to identify molecules (using infra-red light) and in astronomy to identify elements in distant stars.

Dispersion accounts for the colours of the rainbow, where water droplets are the dispersive medium. The primary bow is formed by light which has been reflected once in the drops and the secondary by light which has been reflected twice. For a more detailed explanation, see the prescribed textbook.

5.1.4. Huygens' Principle

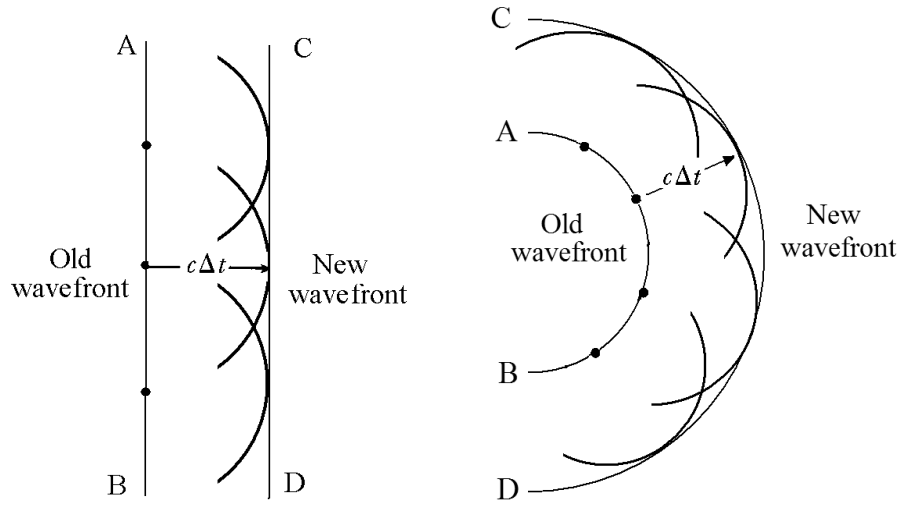
Christiaan Huygens (1629–1695) was a Dutchman and a contemporary of Newton. His principle, which he proposed in 1678, enables us to find the position and shape of a wavefront if its earlier position and shape are known. He used his principle to explain the laws of reflection and refraction of light.

The principle states:

1. Every point on a wavefront can be considered as a source of secondary wavelets that spread out in the forward direction with the speed of the wave itself in the particular medium.

2. The new wavefront at any instant is the envelope (tangent) to the secondary wavelets at that instant.

The principle is illustrated for two common situations in the following diagram (a plane wave on the left and a spherical wave on the right). In each diagram, AB represents the wavefront at some instant and CD represents the wavefront a time Δt later.

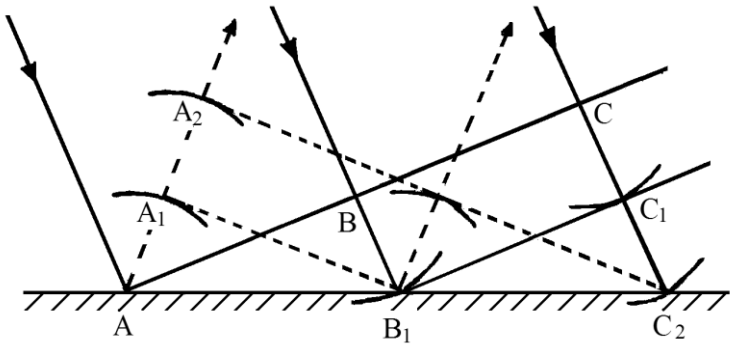


As can be seen, while the wave is travelling in a medium of uniform refractive index (so that its speed does not change), the wavefronts remain plane or spherical, respectively.

Huygens' Principle can also be used to determine what happens when a wave meets the interface with another medium, with a different refractive index.

Reflection from a plane interface

We first use Huygens' Principle to illustrate the reflection of plane waves from the boundary between two media.



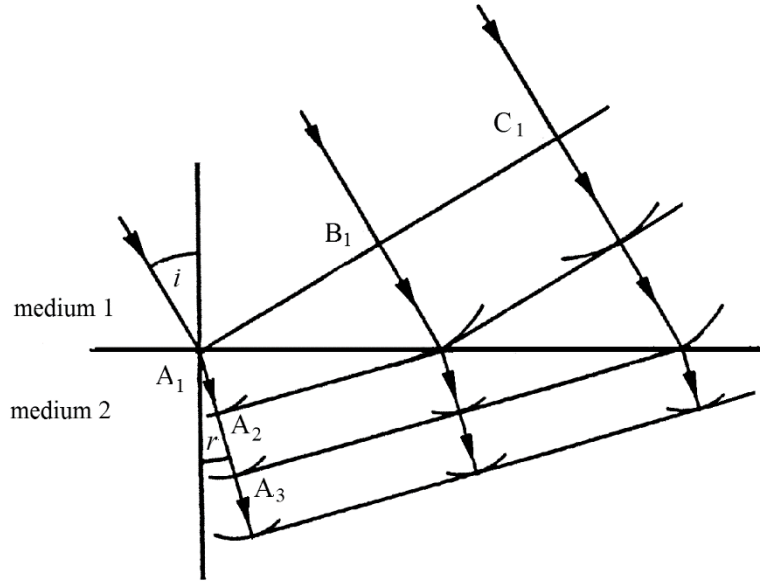
The solid lines indicate the wavefronts (and rays) before reaching the interface or boundary.

The dashed lines show the situation after reflection.

The wavefront ABC is incident on the interface. After time Δt the wavefront is $A_1B_1C_1$, and after a further time Δt it is $A_2B_2C_2$.

Refraction at an interface

The principle can also be applied to examine the change in direction of a plane wave as it travels from one medium into another medium.



Light is travelling from medium 1 to medium 2, in which it travels more slowly.

It can be seen that the slowing down causes the rays and wave fronts to be bent on passing through the interface.

By using Huygens' Principle to identify the position of the wavefronts in situations such as those illustrated here, and drawing rays perpendicular to the wavefronts, we can construct "ray diagrams" which indicate only the direction of propagation of the wave.

In geometrical optics, where we are concerned only with the directional properties of light propagation, we need draw only the ray diagram. This was the approach employed in the earlier discussion of reflection and refraction – it can be seen that this approach is underpinned by Huygens' Principle.

5.2. GEOMETRICAL OPTICS (CURVED INTERFACES)

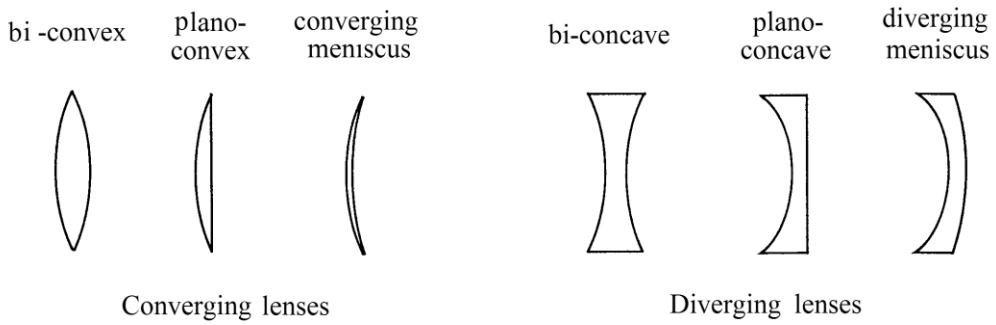
5.2.1. Lenses

Any system in which curved interfaces separate transparent media of different refractive indices is a lens. Examples are the eye and the glass lens in air. Lenses may also be made from some transparent plastics.

When the lens is surrounded by air, light refracts from the air into the lens, crosses through the lens, and then refracts back into the air. Each refraction can change the direction of the light, according to Snell's law.

Types of lenses

We deal only with **thin lenses** (the lens must be thin compared with the object and image distances and the focal length, all defined below). These are usually circular and the two faces are portions of a sphere. Each face can be concave, convex or plane. For example:

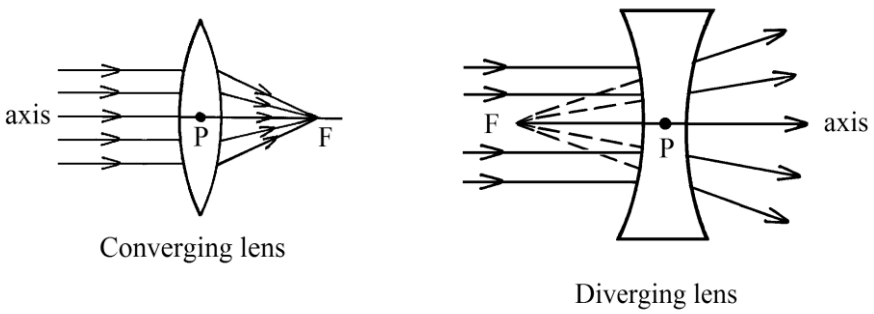


We confine our attention to rays that strike the lens close to its axis and that make small angles with the axis; these are called **paraxial** rays.

Lenses can be **converging** (which are thicker at the centre than at the edges) or **diverging** (which are thicker at the edges).

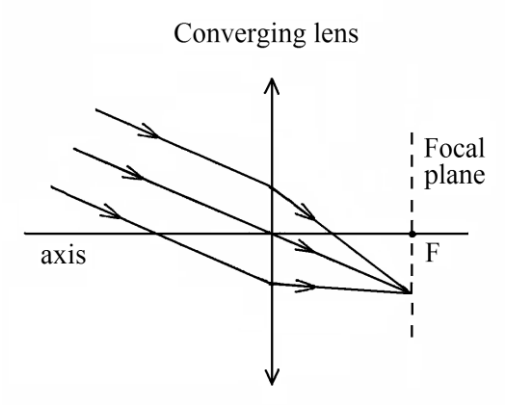
Rays parallel to the axis that pass through a converging lens all pass through a point on the axis on the other side of the lens – the **focal point**.

Rays parallel to the axis appear to diverge from the focal point after passing through a diverging lens.



P is the point where the axis meets the lens, i.e. the **pole** of the lens, and F is the **focal point**.

There is a focal point on each side of the lens and these are always equidistant from the pole if there is the same medium on each side, irrespective of whether the two faces of the lens have the same curvature. The distance from the pole to the focal point is the **focal length** of the lens.

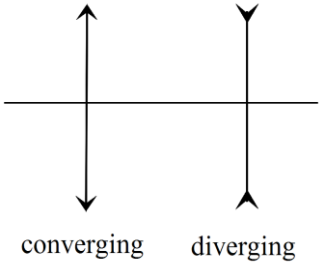


If the beam of parallel rays is not parallel to the axis it will be brought to a focus at a point off the axis but on the **focal plane**.

This is a plane perpendicular to the axis and passing through the focal point.

A similar diagram can be drawn for a diverging lens.

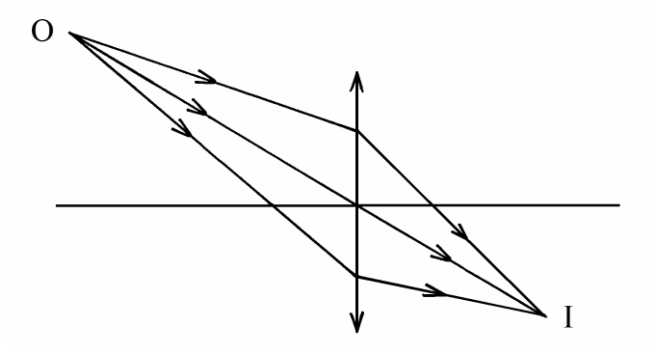
Note how converging and diverging lens are usually represented in diagrams:



Objects and Images

An optical system makes all rays passing through some point on or near the axis pass through some other point on or near the axis, either directly or in projection. These are called **conjugate points**.

The point which is the source of the light is the **object** O and the point conjugate with it is the **image** I.



Since rays are reversible, an object placed at the point labelled I in the diagram will produce an image at the point labelled O.

In other words, the object and image points are interchangeable.

The object is the point associated with **incoming** light (i.e. rays coming in to the lens).

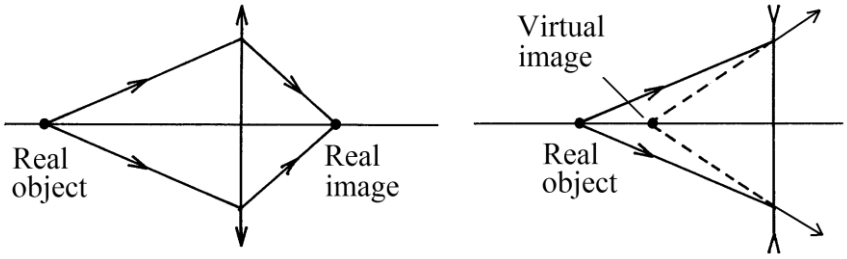
The image is the point associated with **outgoing** light (i.e. rays going out from the lens).

An **extended** object can be thought of as being made up of many point objects. An image is formed of each, collectively producing an extended image.

Objects and images can be **real** or **virtual**, as defined in the following table.

	Real	Virtual
Object	Rays are diverging before striking the lens	Rays are converging before striking the lens
Image	Rays converge when leaving the lens	Rays diverge when leaving the lens

Real and virtual objects and images are illustrated in the following diagrams.



Real images are formed at the point at which rays of light actually intersect (as in the left-hand diagram).

Virtual images are formed at the point from which the rays appear to originate (as in the right-hand diagram).

A **real image** can be cast onto a screen, whereas a **virtual image** cannot. A virtual image can however be seen when the eye focuses the diverging rays onto the retina.

Virtual objects are only of relevance when discussing combinations of lenses (see below).

Ray Tracing

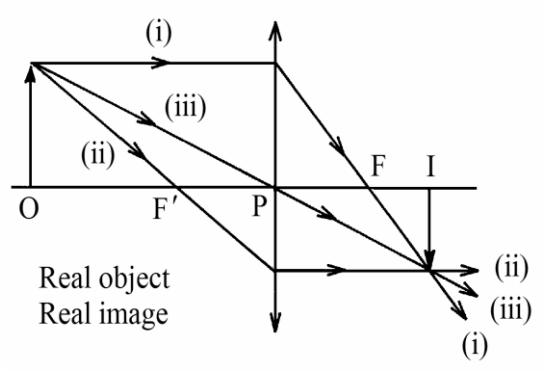
To find the position of an image we must find the point where two or more rays from the object meet again after passing through the lens. Any two of the following can be used:

(i) A ray parallel to the axis will converge to (converging lens) or diverge from (diverging lens) the appropriate focal point.

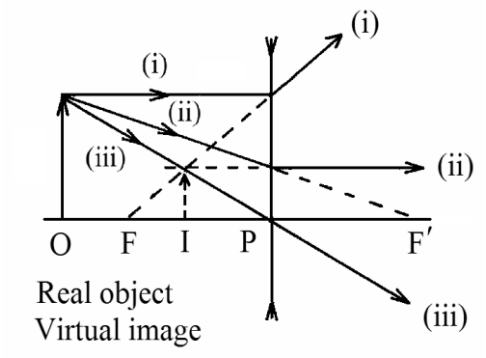
(ii) A ray passing through the other focal point (either directly or in projection) will emerge from the lens parallel to the axis.

(iii) A ray through the pole will continue undeviated.

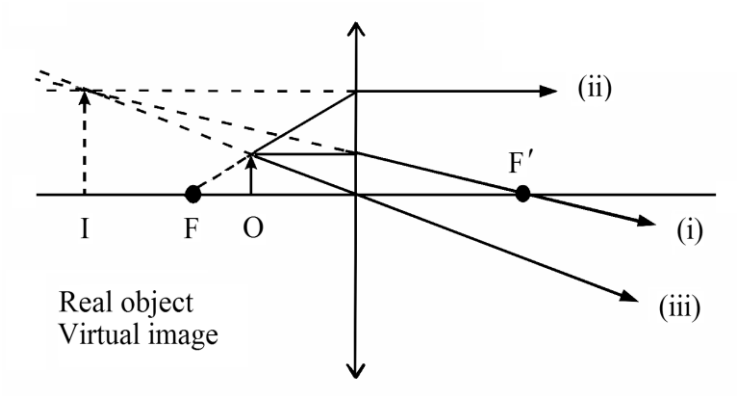
Some examples are shown in the following diagrams.



A converging lens normally forms a real, inverted image of a real object, on the other side of the lens.



A diverging lens forms a virtual image of a real object, with the same orientation and on the same side of the lens.



A converging lens will produce a virtual image if the object is placed between the focal point and the lens.

The image is upright (i.e. not inverted) and is on the same side of the lens as the object.

5.2.2. The Thin Lens

The following sign convention is needed to ensure that the equations we derive are valid for all types of lenses, objects and images.

(i) All distances are measured from the pole of the lens.

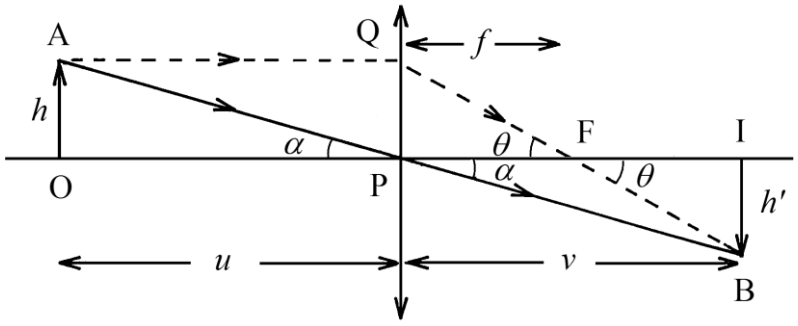
(ii) Distances to real objects and images are positive
Distances to virtual objects and images are negative.

(iii) Focal lengths of converging lenses are positive.
Focal lengths of diverging lenses are negative.

The following symbols are used: f = focal length, u = object distance, v = image distance.

The lens equation

The diagram below shows rays from a real, upright object incident on a converging lens. A real, inverted image is formed. From the sign convention, u, v and f are all positive.



From the triangles AOP and BIP we see that

$$\tan\alpha = \frac{\text{AO}}{\text{OP}} = \frac{\text{AO}}{u} \quad \text{and} \quad \tan\alpha = \frac{\text{BI}}{\text{IP}} = \frac{\text{BI}}{v}$$

or

$$\frac{\text{BI}}{\text{AO}} = \frac{v}{u}$$

Similarly, from triangles QPF and BIF we have

$$\tan\theta = \frac{\text{QP}}{\text{PF}} = \frac{\text{AO}}{f} \quad \text{and} \quad \tan\theta = \frac{\text{BI}}{\text{IF}} = \frac{\text{BI}}{v-f}$$

or, combining equations

$$\frac{\text{BI}}{\text{AO}} = \frac{v}{u} = \frac{v-f}{f}$$

Rearranging this equation, we obtain:

$$\boxed{\frac{1}{f} = \frac{1}{u} + \frac{1}{v}} \qquad \text{(lens equation)}$$

Although derived for a specific situation, this equation holds for all types of lens, object and image provided the sign convention introduced earlier is followed.

Magnification

Linear (or lateral) magnification is defined as

$$|m| = \frac{\text{height of image}}{\text{height of object}}$$

By convention:

The magnification is positive when the object and image have the same orientation.

If the image is inverted relative to the object, the magnification is negative.

In deriving the lens equation, we showed that

$$\frac{\text{height of image}}{\text{height of object}} = \frac{\text{BI}}{\text{AO}} = \frac{v}{u}.$$

For the case considered here the image is inverted relative to the object, so that the sign convention requires that the magnification is negative. Therefore, since u and v are positive:

$$\boxed{m = -\frac{v}{u}} \qquad \text{(linear magnification defined)}$$

This equation holds for any kind of object, image and lens, provided the sign convention introduced for u, v and f is followed.

Power of a lens

The power P of the lens immersed in a medium of refractive index n is defined as

$$\boxed{P = \frac{n}{f}} \qquad \text{(power of lens defined)}$$

If the focal length f of the lens is in metres, its power is in **dioptres**.

- The shorter the focal length the more powerful the lens.
- P has the same sign as f.
- For a lens in air, $P = 1/f$

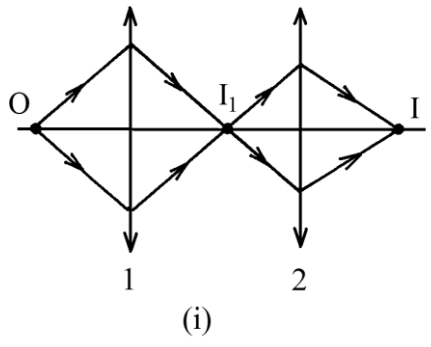
Combinations of lenses

The equations derived above allow us to find the position and magnification of a single lens. In practice, most optical instruments consist of a system of two or more lenses.

If several lenses are used in combination, the image produced by one acts as an object for the next. The following diagrams give several examples.

- In each case the solid lines show the path of rays through the combination of lenses.
- Rays do **not** follow the path indicated by dashed lines, which are included to aid location of virtual objects and images.

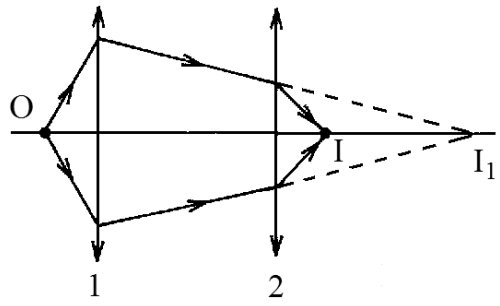
The overall magnification produced by the combination of lenses is the product of the magnifications of the individual lenses, $m = m_1 m_2 m_3 \cdots$



(i)

Rays from the object on the left are converged by the first lens to form a real image between the two lenses.

This image then acts as a real object for the second lens, which converges the rays from the object to form a real image on the right.

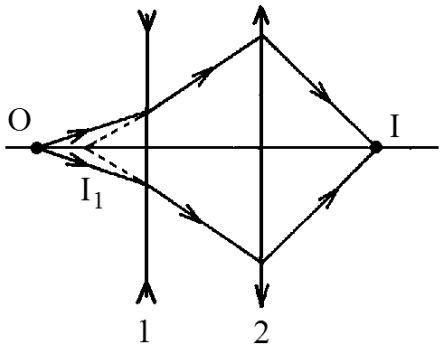


(ii)

If the second lens were not present, the first lens would produce a **real** image labelled I_1 in the diagram; rays do not actually reach this point because of the presence of the second lens.

This real image then acts as a **virtual** object for the second lens (the object is virtual because rays are converging on striking the second lens).

The second lens then forms a final real image at I.

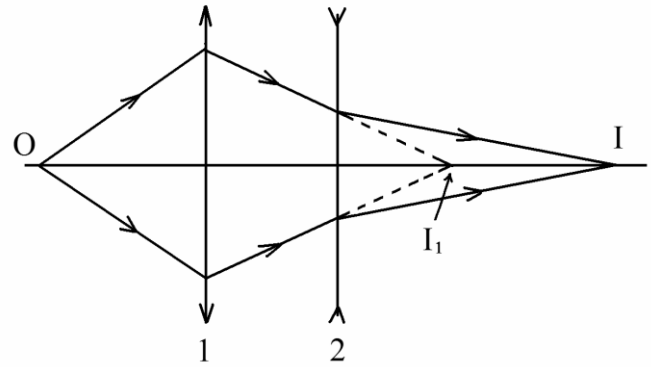


(iii)

The first lens diverges the rays to form a **virtual** image at the point labelled I_1.

This act as a **real** object for the second lens, which then converges the rays to form the final real image to its right.

Note that an object can be real even if the rays do not actually pass through it.



(iv)

As in case (ii), if the second lens were not present, the first lens would produce a **real** image at I_1.

This real image then acts as a **virtual** object for the second lens.

Lens 2 causes the rays to diverge from their original path, but not sufficiently to produce a virtual image.

A real image is formed at I.

In each case the magnitude of the object distance for a lens is the distance from the pole of the lens to the position of the object, whether or not the rays actually pass through the object position, and similarly for the image distance.

For example:

In case (ii), the image distance for lens 1 is the distance from the pole of lens 1 to the point I_1 (and is positive).

The object distance for lens 2 in case (iii) is the distance from the pole of the lens to point I_1 (and is positive).

The object distance for lens 2 in case (iv) is the distance from the pole of the lens to point I_1 (and is negative).

5.2.3. Aberration in Lenses

The theory developed above assumes that:

Rays are paraxial.

Lenses are thin with perfectly spherical surfaces.

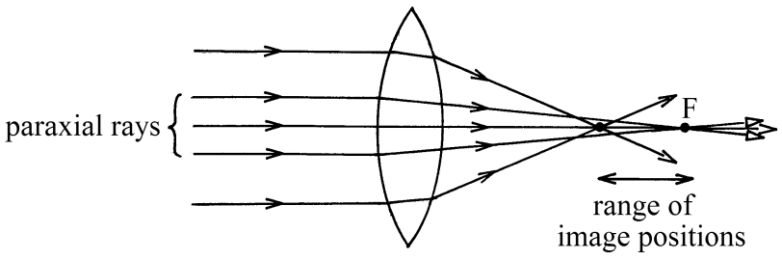
Clearly, these conditions may not be met in real-world applications, with the result that sharply focussed point images may not be formed from point objects.

The departures of real, imperfect images from the ideal predicted by the theory are called aberrations. There are a number of aberrations in lenses (and in the eye), but we deal with only two.

Not all aberrations can be corrected simultaneously, and usually a compromise must be made.

Spherical aberration

Rays far from the axis of a lens with spherical interfaces are not brought to a focus on the focal plane but at points in front of the focal plane.



This causes a variation in image position with distance of incident rays from the axis, so that the image is not sharp.

The effect can be reduced by:

employing apertures (referred to as stops) to cut off the rays far off-axis (as is done in a camera),

careful choice of curvatures of the two surfaces of the lens,

use of combinations of lenses.

Spherical aberration is small in the eye since:

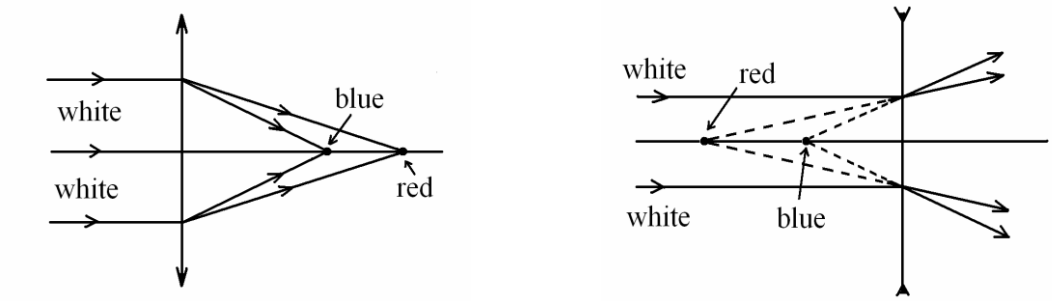
the iris cuts out rays far from the axis,

the refracting surfaces are slightly hyperbolic,

the refractive index of the eye lens decreases with distance from centre (the edge is weaker than the centre).

Chromatic aberration

Because of the slight variation of the refractive index with wavelength, when light of more than one colour passes through a lens, different colours will be focussed at different points. Thus the lens will form a spread-out coloured image of a white object.



For a converging lens the focal length is longer for red light than for blue light, as shown in the diagram on the left (which, for clarity, exaggerates the difference).

The chromatic aberration of a diverging lens is opposite to that of the converging lens.

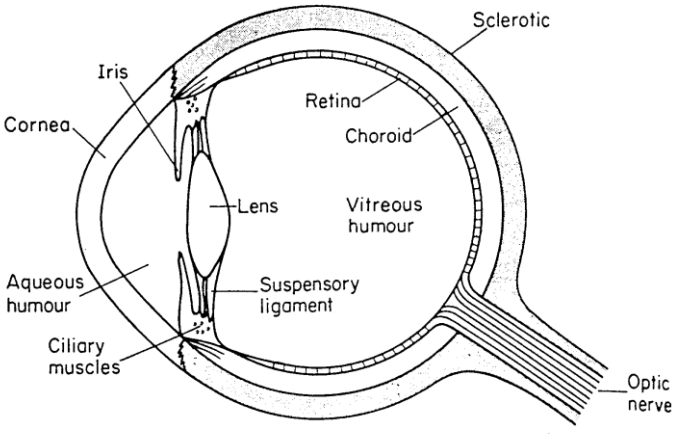
The effect can be reduced by combining two (or more) lenses made of different materials with different refractive indices to form an achromatic doublet. One lens must be converging and the other diverging; otherwise the dispersion will be increased. The effect cannot be completely eliminated simultaneously for all colours but can be greatly reduced.

This is not a serious defect in the eye which focuses near the middle of the visible spectrum (about 556 nm).

5.2.4. The Human Eye

Structure and operation of the eye

Light enters the eye through the **cornea** (a tough transparent skin) where it is refracted. It then goes through a flexible crystalline converging **lens** where it is refracted again. Most of the refraction actually takes place at the cornea, not the lens. This refraction causes an image to be formed on the sensitive **retina**, which transmits impulses to the brain via the **optic nerve**.



The **iris**, or coloured portion of the eye, is a muscular diaphragm that automatically adjusts the size of the **pupil**, or circular opening in its centre, according to the intensity of the light falling on it.

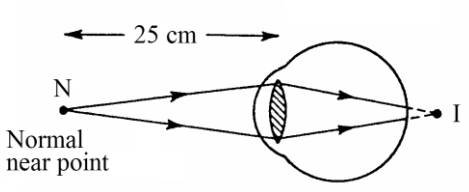
Focussing of the image on the retina is effected by an alteration in the focal length of the eye lens. This is called **accommodation** and is brought about by the ciliary muscles, which vary the thickness, and hence the focal length, of the lens.

When the eye focuses on a distant object, the ciliary muscles are relaxed. This causes the eye lens to become flattened, thereby increasing its focal length. For an object at infinity the focal length of the eye is equal to the distance between the lens and retina, which is about 1,7 cm.

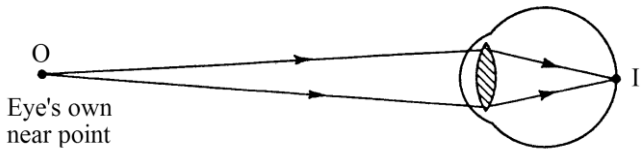
Defects of vision and their correction

The normal eye can accommodate for clear vision of objects from the **far point** (infinity) down to the **near point**, which increases with age but is on average about 25 cm from eye. Typically the distance to the near point is about 18 cm at age 10, increasing to 500 cm or more at age 60.

A **long-sighted** (hyperopic) person can see distant objects clearly but the near point is farther than 25 cm from the eye. This is caused by a mismatch between the focusing power of the lens-cornea system and the length of the eye.

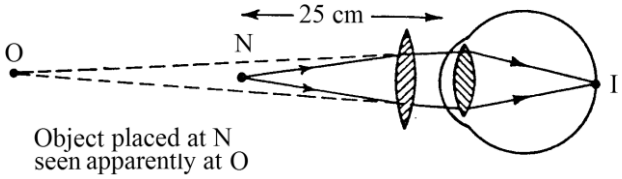


Light rays from the normal near point reach the retina before they have been converged to form a sharp image: the eye ball is too short.



The eye's own near point is further than 25 cm from the eye.

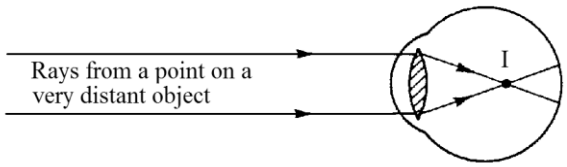
Correction is effected by a converging spectacle lens which forms a virtual image of a close object at the eye's own near point. The eye then focuses on this image.



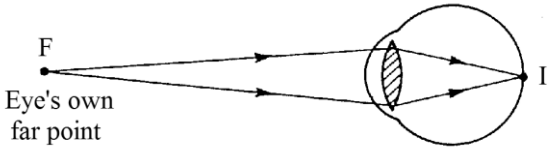
The converging lens causes rays to converge as though coming from O.

Presbyopia (literally "old-age vision") is somewhat similar to long sightedness. As the eye ages, it becomes less able to accommodate and the near point moves out. This is due to a weakening of the ciliary muscles and a hardening of the lens material. The symptoms and correction are the same as for long sightedness.

A **short-sighted** (myopic) person can focus on nearby objects but not on distant objects.

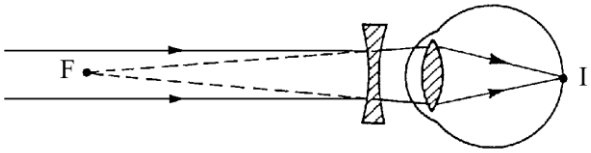


The eye ball is too long so that light from infinity is focussed in front of the retina.



The actual far point of such an eye is much closer to the eye than infinity. It may be as close as 1 m.

Correction is effected by a diverging lens which forms a virtual image at the eye's own far point of an object at infinity.

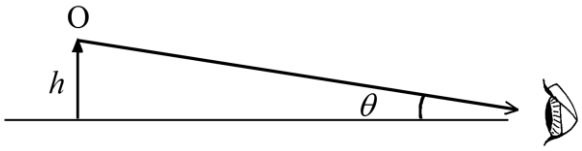


The diverging lens causes rays from a distant object to diverge as though coming from F.

5.2.5. Some Optical Instruments

The function of an optical instrument is to increase the apparent size of an object that is either small or far from the eye.

The size of the image that an object produces on the retina depends on the angle that the image occupies in the field of view of the eye. This is called the **angular size** of the object, θ in the diagram below.



The value of θ clearly depends on how close to the eye the object is placed.

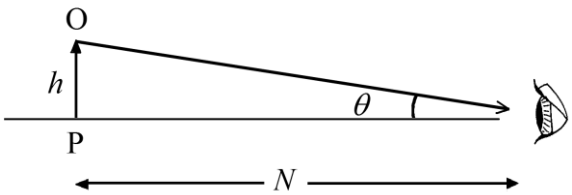
The **angular magnification** of an optical instrument is defined as the ratio of the angular size θ' of an object viewed with the instrument to the maximum angular size $\theta_{\max}$ that can be achieved without it:

$$m_\theta = \frac{\theta'}{\theta_{\max}} \qquad \text{(angular magnification defined)}$$

The simple magnifying lens

This is a single converging lens whose function is to increase the apparent size of a small object whose distance from the eye can be varied.

The normal human eye can focus a sharp image of an object onto the retina if the object is no closer to the eye than the eye's near point, labelled P in the following diagrams.



The angular size of the object O when placed at the near point, a distance N from the eye, is

$$\theta \approx \tan\theta = \frac{h}{N}$$

where we have used the small angle approximation (for a small object the angle must be small).

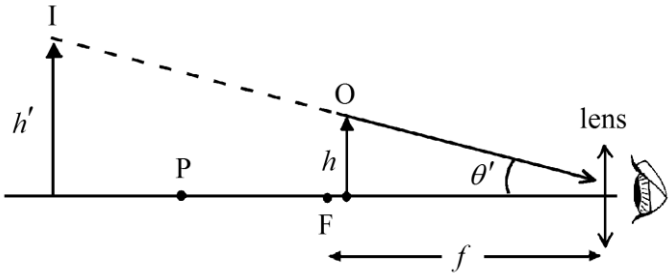
By moving the object closer to the eye, you can increase the angular size and hence the possibility of distinguishing details of the object. However, because the object is then closer than the near point, it is no longer in focus; the image on the retina is no longer clear.

Therefore the **maximum angular size** that can be achieved with the unaided eye is

$$\theta_{\max} = \frac{h}{N} .$$

If the object is moved closer to the eye than the near point, the retinal image can be brought into focus again by looking at the object through a converging lens (the magnifying lens), placed so that the object is inside the focal point of the lens.

The lens then produces a virtual, upright and enlarged image I of the object, which is further from the eye than the image. The focal length of the lens must be chosen so that the image I is beyond the eye's near point P; the eye is then able to focus on this image.



If the object is placed **just** inside the focal point of the lens, the image is formed far from the eye.

This is the ideal situation for comfortable viewing.

The angular size of the virtual image when the magnifying lens is placed directly in front of the eye is

$$\theta' \simeq \tan\theta' = \frac{h}{f} ,$$

which is larger than $\theta_{\max}$, since $f < N$.

From the definition of angular magnification, the magnification of the simple magnifying lens is therefore $\dfrac{h}{f} \Big/ \dfrac{h}{N}$, giving

$$\boxed{m_\theta = \frac{N}{f}} \qquad \text{(magnification of simple magnifier)}$$

This formula is valid if the object is placed at the focal point of the lens so that the image produced by the lens is at infinity. The magnification can be increased somewhat if the final image is formed closer to the near point (but still beyond it); it becomes $(N/f)+1$ if the final image is actually at the eye's near point.

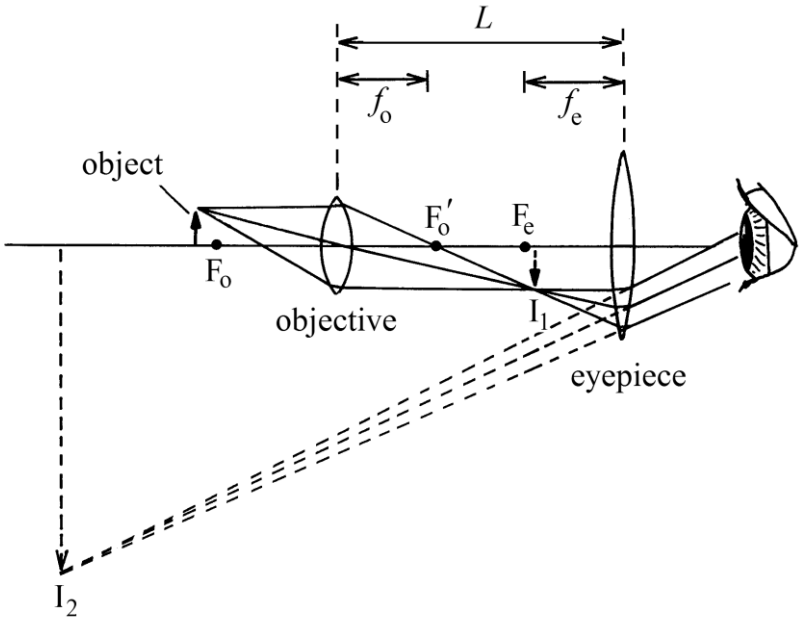
In practice, the maximum magnification that can be achieved is about 4, because of aberrations of the lens. This can be increased to about 20 by using two or more lenses to reduce aberrations.

The compound microscope

In its simplest form, the microscope comprises two converging lenses of short focal length, the objective and the eyepiece (with focal lengths less than about 1 cm and a few cm, respectively). The distance L between the two lenses is much greater than either focal length.

The microscope is used to view small objects that are placed close to the objective. The object must be placed **just outside** the focal point of the objective, which then produces a much enlarged, real and inverted image I_1 of the object. The distance L between the two lenses is adjusted so that this image is formed **just inside** the focal point of the eyepiece.

The image I_1 produced by the objective acts as an object for the eyepiece which produces the virtual image I_2 which is even further enlarged. This image I_2 is the image seen by the eye.



The magnification M of the microscope is the product of the magnifications of the two lenses. The linear magnification produced by the objective is

$$m_o = -\frac{v_o}{u_o} \simeq -\frac{L - f_e}{f_o}$$

Since the image produced by the objective is just inside the focal point of the eyepiece, the latter acts like a simple magnifying lens. The angular magnification of the eyepiece, for an image produced far from the eye, is therefore

$$m_e \simeq \frac{N}{f_e}$$

where N is the distance to the eye's near point. The overall magnification $M = m_o m_e$ of the microscope is therefore

$$\boxed{M = -\frac{L - f_e}{f_o}\frac{N}{f_e}}$$

with the image being inverted with respect to the object.

The usefulness of the microscope is limited by two factors (in addition to lens aberrations).

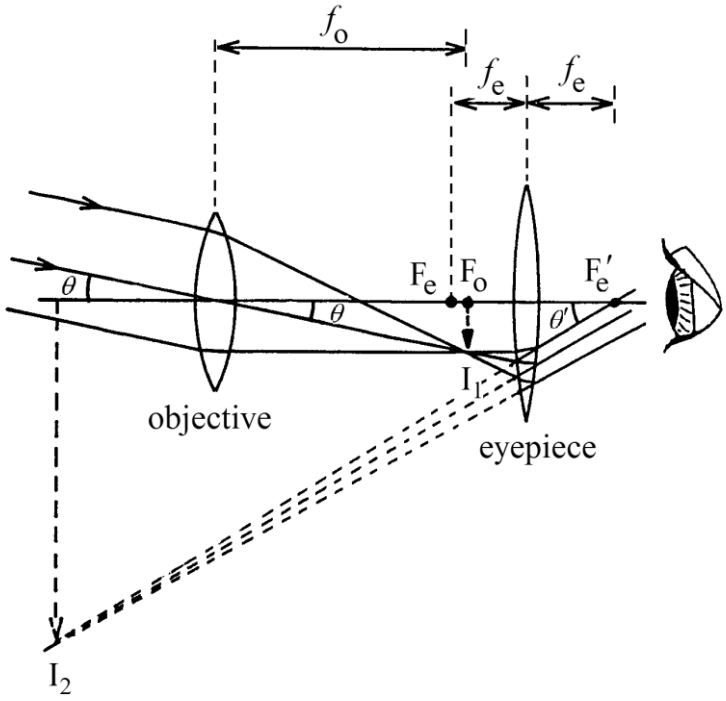
(i) The object being viewed must be larger than the wavelength of the light used to view it. For visible light this means greater than about 1 μm.

(ii) In practice, the magnification is limited by diffraction of light as it passes through the lenses, as will be discussed later.

The astronomical telescope

We discuss here the **refracting** type of astronomical telescope, which consists of two converging lenses, the objective and the eyepiece, located at opposite ends of a long tube. This differs from the reflecting telescope, which uses a lens and a curved mirror to form an image.

Rays from a very distant object are essentially parallel on reaching the telescope. A real, inverted image I_1 is therefore formed in the focal plane of the objective. This image acts as an object for the eyepiece, which produces a final virtual, greatly magnified image I_2. It is this image that the eye focuses on.



In normal relaxed viewing the distance between the two lenses is adjusted so that the focal planes of objective and eyepiece coincide and the final image is at infinity (for clarity, the focal points are shown slightly separated in the diagram). The two lenses are then separated by the distance $f_o + f_e$ which is the length of the tube of the telescope.

The function of the telescope is to increase the angle which a distant object appears to subtend at the eye, and therefore produce the same effect as if the object were either larger or else closer to the eye.

The telescope's total angular magnification is given by $M = \dfrac{\theta'}{\theta}$. From the diagram the angle subtended at the eye by the final image is:

$$\theta' = \frac{h}{f_e}$$

where h is the size of the intermediate image I_1, whereas the angle that would be subtended by the object is:

$$\theta = \frac{h}{f_o}$$

Thus the angular magnification is

$$\boxed{M = \frac{\theta'}{\theta} = \frac{f_o}{f_e}}$$

So for high magnification the objective should have a long focal length and the eyepiece a short one.

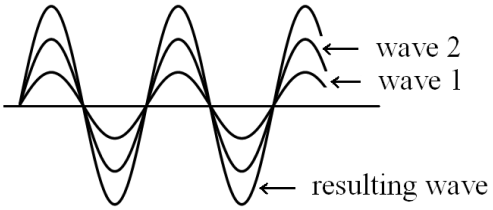
Note that the final image is inverted – this is not significant for an astronomical telescope. For a terrestrial telescope, where it is desirable to have an upright image, an additional lens is used to invert the image.

5.3. PHYSICAL OPTICS

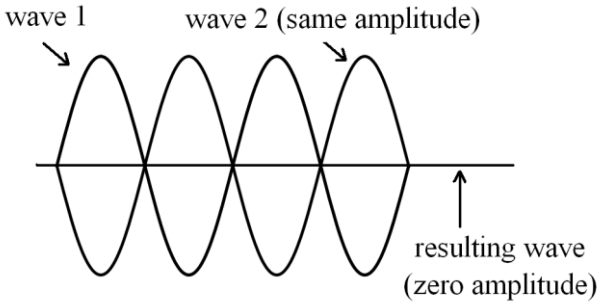
The phenomena considered in this section can be explained only if light is treated as a wave. These are interference, diffraction and polarization

5.3.1. Interference

Two waves of the same wavelength travelling in the same direction can interfere constructively (if they are in phase) or destructively (if they are exactly out of phase).



Constructive interference: waves 1 and 2, which are **in phase**, combine constructively to produce a resulting wave of the same phase but larger amplitude.



Destructive interference: waves 1 and 2, which have the same amplitude but are **exactly out of phase,** combine destructively to produce a wave with zero amplitude.

If two light waves interfere constructively at some points in space and destructively at others, bright and dark regions can be observed; these are called **interference fringes**.

Certain conditions must be satisfied if visible interference effects are to occur.

(1) The two light waves must originate from a **single** source of light.

(2) The path difference between the two interfering waves must not be too great.

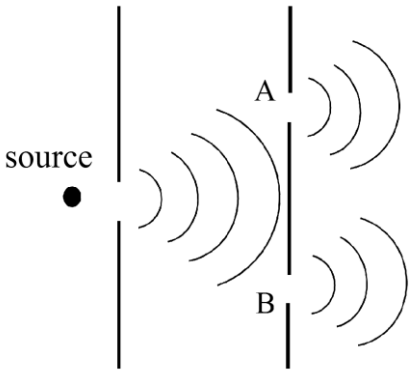
(3) The two waves must have identical wavelengths, and also equal amplitudes (otherwise complete destructive interference cannot occur).

The first condition results from the fact that light radiation is not emitted from a source as a continuous wave, but as a series of wave packets, each of approximate length 10^{-8} s. These wave packets all have the same wavelength and frequency, but differ randomly in phase.

To produce a stable interference pattern, the two travelling waves that interfere must maintain a phase difference between them that is constant in time. Such waves are said to be **coherent**.

If two different light sources are used to produce the two travelling waves, the light waves from one source are emitted independently of the waves from the other source; the phase difference between them cannot therefore be constant. Consequently the states of constructive or destructive interference will have durations of the order of 10^{-8} s; the eye cannot follow such rapid changes and no interference effects can be observed. Ordinary light sources are said to be **incoherent**.

One mechanism for producing two coherent light sources is shown in the diagram below (an alternative is to use a laser – a source that automatically produces coherent light).



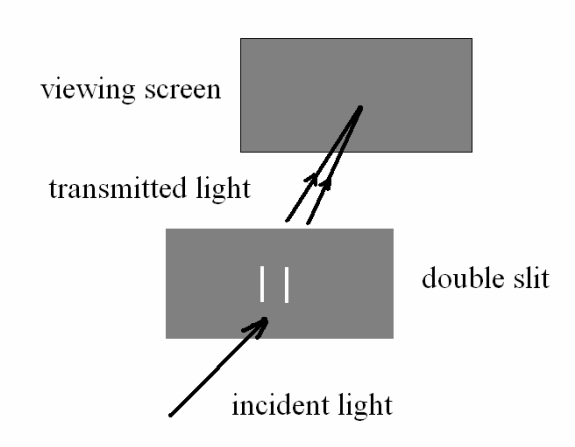
Light from a monochromatic source is collimated by passing it through a narrow slit and then allowed to fall on a screen containing two narrow slits.

The collimating slit creates a single wave-front that illuminates the slits A and B coherently at any instant.

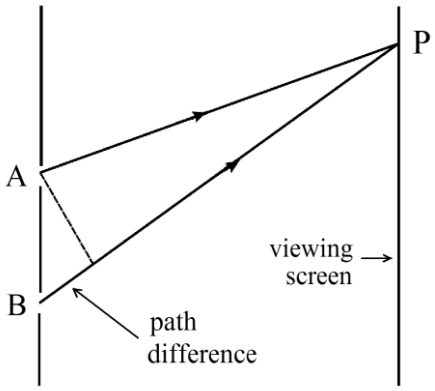
The light emerging from A and B is coherent because a single source produces the original light beam and the function of the two slits is simply to separate the original beam into two.

Any random phase change in the light emitted by the source occurs in both beams simultaneously, so that a constant phase difference is maintained.

The diagram below shows a schematic representation of the apparatus used to produce an interference pattern.



After passing through the slits A and B the two waves will produce an interference pattern that can be observed on a viewing screen placed behind the slits. The following diagram shows the apparatus viewed from above.



Two waves reaching a point P on the viewing screen will in general have travelled a different distance from the source; this is called the **path difference**.

If the times taken by the two beams to reach point P via slits A and B differ by more than about 10^{-8} s, they must originate from different wave packets.

The two beams will no longer be in phase at P – their phases will differ randomly and no interference pattern will be observed.

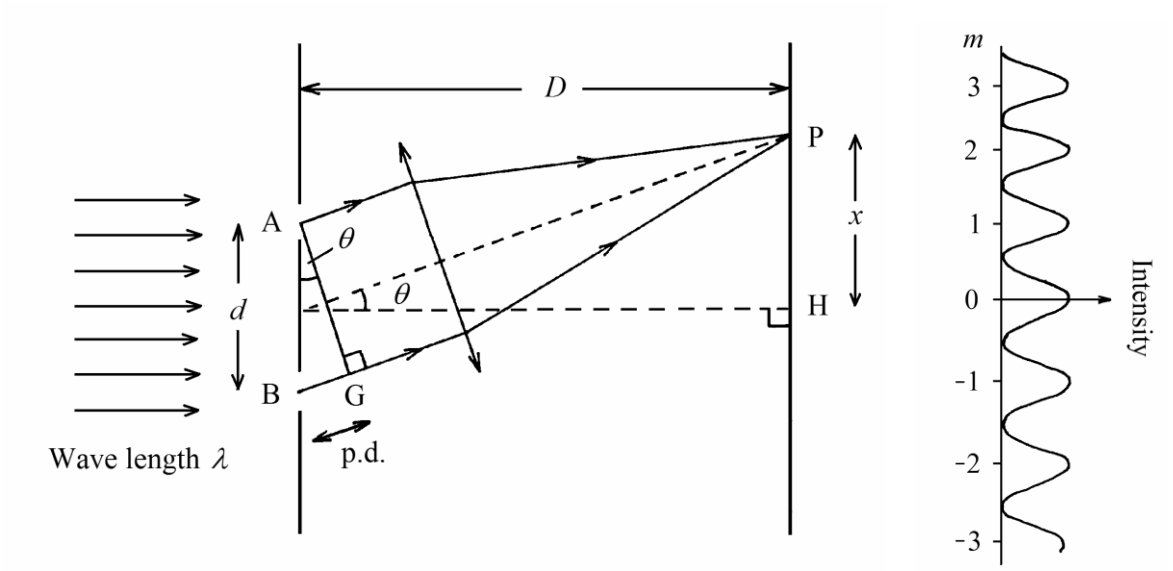
This means that the path difference must not be too large.

Double slit interference

Double slit interference was investigated by Thomas Young (1773–1829). His experiments in 1801 proved that light is a wave; all types of wave, including sound waves and water waves can undergo interference.

A parallel monochromatic beam of light falls normally on two parallel, narrow slits A and B a small distance *d* apart. The two beams emerging from the slits are in phase.

The transmitted light is focused on a screen a distance *D* behind the slits (with $D \gg d$) by means of a converging lens. Note that the diagram below is not drawn to scale.



Consider two rays passing through slits A and B which subsequently reach point P on the screen. The path difference p.d. for the two rays is, from the diagram,

$$\text{BG} = d\sin\theta\,.$$

If the position of P is such that the path difference is an integral number of wavelengths (i.e. 0, λ, 2λ, …), waves from A and B arrive at P **in phase** and will interfere **constructively** producing a **bright fringe**. For bright fringes, therefore, p.d. = $m\lambda$, m = ±1, ±2, ±3,… or

$$\boxed{d\sin\theta_m = m\lambda, \quad m = 0, \pm 1, \pm 2, \cdots} \quad \text{(two-slit interference, bright fringe)}$$

where *m* is called the order of the fringe.

If the position of P is such that the path difference is a half integral number of wavelengths (i.e. $\lambda/2, 3\lambda/2, 5\lambda/2, \cdots$), waves from A and B arrive at P **exactly out of phase** and will interfere **destructively**, producing a **dark fringe**.

The central position at H on the screen is bright, with path difference = 0, and as we go away from it on each side we get alternating dark and bright fringes, as shown in the intensity pattern above.

The distance from H to the *m*'th bright fringe is, from the diagram, $x_m = D\tan\theta_m$. If θ_m is small (less than about 10°) then $\sin\theta_m \approx \tan\theta_m$ so that $x_m = D\sin\theta_m$.

But $\sin\theta_m = m\lambda/d$ from the condition for constructive interference.

Therefore

$$\boxed{x_m = m\frac{\lambda D}{d}, \quad m = 0, \pm 1, \pm 2 \cdots}$$

The separation of two adjacent bright fringes is

$$x_m - x_{m-1} = \frac{m\lambda D}{d} - \frac{(m-1)\lambda D}{d}$$

leading to

$$\boxed{\text{Fringe separation } = \frac{\lambda D}{d}}$$

For small θ the fringes are therefore of equal width and equally spaced.

Note that:

- No energy has been lost in the dark fringes; it has been re-distributed to the bright fringes. The total intensity falling on the screen is the same as if there had been no interference.
- For the interference pattern to be visible, the fringe separation must not be too small. This requires that the inter-slit spacing *d* is not too large compared with the wavelength of the light. In practice, *d* will be a fraction of a millimetre, whereas the wavelength is of the order 500 nm.

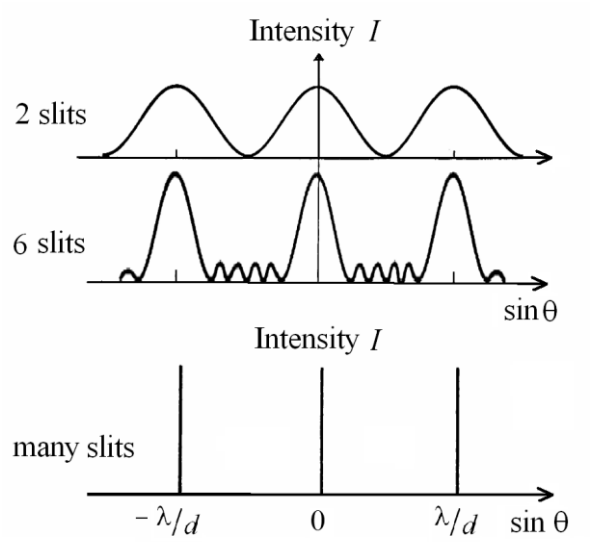
Diffraction grating

The diffraction grating is based on an extension of two-slit interference. It consists of a very large number of slits instead of only two. Gratings are often made by ruling very fine lines, called rulings, on a glass sheet (typically thousands of lines per cm); the untouched spaces between the lines act like slits.

The equations derived above for the double slit also apply in the case of the diffraction grating, with the slit spacing *d* being interpreted as the distance between adjacent slits. For the diffraction grating, the grating spacing is therefore

$$d = \frac{1}{\text{number of lines per unit length}}$$

The fringes produced by the grating are much sharper than those from two slits only and, because of the redistribution of energy, the intensity at the peaks is consequently much larger. The peaks are effectively extremely bright, well-separated lines.



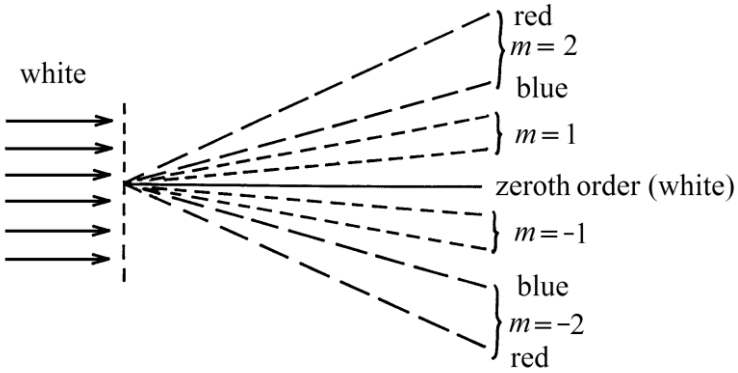
This is illustrated in the diagram, which shows that intensity pattern for two slits, six slits, and a very large number of slits. Note that since d is much smaller for a grating, the angular separation of the intensity maxima is in practice much larger (and the peaks are much more intense).

Diffraction gratings are widely used to determine the wavelengths of light emitted by sources of light ranging from lamps to stars: the wavelength can be deduced from the fringe separation using the formula derived above.

If white light is used instead of monochromatic light, a different value of θ satisfies the bright-fringe equation $d\sin\theta = m\lambda$ for each different wavelength (i.e. colour). Each bright fringe therefore becomes a spectrum.

However, at the centre of the pattern (where $m = 0$) the path difference is zero for all wavelengths, so all colours interfere constructively giving rise to a central white fringe.

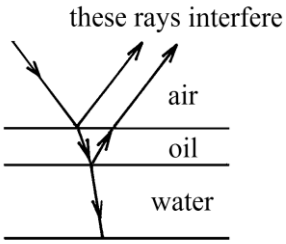
The resultant pattern will be a central white fringe bordered symmetrically on either side by differently spaced coloured fringes, i.e. a white-light spectrum is formed.



In a given order (i.e. m fixed) θ_m is larger for larger λ (from $d\sin\theta_m = m\lambda$). Therefore, red light is deviated more than blue (this is opposite to the spectrum produced by a prism). In fact the red end of one order can, under certain circumstances, overlap the blue end of the next (higher) order.

Thin-film interference

An interference pattern can also be produced if light is reflected from the two surfaces of a thin film of transparent material surrounded by a medium of different refractive index; the thickness of the film must be comparable with the wavelength of the light. Examples are soap bubbles and thin layers of oil or petrol on water.



The example of a thin oil film on a wet road is shown in the diagram.

Some light is reflected at each interface and the reflected waves interfere.

The path difference may be correct for the destructive interference of, say, red light and then the complementary colour, blue-green, will be seen. (Two colours are said to be complementary if when added together they produce white light.)

For other thicknesses of oil film, destructive interference for other colours will occur.

There are two factors (in addition to the path difference) that determine whether constructive or destructive interference will occur in thin-film interference.

(i) An electromagnetic wave travelling from one medium towards another medium of **higher** refractive index undergoes a **phase change** of 180° on reflection (this is equivalent to the wave being inverted, or adding a half wave length to the path length). No phase change occurs if a wave is reflected at the interface with a medium of lower refractive index.

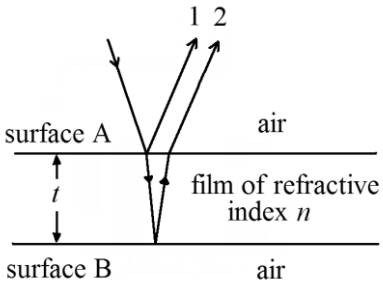
The *transmitted* light does not experience any phase change.

(ii) The wave length of light in a medium of refractive index n is reduced compared to the wave length in air by a factor n.

$$\lambda_n = \frac{\lambda}{n}$$

To illustrate this we consider the simple case of a film of transparent material of refractive index n surrounded by air. Monochromatic light is assumed.

The equations that will be derived assume that light is incident normally at interfaces; in the diagrams near-vertical rays are drawn, but they are shown separated for clarity.



Light reflected from surface A can interfere with light reflected from surface B.

The extra distance travelled by ray 2 compared with ray 1 is $2t$.

When reflected from surface A, ray 1 undergoes a phase change of π radians (or 180°) with respect to the incident ray, because the reflecting medium has a larger refractive index. Ray 2, which is reflected from surface B, undergoes no phase change since the reflecting medium has a smaller refractive index.

Therefore ray 2 will be 180° out of phase compared with ray 1; this is equivalent to adding $\lambda_n/2$ to the path length. The effective path difference between the two rays is therefore $2t + \lambda_n/2$.

For **destructive** interference, i.e. **dark** fringes, the path difference must be an odd number of half wavelengths (i.e. $\lambda_n/2,\ 3\lambda_n/2,\ 5\lambda_n/2, \cdots$).

Therefore we require:

$$2t + \lambda_n/2 = (m + \tfrac{1}{2})\lambda_n, \quad m = 0,1,2,\cdots$$

and it follows from $\lambda_n = \lambda/n$ that the condition on the film thickness *t* for dark fringes is:

$$\boxed{2nt = m\lambda, \quad m = 0,1,2.....} \qquad \text{(dark fringes)}$$

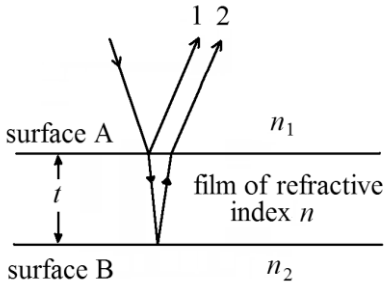
Similarly, for **constructive** interference the path difference must be an integral number of wave lengths, so that the condition for constructive interference is

$$\boxed{2nt = \left(m + 1/2\right)\lambda, \quad m = 0,1,2.....} \qquad \text{(bright fringes)}$$

Note that ray 2 travels a distance 2*t* in the medium of refractive index *n*; this distance is called the **geometrical path length**. The quantity 2*nt* that appears in these equations is called the **optical path length**; it is the distance that ray 1 travels in air in the same time interval.

- Clearly, for an interference **pattern** of alternating dark and bright fringes to be produced, there must be a variation in the film thickness (which can be produced in a number of ways).

A more general case is shown below; a thin film of material of refractive index *n* separates media of refractive indices n_1 and n_2.



The conditions for constructive and destructive interference depend on whether a phase change of 180° occurs at the interfaces, which in turn depends on the relative magnitudes of the refractive indices *n*, n_1 and n_2.

- $\lambda_n/2$ must be added to the geometrical path difference if $n_1 < n$ and a further $\lambda_n/2$ if $n < n_2$.

In summary:

- If there is a phase change at **either** interface (but not both), the conditions for dark and bright fringes are as derived above.
- If there is a phase change at **neither** interface or at **both** interfaces, the conditions must be **interchanged**.

The best approach to solving any problem involving thin-film interference is to derive the necessary condition, as is done above for the case of a thin film in air.

Other examples of thin-film interference are:

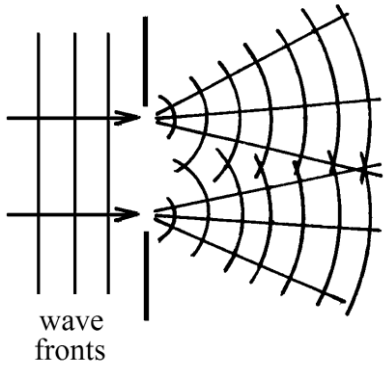
The colours in a soap bubble; gravity causes the film thickness to vary, so that an interference pattern is produced.

The colours in the feathers of some birds (especially peacocks) and in butterfly wings. Note that there are also other mechanisms producing colours in birds, e.g. pigments and light scattering.

Coated lenses. A thin layer of a material of the correct thickness and refractive index is coated onto the surface of a lens. Rays reflected from the air/coating and coating/lens interfaces interfere destructively, thereby greatly reducing the amount of light reflected and increasing the amount transmitted by the lens. This gives a brighter and sharper image.

5.3.2. Diffraction

An interference pattern can also be produced when light passes through a single slit or aperture (or when light passes around an obstacle) provided the size of the slit is not too large compared with the wavelength of the light.



Consider light passing through a narrow slit.

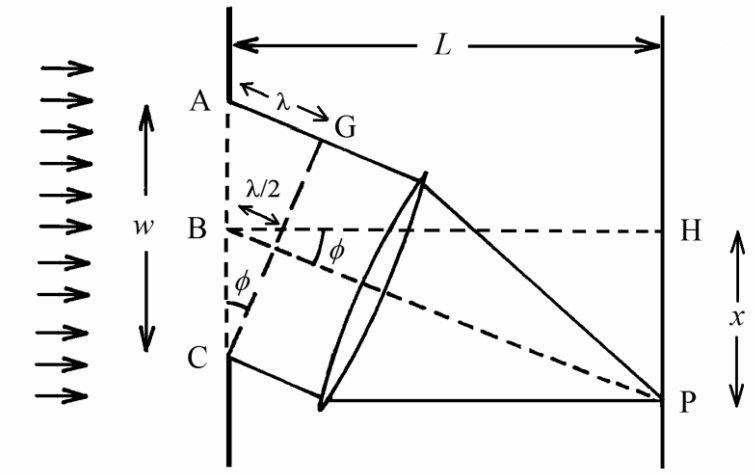
According to Huygens' principle, each point across the slit can be regarded as a source of secondary wavelets (two such wavelets are shown in the diagram).

These secondary wavelets spread out in the region behind the slit. This spreading out after passing through a slit is called **diffraction**.

A diffraction pattern is created by interference between wavelets from different points along the slit.

Single-slit diffraction

Consider parallel monochromatic light falling normally on a narrow parallel-sided slit of width w. The diffracted light is focused by a lens on a screen a distance L behind the slit, with $L \gg w$.



Note that this figure is not drawn to scale.

It is assumed that each point between A and C is a source of secondary wavelets (Huygens' principle).

At H, the point on the screen opposite the centre of the slit, all wavelets arrive **in phase** ($L \gg w$, and so all the wavelets travel the same distance) and interfere constructively to give a bright fringe.

Now consider a point P on the screen such that PA – PC = λ, or

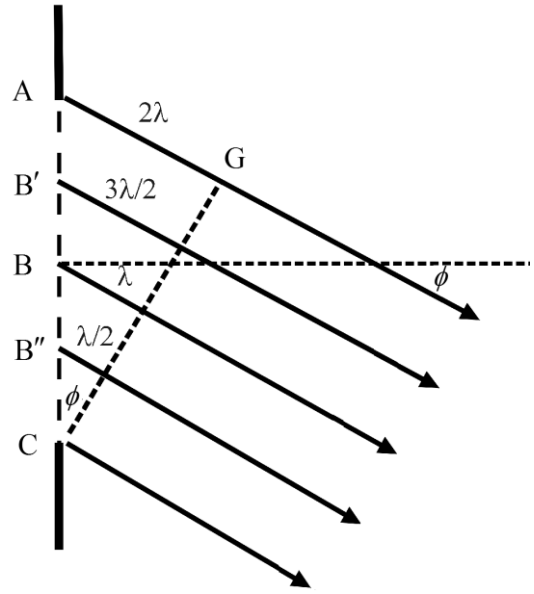
$$\sin\phi = \frac{\text{AG}}{\text{AC}} = \frac{\lambda}{w}.$$

Then PB – PC = $\lambda/2$, and so wavelets from B and C interfere destructively at P, giving zero intensity at P.

The two secondary wavelets from just above B and C will also interfere destructively at P. The process is repeated for successive pairs of secondary wavelets, until A and B are reached; each pair will interfere destructively, and so the total intensity at P will be zero.

Hence, there is a **dark fringe** for $\sin\phi = \frac{\lambda}{w}$. If ϕ is made smaller or bigger by a small amount then wavelets cannot be paired off and the resulting intensity will no longer be zero.

We now consider a point P such that PA – PC = 2λ, equivalent to $\sin\phi = \frac{2\lambda}{w}$. Rays leaving the slit are shown enlarged in the diagram below.



A dark fringe is also formed in this case.

This can be seen by taking two pairs of waves from B" and C and from B' and B, respectively.

The path difference for each pair is $\lambda/2$ and the same argument as before is used.

Thus there is a dark fringe for $\sin\phi = \frac{2\lambda}{w}$.

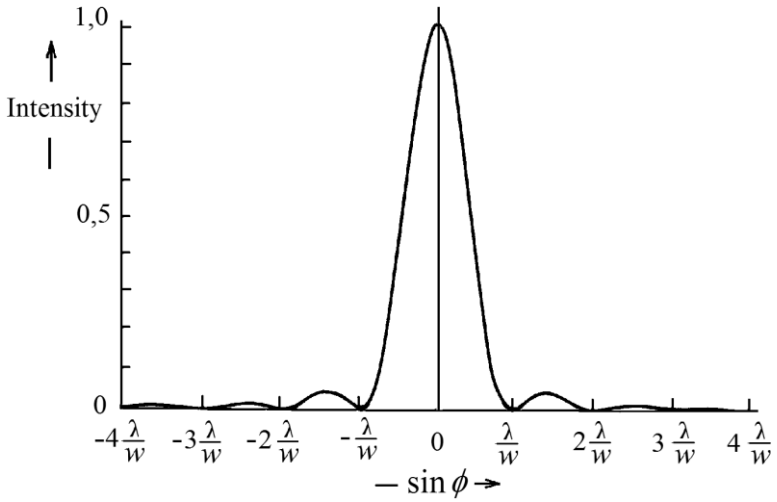
This will happen again for $\sin\phi = \frac{3\lambda}{w}$ etc.

The condition for **dark** fringes is, therefore,

$$\boxed{\sin\phi_m = \frac{m\lambda}{w}, \quad m = \pm1, \pm2, \pm3 \ldots \text{ (not } m = 0)} \quad \text{(single-slit diffraction, dark fringes)}$$

- The equation for dark fringes in the single-slit case is very like the one for bright fringes in the double-slit case, and so care must be taken not to confuse them.

In between the dark fringes are bright fringes, with maxima *approximately* midway between the dark fringes.



Note that the central maximum is twice as wide as the others and considerably brighter. These features are shown in the diagram above.

If ϕ_m is small then

$$\sin\phi_m \approx \tan\phi_m = \frac{x_m}{L}$$

and the distance to the *m*'th dark fringe is given by

$$\boxed{x_m = m\lambda\frac{L}{w}, \quad m = \pm1, \pm2, \pm3....}$$

Note the following:

- The diffraction caused by an **obstacle** in the path of light is the same as produced by an aperture of the same shape and dimensions.
- A **circular** aperture or obstacle (such as a lens in an optical instrument) gives a diffraction pattern of concentric circular fringes, where the angular deviation of the first-order dark fringe is given by

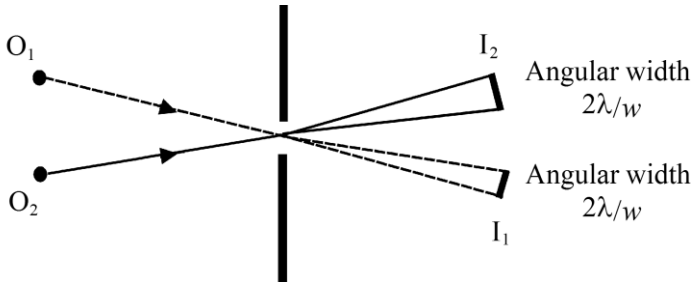
$$\boxed{\sin\phi = 1,22\frac{\lambda}{w}}$$

- Whenever a double slit is used to produce an interference pattern, then the single-slit diffraction pattern (produced by either slit) will be superimposed on the double-slit interference pattern discussed earlier.
  The single-slit pattern will be wider than the double-slit one (since we must have $w < d$) and may cause missing orders in the double-slit pattern if a maximum of one pattern coincides with a minimum of the other.

Resolving power and Rayleigh's criterion

When light passes through a small aperture, it is diffracted and spread out. The smaller the aperture the greater the spread, since $\sin\phi \propto \lambda/w$. A measure of the angular width of the image is $2\lambda/w$, the width of the central maximum of the interference pattern (remember that most of the light intensity is in the central maximum – see the previous diagram).

If there are two incoherent sources of light (i.e. two objects), each will produce its own diffracted, spread-out image.

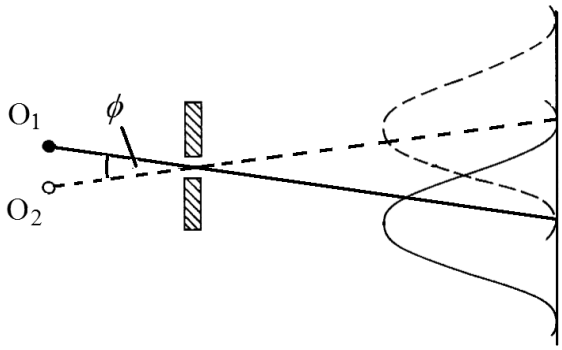


An important property of any optical instrument, including the eye, is its ability to separate two images of this kind. This is the **resolving power** of the instrument or eye.

If the two sources are too close to each other, their images will overlap to such an extent that they cannot be resolved. It is resolution that sets an upper limit on the magnifying power of microscopes and other optical instruments.

Rayleigh's Criterion provides a way of estimating the resolving power of an optical instrument. Two objects are said to be just resolved if the central maximum of the diffraction pattern due to one coincides with the first minimum of the diffraction pattern due to the other (and vice-versa).

This is illustrated in the following diagrams. Two point sources of light at O_1 and O_2 each produce a diffraction pattern on a screen behind a single slit.

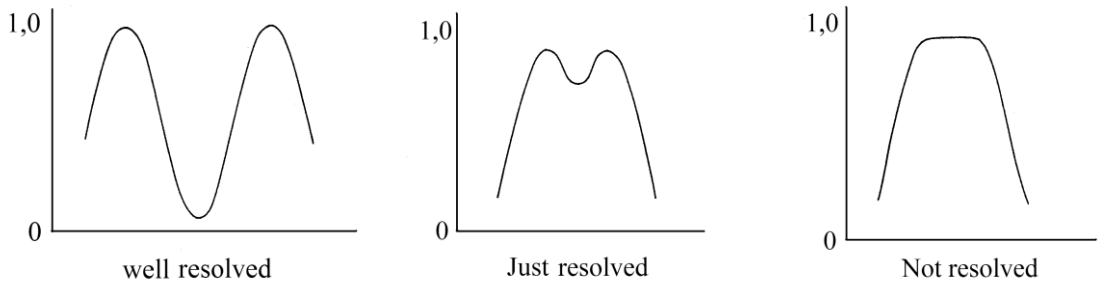


The solid curve shows the central maximum due to the source at O_1 and the dashed curve the central maximum of the diffraction pattern due to the source at O_2.

The first minimum of one pattern coincides with the central maximum of the other.

The angle ϕ in this situation is such that $\sin\phi = \frac{\lambda}{w}$ for a slit of width w or $\sin\phi = 1,22\frac{\lambda}{w}$ for a circular aperture of diameter w.

Plotting the total intensity for the two sources as a function of position on the screen (or the retina of the eye) gives the following curves.



$\sin\phi > \lambda/w$ $\sin\phi = \lambda/w$ $\sin\phi < \lambda/w$

The conditions indicated refer to a rectangular slit; for a circular aperture a factor of 1,22 must be included.

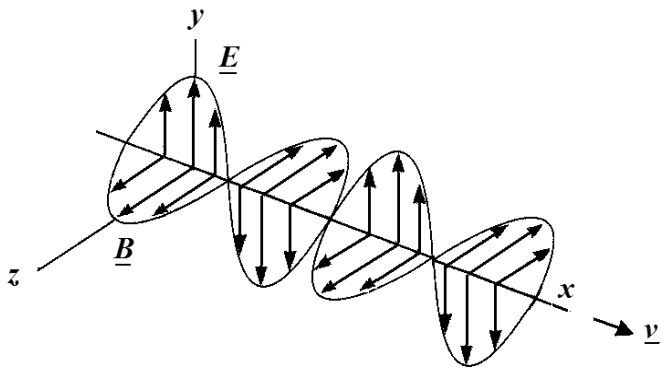
5.3.3. Polarization

The polarization of electromagnetic waves

The electric and a magnetic vectors associated with an electromagnetic wave, including light, are at right angles to each other and also to the direction of wave propagation.

The polarization of light, which is a phenomenon that can be explained only if light is a transverse wave, is associated with the electric vector (i.e. the magnetic vector is irrelevant as far as polarization is concerned).

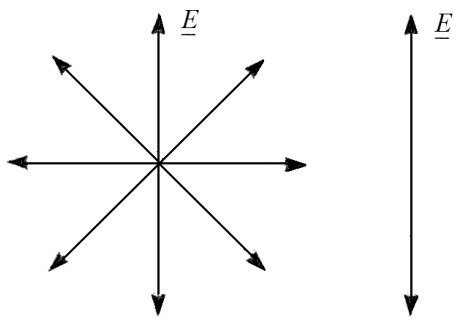
A wave is said to be **plane polarized** (or **linearly polarized**, or simply **polarized**) if the electric field vibrates in the same direction at all times at a particular point in space.



In the diagram above, with the electric field pointing in the *y* direction and the velocity vector in the *x* direction, the wave is said to be linearly polarized in the *y* direction. The *xy* plane, which is formed by the electric field vector and the direction of propagation of the wave, is called the **plane of polarization**.

Light from a normal source is emitted in wave packets; the plane of vibration of the electric vector in the waves varies randomly from one wave packet to the next, so normally light is **unpolarized**.

In the diagram below a light beam is viewed along the direction of propagation, perpendicular to the page.



On the left is a schematic representation of an unpolarized light beam; the electric vector can vibrate in any direction with equal probability.

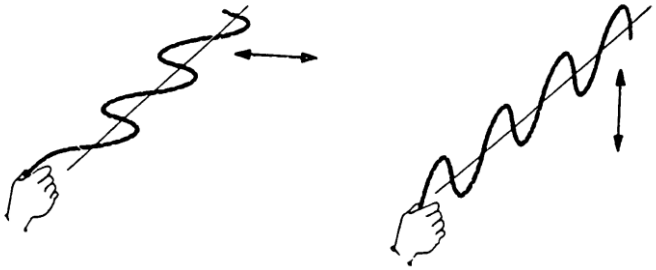
On the right is a plane polarized light beam with the electric vector vibrating only in the vertical direction.

Polarized light can be produced in several ways, which include by absorption, by reflection and by scattering, each of which we consider below.

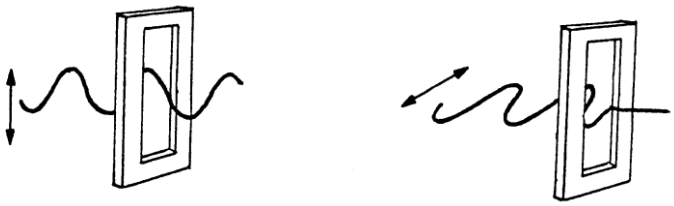
Polarization by absorption

This is the most common technique for polarizing light.

Consider the analogy of waves being sent down a rope. The vibrations shown in the diagram are in either the horizontal or the vertical plane; in each case the wave is plane-polarized since the oscillations are confined to one plane only.



If a sufficiently narrow slit is placed in the path of the wave, the wave will pass through it if its direction of vibration is parallel to the slit, or it will be stopped completely if its direction of vibration is at right angles to the slit.

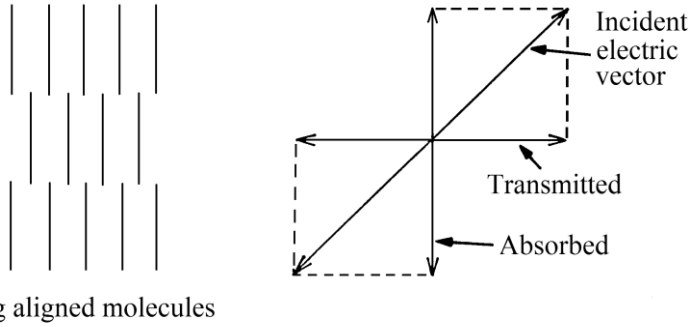


Note that the vibrations of a **longitudinal** wave are **along** the direction of propagation and so no orientation of the slit would affect them.

Certain types of material behave in a similar way as far as the **electric** vector of a light wave is concerned. Examples of such materials are the naturally occurring crystal, tourmaline, and the artificially prepared Polaroid sheet (invented by E H Land in 1932).

The discovery that light can be polarized in this way provided the evidence for light being a **transverse** wave.

The material from which Polaroid sheets are made is fabricated in thin sheets of long-chain hydrocarbons; these are stretched during the manufacturing process so that the molecules are aligned.

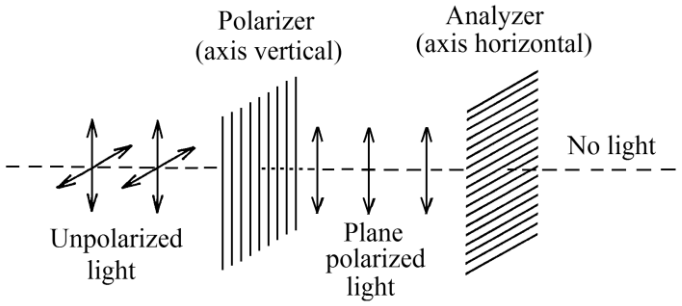


Long aligned molecules

The electric vector of light passing through a sheet can be resolved into components parallel and perpendicular to the aligned molecules. The parallel component causes free electrons in a molecule to oscillate along its length. The molecules therefore readily absorb from the light those components parallel to their length and transmit components perpendicular to their length. For this reason, the direction perpendicular to the molecular chains is referred to as the **transmission axis**.

The transmitted intensity is less than the incident intensity since some light has been absorbed, but now the electric vectors in all the wave packets emerging from the sheet vibrate in one plane only; the light is plane-polarized.

A sheet of Polaroid can be used as a polarizer (to produce plane polarized light) or as an analyzer (to determine the plane of polarization of polarized light).



- If the two sheets of Polaroid in the diagram above have their transmission axes parallel, the light transmitted by the polarizer will also be transmitted by the analyzer.
- If however the two sheets of Polaroid have their axes at right angles (called crossed Polaroids), as shown in the diagram, then no light is transmitted.
- Some materials, for example sugars, cause the plane of polarisation to be rotated when light passes through them; such materials are called **optically active**.

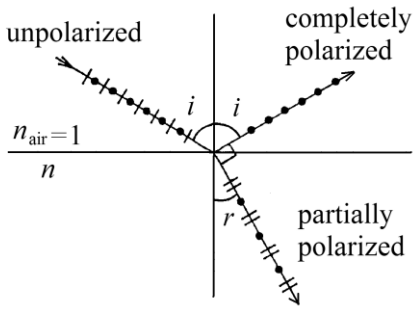
  The angle through which the plane is rotated can be determined by placing a sample of material between the polarizer and analyzer and measuring the angle through which the polarizer must be rotated to again cut out all transmitted light.

Polarization by reflection

Light incident on a transparent medium is partially reflected and partially transmitted. If the medium is a dielectric (including glass and water), the reflected and refracted rays are partially plane polarized, the degree of polarization depending on the angle of incidence.

- Light is said to be **partially plane polarized** when the electric fields oscillating along one direction have greater amplitudes than those oscillating along other directions.

For the special case when the reflected and refracted rays are at 90° to each other, the **reflected ray** is **completely polarized**. This is illustrated in the following diagram, where the dots and lines attached to the rays indicate the direction of vibration of the electric vector.



- The reflected light is completely polarized, with the electric field vector parallel to the surface.
- The transmitted light is partially polarized, with the electric field having greater amplitude in the plane of the diagram.

We determine the condition for this to happen as follows. As shown in the diagram above, unpolarized light is incident on the material at angle of incidence *i*. Since the reflected and refracted rays are perpendicular

$$i + r = 90° \quad \rightarrow \quad r = 90° - i\,.$$

From Snell's law at the interface,

$$\sin i = n \sin r$$

where we have used $n_{air} = 1$. Hence

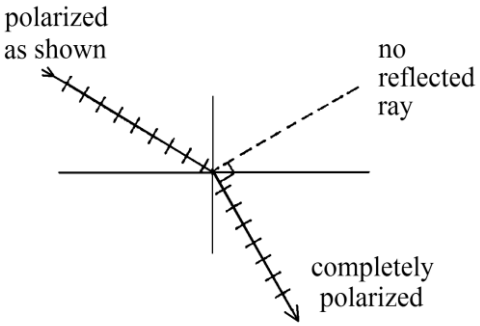
$$n = \frac{\sin i}{\sin(90° - i)} = \frac{\sin i}{\cos i} = \tan i$$

Complete polarization of the reflected ray therefore occurs when

$$\boxed{\tan i_B = n} \qquad \text{(Brewster's angle)}$$

where i_B is the **Brewster angle**, named after its discoverer Sir David Brewster (1781–1868).

If the light incident at the Brewster angle is itself polarized with the electric field oscillating in the plane of the diagram, all the light is transmitted and there is no reflected light.



Polarization by reflection occurs frequently in nature when, for example, sunlight is reflected from a horizontal surface covered by water or snow. The reflected electric vector has a large horizontal component. Sunglasses made from polarizing material can be used to reduce the glare from the reflected light; the transmission axis of the lenses must be vertical to absorb the horizontal component of the reflected light.

Polarization by scattering

Light scattered by particles small compared with the wavelength of the light (e.g. sunlight scattered by air molecules and dust particles in the atmosphere) is partially plane polarized.

Bees and homing pigeons are believed to detect this polarized light and use it for navigation.

The scattered intensity increases with increasing frequency of light, $I \propto f^4$, and so more blue light than red light is scattered (this is Rayleigh scattering). This accounts for the blue of the sky and the red of sunsets.

For information about the scattering process and how it polarizes light, see the textbook.

OPTICS

LECTURE EXAMPLES

University of the Witwatersrand, Johannesburg

School of Physics

Physics I D (PHYS1001/6)

Lecture Examples 5.1 – [2018] **Geometrical Optics I**

1. A pole 3,0 m long stands vertically on the horizontal floor of a pool containing water to a depth of 1,5 m. Rays from the sun make an angle of 40° with the pole.
   Calculate the length of the shadow of the pole on the floor of the pool. [2,09 m]

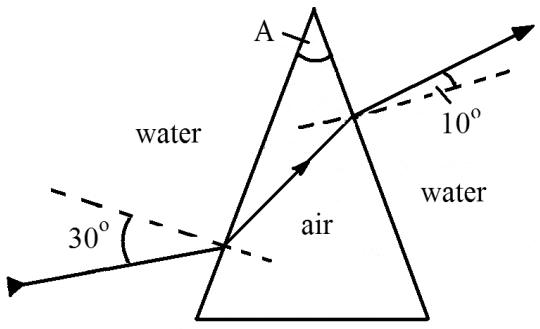
   Hint: consider the path of the ray from the sun that just skims the top of the pole, before passing through the air/water interface to reach the floor of the pool.

2. A submarine is below the surface in clear water and a helicopter is vertically above it. Which one of the following statements is true?

   (a) As the helicopter descends the submarine seems to come closer to the surface.
   (b) To an observer on the submarine the helicopter seems to be closer than it really is.
   (c) If the helicopter flies horizontally total internal reflection will never prevent it from losing sight of the submarine.
   (d) When the submarine dives the critical angle for total internal reflection decreases and it will therefore be less easy to spot.
   (e) If the submarine moves horizontally, because of total internal reflection it will eventually lose sight of the helicopter.

   [c]

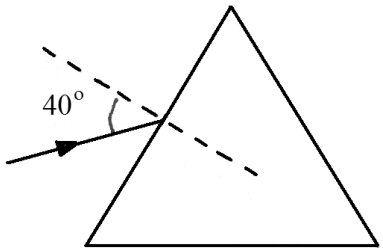
3. A hollow prism, containing air and made from parallel-sided sheets of glass, is immersed in water. Light incident at an angle of 30° emerges at an angle of 10° as shown. The dashed lines are perpendicular to the surfaces.
   Calculate the refracting angle *A* of the prism if the refractive index of water is 1,33. [28,3°]



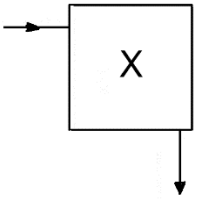
   Hint: first use Snell's law to determine the angle of refraction at the first interface and then the angle of incidence at the second interface. Then relate these angles geometrically to the refracting angle *A*.

4. A ray strikes one face of an equilateral prism in air at an angle of incidence of 40,0°. The refractive index of the prism material is 1,80. Determine the subsequent path of the ray until it emerges from the prism, showing all relevant angles on your sketch.

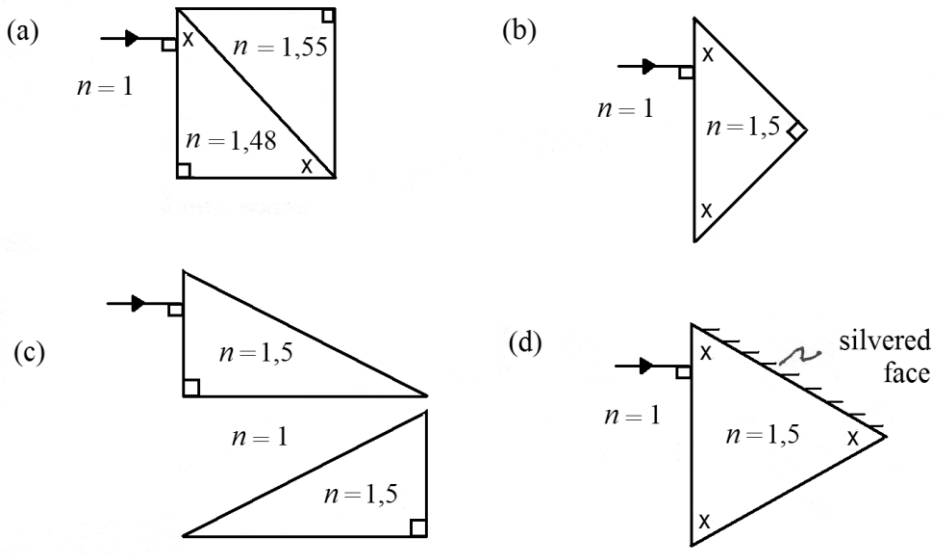
[Emerges at 40° to normal at lower interface]



5. When a beam of light is incident on X from the left, some light emerges as shown at 90° to the incident beam.



Which one of the following could X represent (x indicates angles that are equal within a given triangle)? [a]



6. A beam of light strikes a plane block of glass at an angle of 50° to the normal. The beam contains two wavelengths of 500 nm and 700 nm. The wavelengths in the glass are found to be 338 nm and 476 nm respectively.

What is the angle between the refracted rays? [0,19°]

Hint: first use the wavelengths given to determine the refractive index of glass for each of the two beams into which the light beam splits.

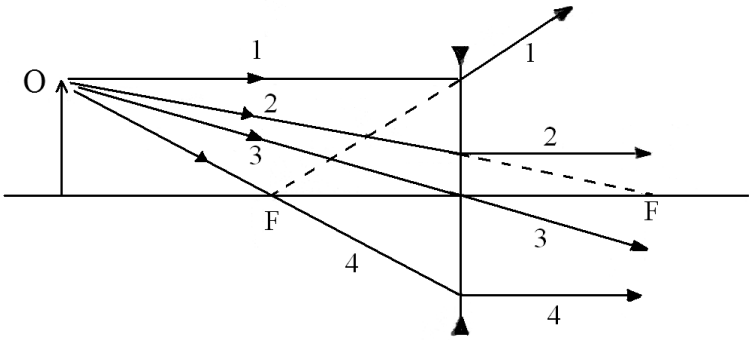
University of the Witwatersrand, Johannesburg

School of Physics

Physics I D (PHYS1001/6)

Lecture Examples 5.2 – [2018] **Geometrical Optics II**

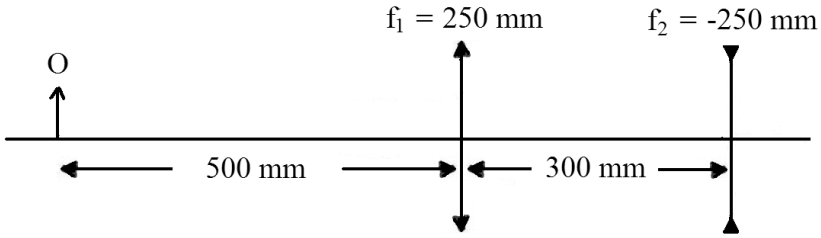
1. In the diagram below which ray(s) is/are not correctly drawn, or are they all correctly drawn? [4]



2. In a drive-in cinema the screen is situated 100 m from the film projector. The image on the screen measures 6,0 m × 4,0 m.
Calculate the focal length of the projection lens required if the frames on the film measure 72 mm × 48 mm. [1,17 m]
Hint: what magnification must be produced by the lens?

3. A converging lens of focal length 24 mm is used to examine a biological specimen on a microscope slide. The lens forms an image of the sample 126 mm from the lens.
How far is the lens from the sample if the image is (i) real and (ii) virtual, and what is the magnification in each case? [29,6 mm, –4,25 (inverted); 20,2 mm, +6,25 (upright)]

4. A converging lens of focal length 250 mm is placed 300 mm in front of a diverging lens of focal length –250 mm. An object of height 5 mm stands 500 mm in front of the converging lens, as shown below.



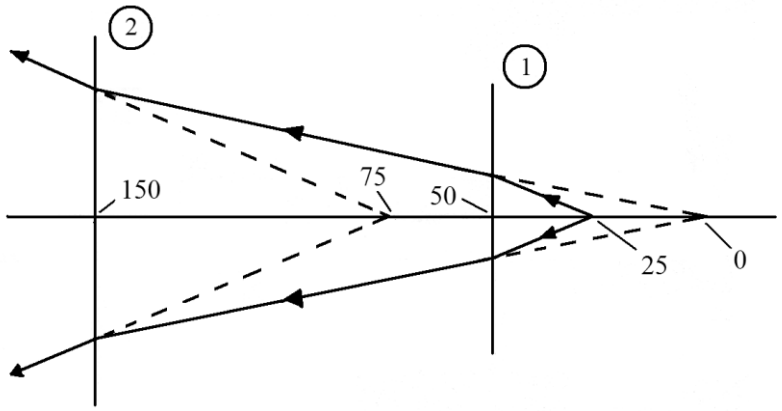
(i) What is the position of the final image? Is it real or virtual?
[1,00 m behind the diverging lens; real]

Hint: first calculate the position of the image formed by the first lens. Where is it relative to the second lens?

(ii) What is the height of the final image? Is the image upright or inverted?
[–25 mm, inverted]

5. A telephoto lens combination consists of a converging lens of focal length 300 mm and a diverging lens of focal length –100 mm, the separation between the lenses being 275 mm. The diverging lens is between the converging lens and the film.
Where should the film be placed in order to photograph an object which is 10 m in front of the converging lens? [52,2 mm behind the diverging lens]

6. The diagram shows two lenses mounted next to a metre stick. The numbers give the positions of the indicated points in mm. Subscripts 1 and 2 refer to objects and images for lenses 1 and 2 respectively.



(i) Which one of the following is correct? [d]

(a) $u_1 = +25$ $v_1 = +50$ $u_2 = +25$ $v_2 = -150$
(b) $u_1 = -25$ $v_1 = +150$ $u_2 = -25$ $v_2 = +150$
(c) $u_1 = +25$ $v_1 = -50$ $u_2 = -150$ $v_2 = -75$
(d) $u_1 = +25$ $v_1 = -50$ $u_2 = +150$ $v_2 = -75$
(e) $u_1 = +50$ $v_1 = -25$ $u_2 = +125$ $v_2 = +750$

(ii) Which one of the following is correct? [a]

(a) $f_1 > 0$ $f_2 < 0$ (b) $f_1 < 0$ $f_2 > 0$
(c) $f_1 < 0$ $f_2 < 0$ (d) $f_1 > 0$ $f_2 > 0$
(e) It is impossible to determine the type of lens from the data given.

7. A near-sighted man cannot clearly see objects more than 2,0 m away.
(i) What power spectacles (assumed to be in contact with his eyes) does he need to see distant objects? [–0,50 dioptre]
(ii) If his near point without spectacles is 100 mm, what is it with them? [105 mm]

Hint: the spectacle lens must take a real object placed at his new near point and produce an image at the eye's own (unaided) near point, which the eye then focuses on.

University of the Witwatersrand, Johannesburg

School of Physics

Physics I D (PHYS1001/6)

Lecture Example 5.3 – [2018] **Physical Optics**

1. A double-slit interferometer has a screen 1,0 m from the slits, and is illuminated by a parallel beam of normally-incident light with a wavelength range 450 mm $\leq \lambda \leq$ 600 mm.

   (i) Is the red end of the spectrum of a given order further from the zero-order line than the blue end? [Yes]

   (ii) The interference pattern on the screen is such that 3,0 mm from the zero-order line the red end of one order is found to be coincident with the blue end of the next order. What is the separation of the slits? [0,60 mm]

2. A pair of narrow slits is illuminated normally with light of wavelength 640 nm. When a piece of transparent material 12 μm thick is placed opposite one slit the third-order bright image is found where the ninth order used to be.

   Calculate the refractive index of the material. [1,32]

   Hint: first write down the expression for the angular position of the 9th order image. Then derive the condition for the angular position of the 3rd order image when the slab of transparent material has been inserted (remember that the slab changes the optical path difference).

3. A pair of narrow slits is illuminated normally with monochromatic light and its interference fringes observed on a screen (with air in the gap between the screen and slits).

   Water (of refractive index 4/3) now fills the whole space between slits and screen. What order bright image will now fall where the ninth order used to be? [12th]

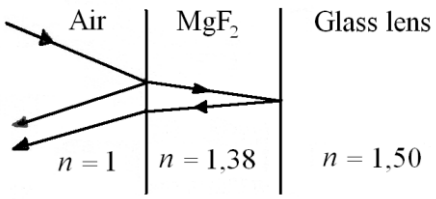
   Note: the answer does not depend on the slit separation or wavelength of the incident light; use symbols for these quantities.

4. A diffraction grating is ruled with 400 lines per mm, and is illuminated with light from atomic hydrogen at normal incidence. The α and δ lines of hydrogen have wavelengths of 656 nm and 410 nm respectively.

   (i) Calculate the angular separation, in the second-order spectrum, between the α and δ lines. [12,5°]

   (ii) What is the highest order that is possible for each of these lines? [α: 3, δ: 6]
   Hint: $\sin\theta$ cannot exceed 1.

5. In order to reduce reflection from an optical surface, such as the surface of a lens, the surface is often coated with a thin film of MgF_2 (of refractive index 1,38).



(i) Determine the minimum thickness of a coating that will minimize reflection for normal incidence of light near the middle of the optical spectrum, with $\lambda = 552$ nm. [100 nm]

Hint: you must first decide what the condition for destructive interference is for the combination of refractive indices shown in the diagram above.

(ii) What wavelength of light nearest 552 nm will undergo a maximum reflection for this thickness of coating? [276 nm]

6. Light of wavelength 560 nm is incident normally on a single slit, producing a diffraction pattern on a screen 1,00 m behind the slit.

When the separation of the slit and screen is increased to 1,25 m and the fringe pattern is refocused, it is found that the third-order minimum has moved 5,0 mm from its previous position on the screen.

Calculate the width of the slit. [84 μm]

7. The rails on railways in South Africa are about 1,1 m apart. How high can a plane be above the railway before the pilot can no longer resolve the rails?

Assume that the pupil of the eye has a diameter of 3,0 mm and that the wavelength of the light is 550 nm. [4,9 km]

PHYS 1001/1006 TUTORIALS
YEAR 2018
4th BLOCK

TUTORIALS TO PREPARE

WEEK: 10 September: **No Tutorials converted to lecture**
WEEK: 17 September: **Geometrical Optics I**
WEEK: 24 September: **Geometrical Optics II**

- All students are expected to prepare solutions to these questions listed above prior to attending the tutorial session. Tutors will do a quick check. Preparation of tutorials will be recorded in the class registers as Not Prepared (NP), Partially Prepared (PP) and Well Prepared (WP).

- Students should form mini-groups of 3 and discussion within the groups should commence immediately on arrival at the tutorial venue. This discussion should be for a minimum of 20 minutes. Tutors will float around groups and assist students and will only give feedback on questions to the entire class that he/she might find as problematic. Note tutors are expected to give an overview of the solution and not complete solutions. Any problems not addressed will be carried over to the next week for discussion.

- A tutorial test will then be given at the end of each session (approximately 10 minutes). Tutors are expected to give feedback on the tutorial test at the beginning of the next session.

PHYS1001/1006 Course Co-ordinator

University of the Witwatersrand, Johannesburg

School of Physics

PHYS1001/6 (Physics I D)

Tutorial 5.1 – [2018] **Geometrical Optics I**

Question 1

The velocity of light in a vacuum can be determined by measuring the change in wavelength, from λ_1 to λ_2, when light passes from a medium of refractive index n_1 to one of refractive index n_2.

Derive an equation for the velocity of light in a vacuum, c, in terms of these quantities and the frequency f of the light.

Question 2

A man whose eyes are 1,8 m above the ground stands 2,4 m from the edge of a brim-full swimming bath 2,0 m deep.

How far from the wall nearest him is a point on the bottom of the swimming bath that he can just see? Refractive index of water = 1,33.

[1.51 m]

Question 3

A small object is placed at the bottom of a tank of water which is 200 mm deep.

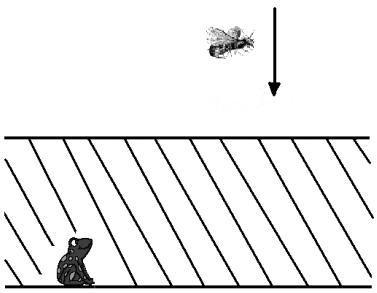
(i) Calculate its apparent depth below the water surface.

(ii) A block of flint glass 50 mm thick is placed over the object in the water so that the surface of the water is 150 mm above the top of the block. Calculate the apparent depth of the object below the water surface.

Refractive index of flint glass = 1,67. Refractive index of water = 1,33.

[(i) 150 mm (ii) 143 mm]

Question 4



A frog underneath the surface of a pond sees a fly approach it from above the surface.

Which of the following statements is/are correct?

(i) Because of total internal reflection the frog cannot see the fly until the fly gets to a certain minimum distance above the water surface.

(ii) To the frog the fly appears to be higher above the water surface than it really is.

(iii) To the fly the frog appears to be further below the water surface than it really is.

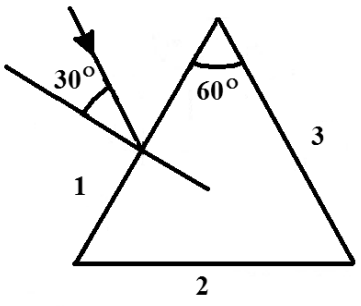
(iv) Because of total internal reflection the fly cannot see the frog if the fly is less than a certain distance above the water surface.

(v) The path of the rays from the frog to the fly depends on the colour of the frog.

[(ii), (v)]

Question 5

A ray is incident at 30° to the normal of one face of an equilateral prism made of glass of refractive index 1,60, as shown in the diagram.



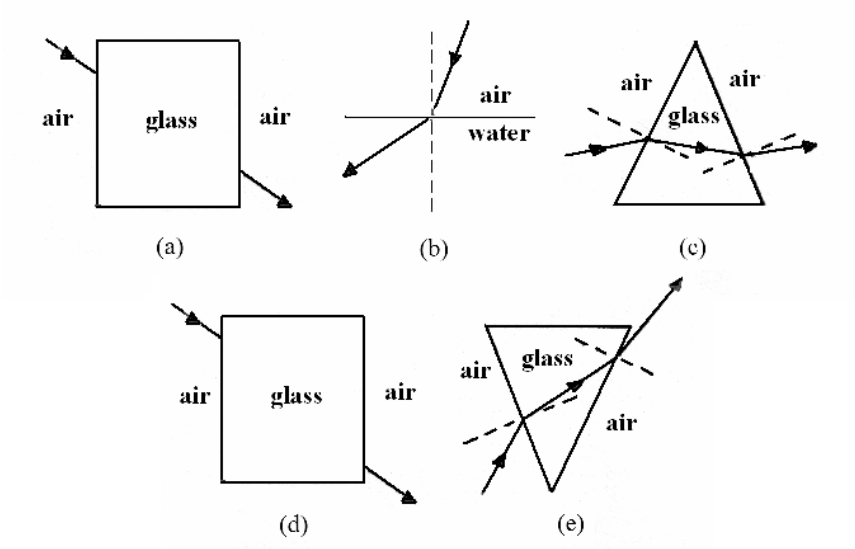
(i) Determine by calculation whether the ray will be totally-internally reflected at face 2.

(ii) Calculate what angle the ray will make with the normal when it emerges from face 3.

[(i) It will be (ii) 30°]

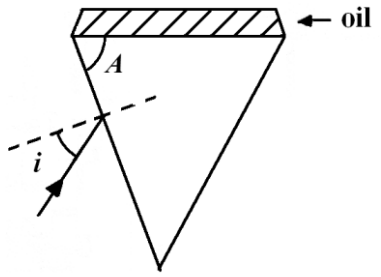
Question 6

Which one of the diagrams below is correct?



[e]

Question 7



Light is incident on one face of a glass prism in air, as shown in the diagram. The refractive index of the prism is 1,60. The top face of the prism is covered with a parallel-sided layer of oil.

(i) Calculate the refracting angle *A* of the prism if the light is just totally reflected at the oil/air interface when the angle of incidence *i* is 70°.

(ii) Calculate the refractive index of the oil if the light is just totally reflected at the glass/oil interface when *i* is 13°.

[(i) 74.6^{o} (ii) 1.47]

Question 8

(i) For a prism with a refracting angle of 60° the angle of minimum deviation is 37,18°. Calculate the refractive index of the prism material.

(ii) On either side of the position for minimum deviation there are two values of the angle of incidence for which the angles of deviation are equal. For this prism, calculate the deviation for an angle of incidence of 63,46°, and determine the other value of the angle of incidence which gives the same deviation.

[(i) 1,50 (ii) 40.00^{o} , 36.54^{o}]

University of the Witwatersrand, Johannesburg

School of Physics

PHYS1001/6 (Physics I D)

Tutorial 5.2 – [2018] **Geometrical Optics II**

Question 1

A transparency measuring 24 mm × 36 mm is to be projected so that the image is 1,2 m × 1,8 m. The projector lens has a focal length of 50 mm.

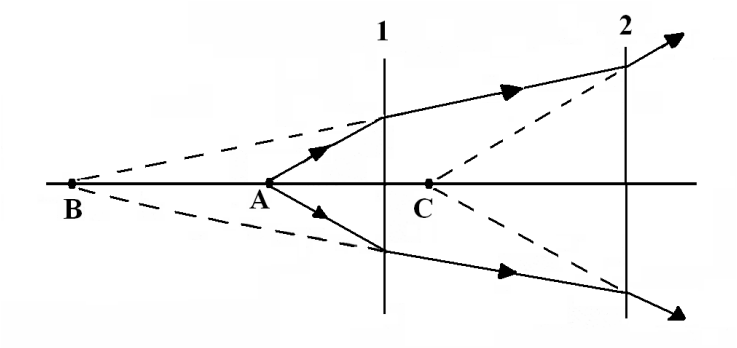
(i) How far from the lens must the slide be placed?

(ii) How far from the lens must the screen be placed?

[(i) 51 mm (ii) 2.55 m]

Question 2

The diagram shows rays passing through two lenses 1 and 2.



(i) Determine whether:

A is a real or virtual object for lens 1;

B is a real or virtual image from lens 1;

B is a real or virtual object for lens 2;

C is a real or virtual image from lens 2.

(ii) Determine whether each lens is converging or diverging.

Question 3

A telephoto lens for a camera consists of a converging lens of focal length 50 mm and a diverging lens of focal length −50 mm placed 20 mm apart, with the diverging lens nearer the film.

(i) Calculate the distance from the diverging lens to the film if a bird 2 m from the converging lens is in sharp focus on the film.

(ii) Calculate the height of the image if the bird is 300 mm high.

[(i) 83.3 mm (ii) -20.6 mm (inverted)]

Question 4

A camera has a lens of focal length $f_1 = +50$ mm.

(i) How large is the image formed by this lens of an object 10 mm high situated 250 mm in front of the lens?

(ii) A diverging lens, with a focal length $f_2 = -50$ mm, is now placed 25 mm in front of the converging lens. How large is the new final image if the object is 250 mm in front of the diverging lens?

[(i) -2.5 mm (inverted) (ii) -5 mm (inverted)]

Question 5

A near-sighted man cannot focus on objects further than 0.5 m from his eyes, and acquires spectacles which are 20 mm from his eyes.

(i) Calculate the power of the spectacle lenses that he would need to see distant objects.

(ii) If his near point is 30 mm from his eyes without spectacle, what will it be with his spectacles.

[(i) -2.08 dioptre (ii) 30.2 mm]

Question 6

A telescope consists of two converging lenses. The objective has focal length 1 m and the eyepiece focal length 40 mm. The image of a star is formed at the minimum distance of distinct vision (250 mm).

(i) What is the separation of the lenses?

(ii) Explain whether the image is inverted or upright.

[(i) 1.03 m, (ii) Inverted]

University of the Witwatersrand, Johannesburg

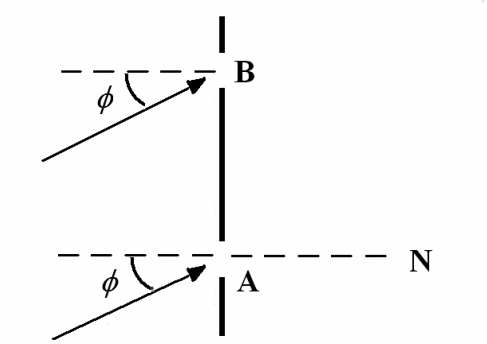
School of Physics

PHYS1001/6 (Physics I D)

Tutorial 5.3 – [2018] **Physical Optics**

Question 1

Light of wavelength 560 nm strikes a double slit AB with slit separation 1,4 μm at an angle $\phi = 30^\circ$ as shown. An observer looks at the interference pattern formed on a screen placed on the opposite side of the double slit.

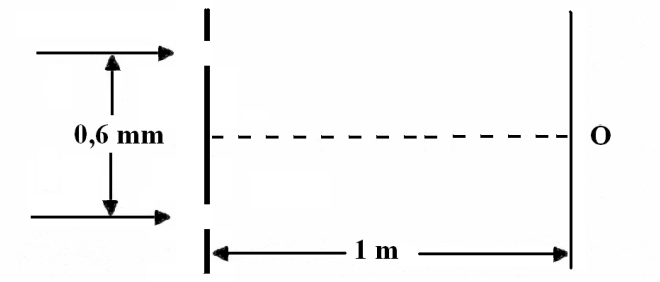


At what angles to the normal AN, within the angle NAB, does he see interference maxima?

[5.7°, 30°, 64.2°]

Question 2

The diagram represents a double-slit interferometer set up in air and illuminated by parallel, normally-incident light. The slits are 0,6 mm apart, and an interference pattern is formed on a screen 1 m from the slits.



(i) When monochromatic light is used the fringe separation on the screen is 0,9 mm. What is the wavelength of the light?

(ii) A thin parallel-sided sheet of mica, of thickness 13,6 μm and refractive index 1,60, is placed over the lower slit. What is the order of the bright fringe now at O?

(iii) The illumination is now changed to white light. Where is the white fringe formed on the screen? Ignore dispersive effects in the mica.

[(i) 540 nm, (ii) 15 (iii) 13.5 mm below O]

Question 3

White light, with wavelengths ranging from 400 nm to 700 nm, shines normally onto a grating ruled with 500 lines per mm.

(i) What is the angular width of the first-order spectrum?

(ii) What is the order of the first spectrum whose red edge overlaps the violet end of the next spectrum?

[(i) 90^{o}, (ii) 2^{nd}]

Question 4

A thin wedge of air is trapped between two glass plates. It is illuminated from vertically above with light consisting of two wavelengths, 400 nm and 600 nm.

For what thickness of the wedge in the range 100 nm to 700 nm will there be a dark fringe in the reflected light? Assume that there are no interference effects within the glass itself.

[600 nm]

Question 5

Light of wavelength 633 nm shines from a laser on to a thin smear of blood on a glass slide. The first-order dark ring formed on a screen 1 m from the slide has a diameter of 0,2 m.

Calculate the diameter of the diffracting blood corpuscles.

[7.7 μm]

Question 6

The Viking Mars Lander had to choose a suitable landing place on mars. In order to do this it had to be able to resolve objects 500 mm apart.

From what distance from the surface of Mars did it have to send back pictures from its TV camera, which had a 100 mm diameter aperture?

Assume that red light of wavelength 630 nm was used.

[65.1 km]

PHYSICS ID [PHYS1001/1006] LECTURE NOTES

6. MODERN PHYSICS

6.1. QUANTUM PHYSICS

At the beginning of the 20th century some scientists felt that they had learned almost all there was to know about physics – all that remained was to fill in the details to explain observed phenomena using existing theories of mechanics, electromagnetism, thermodynamics and statistical physics.

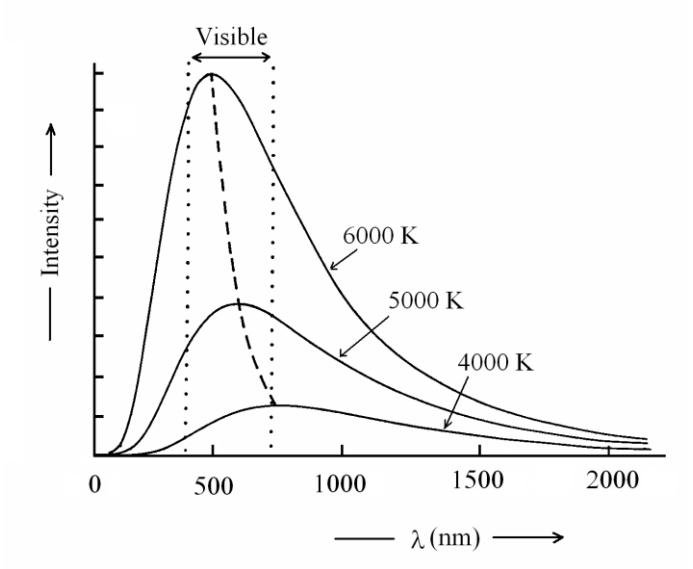
However, there were a number of phenomena discovered late in the 19th century, mostly related to the world of very small objects such as atoms and molecules, which defied description in terms of the physics known at the time.

The eventual understanding of these phenomena required the development of quantum mechanics, which was the culmination of almost 30 years of intense experimental and theoretical activity starting in 1900.

6.1.1. Blackbody Radiation

All objects (including solids and liquids, and even dense gases) at any temperature emit electromagnetic radiation, usually referred to as thermal radiation.

- At normal temperatures, we are not aware of the radiation because of its low intensity.



The diagram shows the spectrum of radiation emitted by an idealised **black body** at various temperatures.

It can be seen that the spectrum contains a continuous range of wavelengths and has a pronounced peak.

- The intensity of the radiation at a particular temperature is the area under the curve for that temperature. According to Stefan's law it depends on the fourth power of the temperature (in kelvin).
- The peak in the curves shifts with temperature. As the temperature increases the peak moves towards lower wavelengths (i.e. higher frequencies).

The wavelength λ_{max} for maximum intensity at any temperature T (in kelvin) is given by a formula first discovered empirically by Wien:

$$\boxed{\lambda_{max} T = 2.90 \times 10^{-3} \text{m.K}} \qquad \text{(Wien's displacement law)}$$

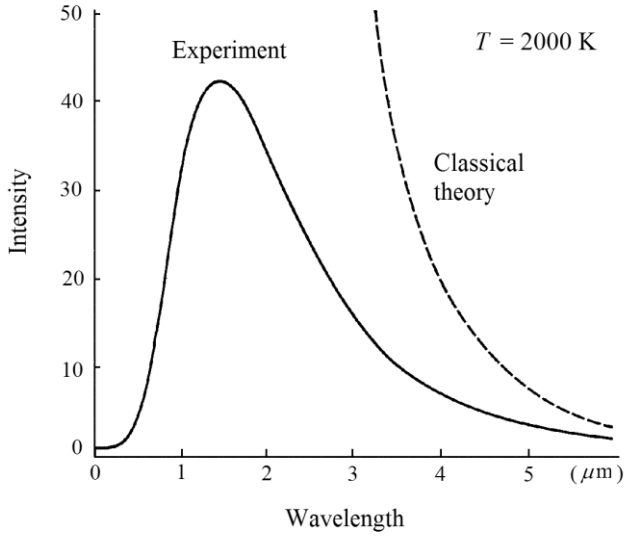
- For a body with a surface at room temperature (about 27°C or 300 K), λ_{max} = 10 μm, which is in the far infrared part of the spectrum, and the intensity of the radiation is very low.
- The peak for an object at 1000 K is still in the infrared, with λ_{max} = 3 μm = 3000 nm, and we can feel the radiation as heat. The intensity in the visible region is sufficient that the object appears to glow red.

- The surface temperature of the sun is about 5800 K, for which $\lambda_{max} = 500$ nm. The peak is therefore well within the visible region of the electromagnetic spectrum (see the curve for 6000 K).

Classical physics was able to explain the existence of thermal radiation:

- Electromagnetic theory predicts that an oscillating electric charge emits electromagnetic radiation, so that thermal radiation was interpreted as being due to oscillations of charges in the molecules of the hot body.
- The thermally agitated charges can have a distribution of frequencies, so that the spectrum of radiation emitted will be continuous in frequency and hence wavelength.
- As the body becomes hotter, the frequency of the oscillations increases and thus the wavelength of the radiation decreases.

Theoretical attempts to explain the **shape of the spectrum** were however unsuccessful.



The diagram shows the prediction of Rayleigh and Jeans (1900) compared with the observed spectrum at $T = 2000$ K.

The theory fits the data well only at large wavelengths.

The discrepancy at small wavelengths became known as the **ultraviolet catastrophe**.

In 1900, Max Planck (1858–1947) proposed an empirical formula that fitted the data. This formula contained an adjustable parameter *h*, now called Planck's constant, whose value Planck found by fitting the formula to the experimental curve; the modern value is $h = 6{,}626 \times 10^{-34}$ J.s.

In order to explain his formula, Planck made the radical proposal that the energy of any molecular vibration with frequency *f* is given by

$$\boxed{E = nhf, \quad n = 1, 2, 3, \cdots} \qquad \text{(quantum hypothesis)}$$

i.e. the molecular energy of vibration is quantised.

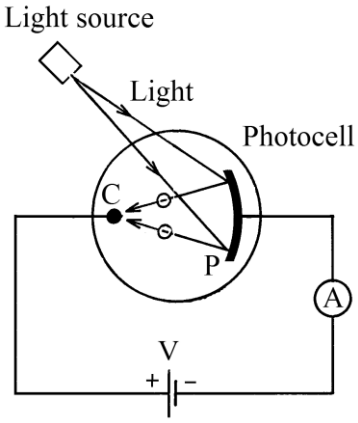
- This is Planck's **quantum hypothesis**, which was the beginning of the development of quantum theory.

Planck's explanation was not generally accepted by scientists (including Planck himself) until Einstein extended the quantum hypothesis a few years later to explain the photoelectric effect.

6.1.2. Photoelectric Effect

The photoelectric effect occurs when light of sufficiently short wavelength shines on a clean metal surface, and electrons (called photo-electrons) are emitted from the surface. The effect was first discovered in 1887 by Heinrich Hertz (1857–1894) and was studied extensively by Lenard in 1900.

The effect may be demonstrated by shining light onto a photocell connected to an ammeter and variable power supply as shown in the diagram.



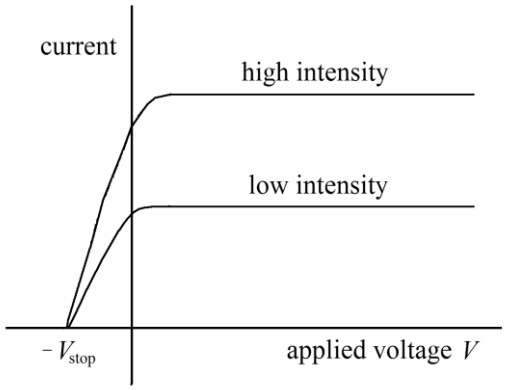
Provided light of a sufficiently high frequency is used, the ammeter indicates that a current is flowing. Thus, electrons must be moving across the photocell from the metal plate P to the collector C.

The following effects are observed as the frequency and intensity of the incident light are varied.

- If the frequency f of the incident light is below a certain minimum value f_0, called the threshold frequency, no electrons are emitted, even for a very intense beam of light.
- For frequencies above f_0, electrons are emitted almost instantaneously (less than 10^{-9} s after the surface is illuminated).
- The number of electrons emitted per second (the current) is proportional to the intensity of the light, as shown in the diagram below.

If the polarity of the supply is **reversed** and its voltage increased slowly, it is found that the current drops. The electrons released from the plate P are being repelled by the collector C, which is now at a negative potential relative to the plate.

At reverse voltage V, only those electrons whose kinetic energy on leaving the plate P exceeds eV can reach the collector; those with smaller energies will be turned back.

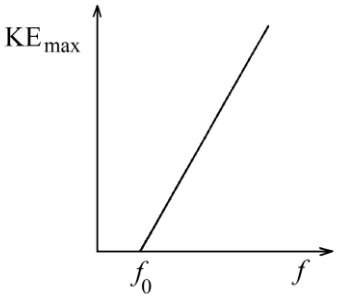


When the reverse voltage is equal to or greater than some value V_{stop} (the **stopping potential**) the current drops to zero.

The value of V_{stop} is found to be independent of the intensity of the incident light.

The maximum kinetic energy of the electrons emitted from the plate, KE_{max}, is related to the stopping voltage through the relationship

$$\boxed{KE_{max} = eV_{stop}}$$



The maximum kinetic energy is found to depend linearly on the frequency of the incident light, as shown in the graph to the left.

It does not depend on the intensity of the incident light.

We now discuss the photoelectric effect in terms of the wave theory of light, the standard theory at the beginning of the 20th century, and particle theory of light introduced by Einstein in 1905. We assume that the incident light is monochromatic.

Photoelectric effect and wave theory of light

Classically, the incident radiation contains an oscillating electric field. A negatively-charged electron in the metal plate should oscillate in response to the field, and thereby absorb energy from the field. If the amplitude of the electron's oscillations is large enough, it should break free from the surface of the target; i.e. it should be ejected from the target to form the observed current.

Wave theory predicts:

(a) The photoelectric effect should occur at any frequency of the incident light, provided the light intensity is sufficiently high; i.e. there should be no cut-off frequency.

(b) The photoelectrons should require a significant amount of time to absorb from the incident radiation enough kinetic energy to escape from the metal, particularly at low intensities.

(c) If the intensity of the incident light is increased, more energy is carried into the metal per unit time; hence electrons of higher kinetic energy should be ejected, i.e. KE_{max} should increase with intensity.

(d) The frequency of the incident light should not affect the kinetic energy of the ejected electrons – KE_{max} should depend **only** on the intensity of the light.

Thus the wave theory of light can explain the formation of the photoelectrons but is unable to explain many observed features of the photoelectric effect.

Photoelectric effect and particle theory of light

Einstein reasoned that if the energy of the molecular oscillations that produce light is quantised, as suggested by Planck, then the energy of the resulting radiation should also be quantised.

- If an oscillator initially in a state of energy *nhf* emits energy, it must make a transition to a state of energy $(n-1)hf$; the conservation of energy requires that radiation of energy *hf* will be emitted.
- Light should therefore travel through space in localised packets or quanta (now called photons) each with an energy $E = hf$, where f is the frequency of the light, which is equal to the frequency of the oscillator.

In a monochromatic beam, all photons have the same energy *hf*. In the photocell, an electron is ejected from the metal by a **single** collision with a **single** incident photon; the electron absorbs all the photon's energy *hf* and the photon disappears.

Some minimum energy W_0 (called the work function of the metal, typically a few eV) is required to get an electron just out of the surface of the metal. If $hf < W_0$ an electron cannot be ejected.

If $hf > W_0$ then, from the conservation of energy, an electron will be emitted with kinetic energy given by

$$\boxed{hf = KE_{\max} + W_0} \qquad \text{(photoelectric equation)}$$

For electrons more tightly bound in the metal, the electron has a correspondingly smaller kinetic energy when it emerges from the metal; thus the kinetic energy predicted by the photoelectric equation is the maximum value possible.

The photon theory predicts:

(a) An increase in intensity of the light beam means that more photons are incident, so more electrons will be ejected. But since the energy of each photon is not changed, the maximum kinetic energy of the electrons is not changed; i.e. $KE_{\max}$ is independent of the intensity of the light.

(b) If the frequency of the light is increased, the maximum kinetic energy of the electrons increases linearly according to $KE_{\max} = hf - W_0$.

(c) It follows from the equation $KE_{\max} = hf - W_0$ that no electrons can be emitted unless $hf > W_0$. Therefore there is a cut-off frequency f_0, where

$$\boxed{f_0 = \frac{W_0}{h}} \qquad \text{(cut-off frequency)}$$

(d) Since an electron is ejected as the result of a single collision with an incident photon, it does not have to wait to absorb sufficient energy to escape; it receives the required amount of energy all at once (provided $hf > W_0$).

These predictions explained the facts as known in 1905. Further experiments carried out by R.A. Millikan in 1913–1914 had results fully in agreement with Einstein's photon theory; in particular, Millikan's experiments were in agreement with the Einstein's photoelectric equation.

Planck's constant can be determined from the slope of a plot of $KE_{\max}$ against frequency; the value obtained agrees with that deduced from the spectrum of black-body radiation.

Photocells have many modern uses; for example:

(i) They can be used in street lights. When ambient light falls on a photocell, the current produced activates a switch that turns off the street light.

(ii) Photocells are used in lift doors etc. to detect a person passing through the door. The person breaks a light beam and the photo-current is interrupted; this signals the door to open.

6.2. ATOMIC PHYSICS

6.2.1. Structure of the Atom

It was known in the 19th century that matter was composed of atoms. However, the precise structure of the atom was still very unclear. In 1897 J.J. Thomson (1856–1940) discovered the electron. This lead to his postulation of the "plum pudding" model of the atom in which the electrons were thought to be embedded in a mass of positive charge.

In 1911, Ernest Rutherford (1871–1937) proposed the **planetary model** of the atom; this gave good physical insight into the structure of the atom.

- Electrons (negatively charged) are thought of as moving in orbits about the nucleus (positively charged), in a way similar to planets orbiting the sun.
- The radius of the electron orbits is approximately 10^{-10} m while the radius of the nucleus is approximately 10^{-15} m. Thus, an atom is mostly empty space. Over 99,9% of the mass of an atom resides in the nucleus.

This model however suffered from two serious defects.

- An atom emits electromagnetic radiation with certain characteristic frequencies. The Rutherford model had no explanation for this type of radiation.
- More seriously, according to this model all atoms should be unstable. An electron in circular motion is obviously accelerating; classically, such accelerating charges emit radiation with a frequency equal to the frequency of the circular motion. As the electron radiates energy the radius of its orbit should steadily decrease and the frequency of revolution will consequently increase. Therefore, the frequency of the radiation should increase until the electron spirals into the nucleus and the atom collapses.

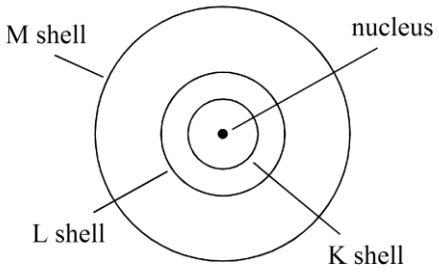
The **Bohr model** of the atom was developed shortly afterwards in order to overcome these defects by including early quantum concepts. In 1913 Niels Bohr (1885–1962) postulated that:

- Electrons move in circular orbits about the nucleus. The electrons are held in the orbit by the electrostatic force exerted on them by the nucleus, and the circular motion can be described using classical mechanics.
- Only certain electron orbits are stable; an electron in such an orbit has a definite, fixed energy and moves in the orbit without radiating energy.
- Light, or other electromagnetic radiation, is emitted only when an electron jumps from one orbit to another of lower energy, a process that cannot be described using classical mechanics.

By imposing a particular condition on the stable electron orbits, Bohr was able to derive expressions for the energy and radius of the allowed orbits in the simplest atom, that of hydrogen, which contains a single electron orbiting a nucleus consisting of a single proton:

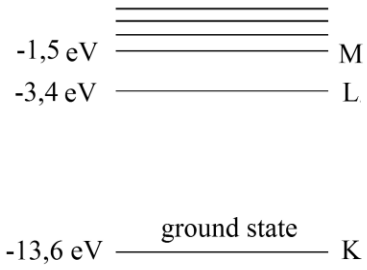
$$\boxed{\begin{aligned} E_n &= -\frac{13{,}6\ \text{eV}}{n^2} \\ r_n &= 5{,}29\times10^{-11}\ n^2\ \text{m} \end{aligned}} \qquad \text{(Bohr theory for hydrogen)}$$

The principal quantum number n can take on all positive integer values.



The diagram shows the innermost orbits of the hydrogen atom (not to scale).

The electron orbits are usually labelled K, L, M, N and so on, in order of increasing radius.



The K shell corresponds to $n = 1$, the L shell to $n = 2$ and so on, in the Bohr formulas for the energy and radius of the hydrogen atom.

The diagram on the left shows schematically the lowest energy states of the hydrogen atom, corresponding to the orbits in the previous diagram.

It should be noted that the Bohr theory is very successful for hydrogen and ions with a single electron, but is unable to explain the properties of atoms with two or more electrons with any accuracy.

6.2.2. Atomic Spectra

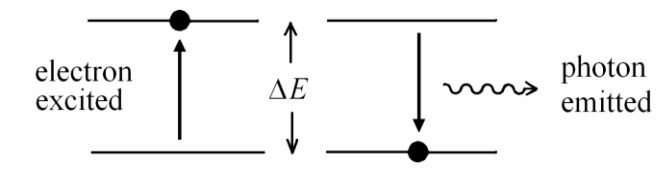
One of Bohr's motivations in developing his theory of atomic structure was to provide an explanation of the line spectra emitted by atoms, particularly hydrogen whose spectrum was accurately known by that time.

Line spectra are spectra containing only certain wavelengths of radiation, and they result from transitions that take place within an atom. Such spectra can be produced by a variety of mechanisms, such as heating a gas of the atoms, collisions between atoms and irradiation of a gas with light.

Consider an evacuated glass tube filled with the atoms of a gas at low to moderate pressures. A voltage is applied to metal electrodes in the tube.

- If the applied voltage is sufficiently large to produce a current in the gas, the tube emits light whose colour is characteristic of the gas (this is how a neon tube works).
- If the emitted light is then passed through a spectrometer it produces a spectrum of discrete bright lines, each line having a different colour (hence different wavelength).

The light is emitted in electronic transitions in the atoms of the gas. Atomic electrons in the gas are raised to higher energy levels by collisions with accelerated electrons. When an electron subsequently falls back to the energy level from which it has been raised by excitation, it emits all the excess energy in a single packet called a quantum of energy or a **photon**.



Since energy must be conserved in this process, the energy of the quantum emitted is

$$\boxed{E_{\text{photon}} = \Delta E = E_{\text{initial}} - E_{\text{final}}}$$

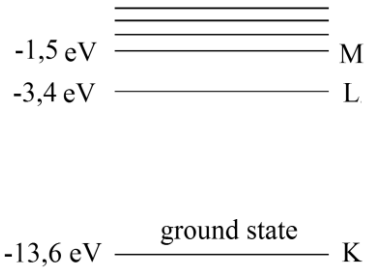
where E_{initial} is the energy of the upper electron state and E_{final} that of the lower electron state. The energy of the emitted photon is related to its frequency and wavelength by

$$\boxed{E_{\text{photon}} = hf = \frac{hc}{\lambda}} \qquad \text{(photon wavelength)}$$

where h is Planck's constant and c is the speed of light.

- Since ΔE for any pair of energy levels varies from one element to another, the frequencies emitted by an element are characteristic of it and can be used to identify it.
- The energy levels involved in **optical** transitions are generally those far out from the nucleus where the energy-level differences are comparatively small (as opposed to X-ray line spectra which will be discussed later).

Consider the example of hydrogen, referring to the energy-level diagram reproduced on the left.



The single electron of the hydrogen atom is usually in the $n = 1$ ground state.

To move the electron to the $n = 2$ state, the atom would have to absorb 13,6 – 3,4 = 10,2 eV of energy.

- The K shell would then be unpopulated, and the atom would after a very short time interval return to its original state by the emission of electromagnetic radiation in the form of a photon of energy 10,2 eV (and wavelength 122 nm, from $E_{\text{photon}} = hc/\lambda$).
- To be excited from the ground state to the n = 3 state, the atom would have to absorb 13,6 – 1,5 = 12,1 eV of energy.
- The electron could subsequently move directly back to the ground state, or it could first move to the $n = 2$ state by emitting a photon of energy 3,4 – 1,5 = 1,9 eV and wavelength 656 nm.

Experiments carried out towards the end of the 19th century showed that the spectrum emitted by hydrogen contained radiation of just a few wavelengths, including those calculated above. In particular, the radiation of 656 nm produces a bright red line when the spectrum is analysed.

The spectra described above are referred to as **emission spectra**; radiation is **emitted** during atomic transitions. An **absorption spectrum** is produced when cooler vapour **absorbs** light of those frequencies which the vapour itself would emit at higher temperatures. The absorption spectrum consists of dark lines on a continuous coloured background, the dark lines resulting from photons absorbed from the incident light.

For example:

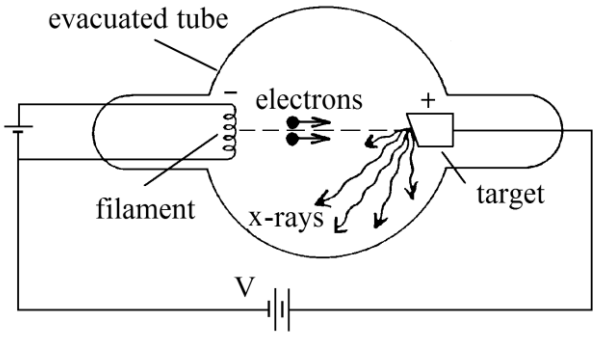
- If white light passes through sodium vapour, the sodium atoms will absorb from the incident light photons of the right frequency to cause transitions in the sodium atoms.
- The transmitted light will then consist of the continuous incident white light minus the frequencies which have been absorbed.
- Each dark line in the absorption spectrum of a particular element coincides exactly with the bright line seen in the emission spectrum of the same element.

6.2.3. X-Rays

In 1895 Wilhelm Röntgen (1845–1923) discovered that when a beam of fast-moving electrons struck the end of the tube in which they had been produced, highly penetrating radiation was emitted. He called this radiation X-rays. In 1912 Max von Laue (1879–1960) established the wave nature of the radiation by diffracting X-rays with a crystal (diffraction is a wave phenomenon).

Production of X-rays

X-rays are electromagnetic waves of very short wavelength (less than about 10^{-8} m), i.e. very high frequency and energy. They are usually produced by allowing high-energy electrons to strike a metal target. A typical X-ray setup is shown below.



A current in the filament causes electrons to be emitted, and these are accelerated towards the target which is held at a much higher potential than the filament. Some of the electrons incident on the target lose kinetic energy in collisions with atoms of the target; this energy is emitted as X-rays, through processes described below.

The accelerating voltage *V* that must be used in an X-ray tube is determined by how the X-rays will be employed. For example, when used for medical diagnosis, voltages up to about 150 kV are appropriate, whereas for medical therapeutic purposes voltages are in the approximate range 250 kV to 4 MV.

The target used in the X-ray tube must have particular properties:

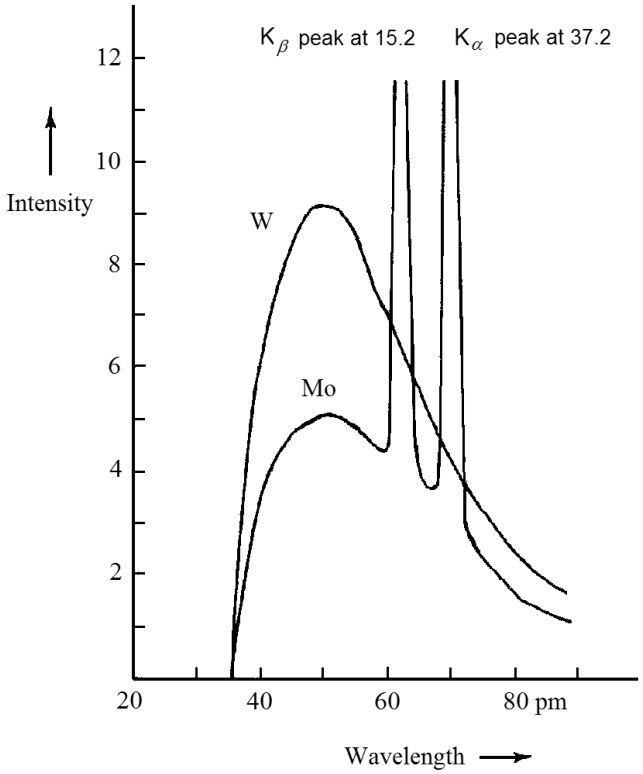
- The target must have a high melting point, since most of the kinetic energy of the electrons (about 99%) is dissipated inside the target as heat. The target is water- or oil-cooled because of the large heat dissipation.
- The target material must have a high atomic number, since the energy-level differences in light atoms are too small to generate X-rays. Tungsten and molybdenum are often used in targets.

- The target and filament are encased in an evacuated tube to prevent collisions between electrons and air molecules.

Origin of the X-Ray spectrum

There are usually two components to the spectrum produced by an X-ray tube: a continuous spectrum and a line spectrum (first observed in 1908), which is superimposed on the continuous spectrum and is not always present.

- The continuous spectrum depends on the voltage applied to the tube.
- The line spectrum is characteristic of the target material.



Both features are illustrated in the diagram on the left, which shows spectra of molybdenum (Mo) and tungsten (W) targets for an accelerating voltage of 35 kV.

Note that at 35 kV no line spectrum is produced for tungsten.

The continuous spectrum

The continuous spectrum is produced by electrons suddenly slowing down in collisions with atoms in the metal target. In such collisions, some or all the electron's kinetic energy will be converted to X-rays.

According to classical physics, whenever a charge is accelerated it must emit radiation. Since the radiation is caused by the electrons being slowed down, it is often called **Bremsstrahlung**, or braking radiation. Classical physics is therefore able to explain the formation of such X-rays.

According to quantum theory, the X-rays emitted by this process emerge in the form of photons, each collision producing a single photon with energy equal to the kinetic energy lost by the electron.

Photons with the largest energy are emitted when an electron loses all its kinetic energy in a single collision. Since the electron of charge $-e$ has been accelerated through a potential difference V, its

kinetic energy before the collision is eV. Therefore, from the conservation of energy in the collision, the maximum photon energy is given by

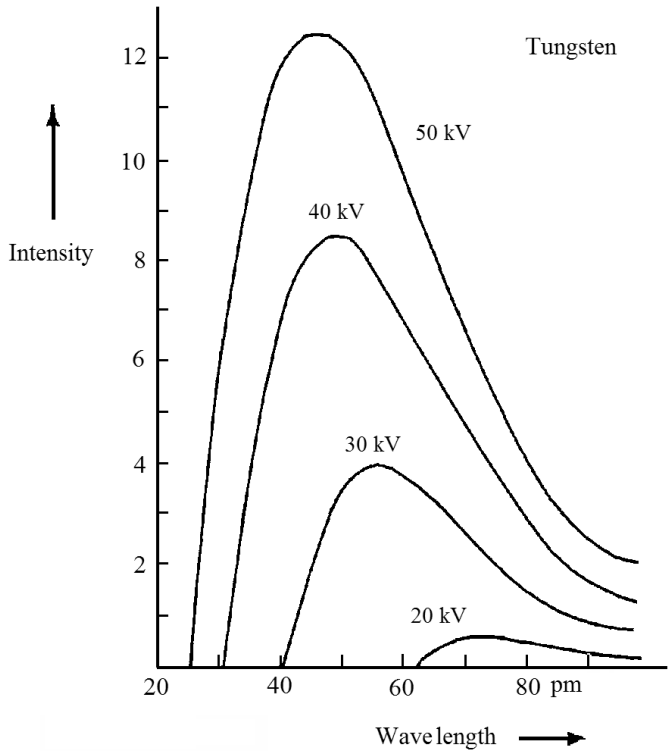
$$eV = E_{max} = \frac{hc}{\lambda_{min}}$$

where λ_{min} is the photon wavelength corresponding to the maximum photon energy:

$$\boxed{\lambda_{min} = \frac{hc}{eV}}$$

It follows that the continuous spectrum should have a short wavelength cut-off, λ_{min}, which is the same for all target elements at a given accelerating voltage; this feature is illustrated in the previous diagram for W and Mo targets.

Electrons that lose smaller amounts of energy in each of several collisions emit lower energy X-rays, and therefore larger wavelengths than the minimum.



The equation for the cut-off wavelength predicts that λ_{min} should decrease as the accelerating voltage is increased.

Plotting intensity versus wavelength for several accelerating voltages, for the same target (tungsten in the diagram on the left), confirms this prediction.

Note that although the classical theory of radiation is able to explain the existence of Bremsstrahlung, it cannot account for the minimum wavelength; this is a purely quantum effect.

The line spectrum

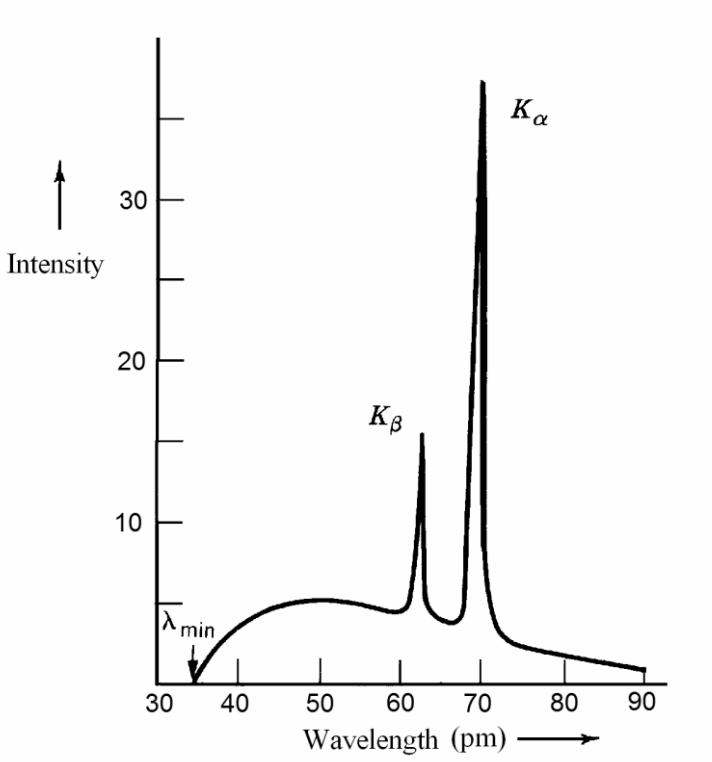
A line spectrum comprises pronounced intensity peaks at certain wavelengths, indicating the enhanced production of X-rays; the wavelengths are characteristic of the target material used. The spectrum originates in the rearrangement of the electron structure of the target atoms after having been disturbed by the bombarding electrons.

- A fast moving electron collides with an electron in an inner shell of a target atom with sufficient energy to remove the electron from the atom.
- The vacancy created in this shell is filled when an electron in a higher level drops down into the lower-energy level containing the vacancy. This normally happens within about 10^{-9} s.

- This transition is accompanied by the emission of a photon whose energy equals the energy difference between the two atomic shells. This is the same process that was described earlier for hydrogen, except that the energy differences are much larger (usually greater than 1 keV) so that wavelengths are much smaller (0,01 – 1,0 nm); X-rays are produced rather than radiation in the visible range.
- This first transition leaves a further vacancy in the higher-energy shell; this is filled by an electron dropping from a yet-higher shell, and a second photon is emitted, and so on.

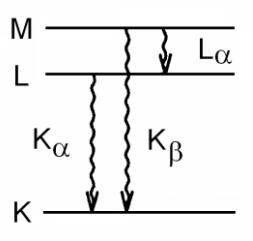
Example: spectrum of molybdenum at 35 keV.

Many of the features described above can be illustrated by consideration of the spectrum for molybdenum at an accelerating voltage of 35 keV, shown in the next diagram.



- The equation $\lambda_{min} = hc/eV$ predicts that for V = 35 kV the continuous spectrum should have a minimum at a wavelength of 35,5 pm, as is observed.
- Sharp lines are observed with wavelengths of about 63 pm and 71 pm. From $E = hc/\lambda$, these correspond to the emission of photons of energy 19,6 keV and 17,4 keV, respectively.
- Using these and other measured transition energies, an energy level diagram can be constructed for molybdenum. The lowest few levels are shown schematically in the diagram below.

The lines are labelled according to the energy levels involved in the transition. The K lines are formed when an electron drops to the K shell, and so on. The lines are further labelled α, β, γ etc. to indicate that the electron falls from the next highest level, the one above that, and so on.



The K_α line is produced by transitions from the L shell to K shell, the K_β line by transitions from the M shell to the K shell, and so on.

K_α = 17,4 keV
K_β = 19,6 keV
L_α = 2,3 keV

The wavelength of the radiation emitted in the L_α transition (about 540 pm) is far beyond the range shown in the diagram for a Mo target.

- If the kinetic energy of incident electrons is too small, no atomic electrons can be ejected from the atom and so no line spectrum will be produced.

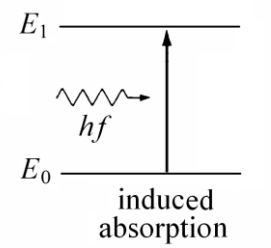
For example, there are no K lines for **tungsten** at an accelerating voltage of 35 kV. However, since it takes less energy to knock an electron out of the L shell, transitions from higher atomic states into the L shell may still take place; the resulting L lines will occur at much larger wavelengths since the energy differences are smaller.

6.2.4. The Laser

The term **laser** stands for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation. The laser was invented in 1960.

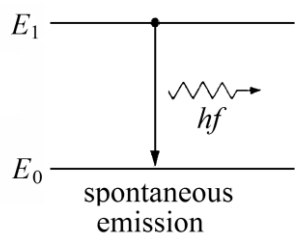
- A laser produces a beam of light whose waves all have the same frequency (they are monochromatic to within about 10^{-6} nm).
- The waves are exactly in phase with one another (they are coherent).
- The beam is also well collimated and so spreads out very little even over long distances.

Transitions between two energy levels in an atom can occur by induced absorption, spontaneous emission, and induced emission, the first two of which were discussed earlier.



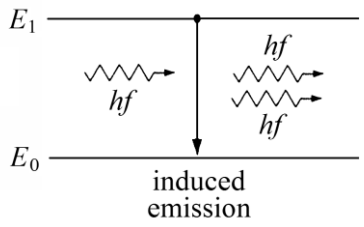
At ordinary temperatures, most atoms in a sample will be in the state of lowest energy E_0, the ground state. If a container of gas is illuminated by light containing photons of all energies (i.e. a continuous spectrum), all those photons whose energy corresponds exactly to the difference in energy between two levels in the atom will be absorbed.

So, for example, an atom can be raised from the ground state to an excited state with energy E_1 by absorption of a photon of the correct energy $E_1 - E_0$. This process is stimulated or **induced absorption**.



Once an atom is in an excited state, there is a constant, fixed probability that it will revert back to a lower level by emitting a photon.

This process is called **spontaneous emission**, and it usually occurs within about $10^{-9} - 10^{-8}$ s, which is the average lifetime of a typical excited state of an atom.



A third process that is important for the operation of a laser is stimulated or **induced emission**, which was predicted by Einstein in 1917.

Suppose an atom is in an excited state, as shown in the diagram, when a photon of energy $E_1 - E_0$ is incident on it.

The incoming photon increases the probability that the atom will return to the ground state, thereby emitting a second photon of exactly the same energy as the incident photon.

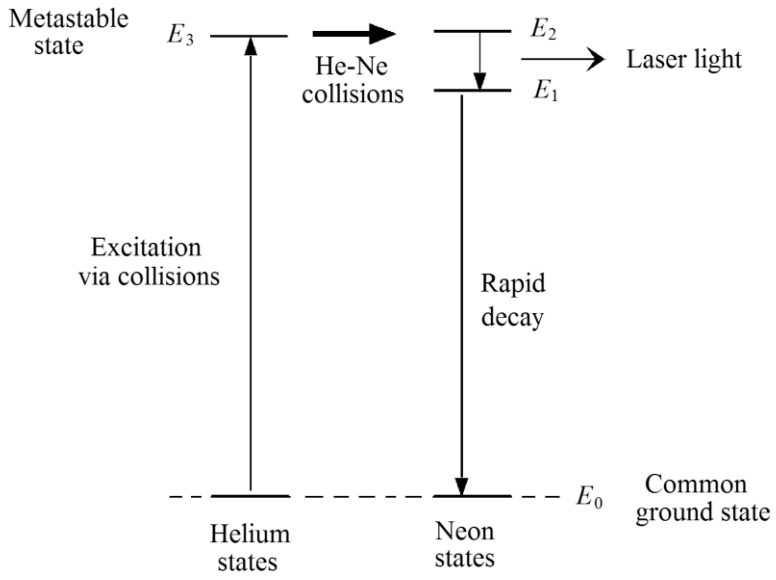
Note that the process of stimulated emission produces two identical photons, the incident photon and the emitted photon, and these are **exactly in phase**. These photons can stimulate or induce other atoms to emit identical photons in a chain of similar processes. The many photons produced in this way are the source of the intense, coherent light produced by a laser.

Several conditions are necessary for the operation of a laser:

- As indicated, normally most atoms in a sample will be in the ground state; when light is incident on a gas of these atoms the net result is the absorption of energy. For the laser to work we require more atoms to be in an excited state rather than the ground state; this is called **population inversion**. How this is achieved depends on the type of laser.
- A further condition is the existence of a **metastable state** in the atom, an excited energy level whose lifetime may be 10^{-3} s or more instead of the usual 10^{-8} s. This is necessary in order that stimulated emission takes place before the excited state decays by spontaneous emission.
- In addition, the emitted photons must be confined within the laser long enough to allow them to induce further emissions from other atoms in metastable states. This is achieved by placing reflecting mirrors at the two ends of the laser tube; one mirror is totally reflecting and the other is slightly transparent to allow the laser beam to leave the laser.

A common type of laser is the **helium-neon laser**, invented in 1961, which contains a mixture of helium and neon gases (in the ratio 20:80). The figure shows simplified energy-level diagrams for the two atoms.

- A high voltage applied to the laser tube causes electrons to sweep through the tube, colliding with atoms of the gas and raising them to excited states, including the metastable state at energy E_3 = 20,61 eV in helium.
- Neon contains a state at energy E_2 = 20,66 eV, which is very close to the energy of the metastable state in helium. Therefore, when a metastable helium atom collides with a neon atom in its ground state, the excitation energy of the helium atom is often transferred to the neon atom which is left in the state of energy E_2.



- The metastability of the helium level E_3 ensures a ready supply of neon atoms in level E_2. In this way the state of neon with energy E_2 becomes more heavily populated than the state at lower energy E_1, i.e. there is a population inversion.
- The coherent beam of wavelength 632.8 nm that the laser emits results from transitions from the state of neon with energy E_2 to the state at energy E_1 (which then decays rapidly to the ground state via intermediate levels not shown in the diagram).

6.3. NUCLEAR PHYSICS

6.3.1. The Nucleus

The atom consists of a small positively-charged nucleus, surrounded at a relatively enormous distance by orbital electrons. The nucleus itself consists of protons (positive) and neutrons (electrically neutral), together called **nucleons**. The existence of the neutron was confirmed experimentally in 1932 by James Chadwick (1891-1974).

Nuclear properties and the nucleon

Some properties of atomic particles are compared in the following table. Note that the masses of the neutron and proton are very similar, each having a mass about 1840 times that of the electron.

Particle	Mass		Charge	
	(kg)	(u)	(C)	$\lvert e \rvert$
Proton	$1{,}673 \times 10^{-27}$	1,00728	$+ 1{,}602 \times 10^{-19}$	+ 1
Neutron	$1{,}675 \times 10^{-27}$	1,00866	0	0
Electron	$9{,}109 \times 10^{-31}$	$5{,}49 \times 10^{-4}$	$- 1{,}602 \times 10^{-19}$	– 1

The **atomic mass** of an *atom* is its mass expressed in (unified) atomic mass units; this is defined in such a way that the mass of the neutral ^{12}C atom is **exactly** 12 u. In terms of the standard SI unit of mass:

$$1 \text{ u} \cong 1{,}661 \times 10^{-27} \text{ kg}$$

$$1 \text{ kg} \cong 6{,}022 \times 10^{26} \text{ u}$$

The mass of a **nucleus** is smaller than that of the corresponding atom by an amount that is approximately (but not exactly) equal to the mass of the electrons in the atom.

A nucleus is described by three numbers:

(i) **Atomic number** *Z*: This is the number of protons in the nucleus, and equals the number of orbital electrons in the neutral atom. Every element has a different *Z*, and this gives the position of the element in the Periodic Table, and determines the chemical properties of the element.

(ii) **Neutron number** *N*: This is the number of neutrons in the nucleus.

(iii) **Mass number** *A*: This is the total number of nucleons (neutrons plus protons) in the nucleus: $A = Z + N$. The mass number is usually the integer nearest to the atomic mass in u.

The symbol for a nucleus is written ${}^{A}_{Z}\text{X}$ or ${}^{A}_{Z}\text{X}_{N}$, where X represents the chemical symbol for the element; e.g. ${}^{1}_{1}\text{H}$, ${}^{4}_{2}\text{He}$, ${}^{238}_{92}\text{U}$.

Isotopes are forms of the same element which differ in the number of neutrons in the nucleus. They have the same atomic number (which characterises the element) but different mass number.

For example, there are three isotopes of hydrogen: ${}^{1}_{1}\text{H}$ (ordinary hydrogen), ${}^{2}_{1}\text{H}$ (deuterium) and ${}^{3}_{1}\text{H}$ (tritium).

(i) Hydrogen is the simplest and lightest atom; a single electron is in orbit around a nucleus consisting of a single proton.

(ii) The next heaviest atom is deuterium; its nucleus (the deuteron) consists on a neutron and a proton, and one electron is in orbit around the nucleus.

(iii) Tritium has an unstable nucleus (the triton) containing one proton and two neutrons, and a single electron is in orbit about the nucleus.

Note that in each case the atom contains one proton and one electron; they are all forms of hydrogen and have similar chemical properties; the nuclear properties are however quite different because of the different numbers of neutrons.

Binding energy

If we add up the mass of the constituent nucleons of any nucleus, the total is greater than the mass of the nucleus by an amount called the mass deficit; if we wish to separate the nucleus into its constituent nucleons we would have to create this extra mass.

Consider the **mass of ^{12}C** as an example: if the masses of 6 hydrogen atoms and 6 neutrons are added together the total is 12,0989 u and not 12,0000 u which, by definition, is the mass of an atom of ^{12}C. (Note that we use the mass of the hydrogen atom here, rather than the mass of the proton, to allow for the fact that the neutral atom contains 6 electrons in addition to the protons and neutrons).

Mass of 6 hydrogen atoms	6 × 1,007825 u = 6,04695 u
Mass of 6 neutrons	6 × 1,008665 u = 6,05199 u
Total mass of nucleons	12,09894 u
Mass of carbon-12 atom	12,00000 u

Thus in order to take the atom apart, we must create extra mass 0,09894 u.

Einstein had already in 1905 proposed the mass-energy equivalence: he showed that matter of mass m can, under suitable conditions, be converted into energy E (and vice versa) given by

$$\boxed{E = mc^2}$$

Using this equivalence, it is easily shown that

$$\boxed{1 \text{ u} = 931{,}5 \text{ MeV}} \qquad \text{(mass-energy equivalence)}$$

The energy equivalent of the mass deficit of a nucleus is called its **binding energy** B. It is equal to the amount of work required to split the nucleus into its component protons and neutrons.

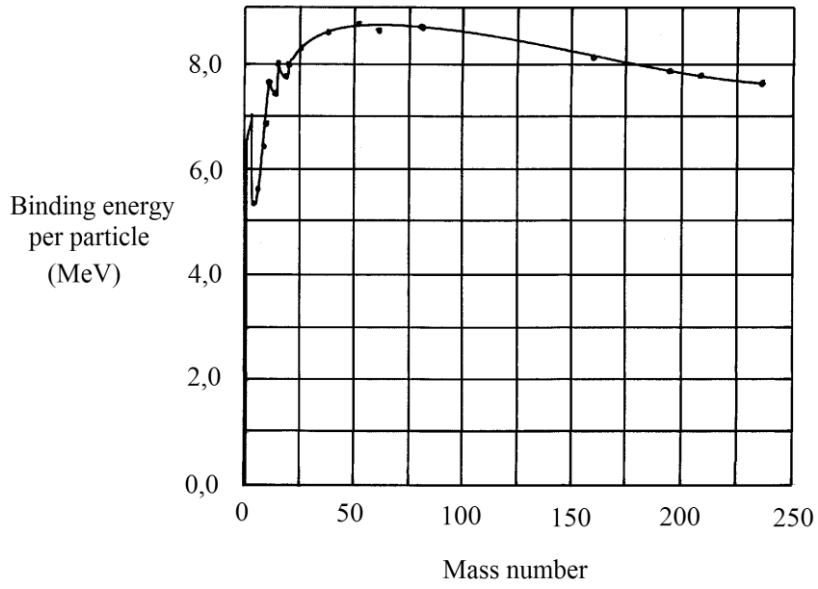
In the case of carbon-12, the energy equivalent of the mass difference 0,09894 u is the binding energy of carbon-12 which, using the mass-energy conversion factor above, is 92,2 MeV – the energy required to split the nucleus ^{12}C into its component six protons and six neutrons.

Note:

- The size of the binding energy is a measure of the strength of the force that holds the nucleons together in the nucleus – the strong nuclear interaction.
- If this force were weaker, nuclei would fly apart because of the repulsive Coulomb force between the protons.

We sometimes use the binding energy per nucleon, B/A, rather than the binding energy itself. For example for carbon-12, B/A = 7,68 MeV.

Plotting B/A versus A for the elements gives the following curve:



For most nuclei (but not the very lightest), $B/A \approx 8$ MeV. The curve peaks at about 8.7 MeV near $A = 60$ and then decreases slowly as mass number increases.

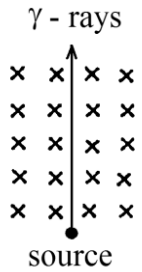
This curve can be used to explain how energy is obtained from fission and fusion reactions (see later).

6.3.2. Natural Radioactivity

Pioneers in this field included: Antoine Becquerel (1852–1908) who discovered radioactivity in 1896, Ernest Rutherford (1871–1937), Pierre Curie (1859–1900) and his wife Marie Curie (1867–1934).

Certain isotopes have nuclei with an excess of energy and they get rid of this excess energy by emitting radiation of various types. These isotopes are said to be radioactive and are called **radioisotopes**.

The most common of the emitted radiations are of three types, which were named α, β and γ rays by Rutherford. If the radiation carries an electric charge, its sign can be determined by applying a magnetic field at right angles to the direction of emission of the radiation.



In these diagrams the direction of the magnetic field is into the page, as indicated by the symbol ×.

γ rays emerging from a radioactive source are not deviated by the magnetic field, indicating that they are uncharged.

α - rays

α rays are deviated to the left, indicating that they carry a positive charge.

There is very little dispersion of the beam of particles, indicating that there is little variation in their speeds.

β- rays

The β particles are deviated to the right, showing that they are negatively charged, and are denoted β^-.

There is a second type of β radiation, less common, that would be deviated to the left, since they carry positive charge. They are denoted β^+.

The considerable dispersion in the paths shows that the β particles have widely varying speeds.

α particles

α particles are helium nuclei, i.e. helium atoms stripped of their orbital electrons. Thus an α particle consists of two protons and two neutrons bound together. Some properties are shown in the following table.

Charge:	Equal to two proton charges (+2*e*).
Mass:	About 4 times that of the hydrogen atom.
Ionization:	Intensely ionizing, i.e. they produce a large number of ion pairs per unit distance as they travel through matter.
Range:	Because of above strong interaction, their range in matter is small, e.g. about 10–100 mm in air and about 1/100 mm in solids.
Velocity:	Up to 1/20 that of light.

A general α decay is written

$$\boxed{{}^{A}_{Z}\mathrm{P} \rightarrow {}^{A-4}_{Z-2}\mathrm{D} + {}^{4}_{2}\mathrm{He}}$$

where P denotes the parent nucleus and D the daughter nucleus. For example:

$${}^{238}_{92}\mathrm{U} \rightarrow {}^{234}_{90}\mathrm{Th} + {}^{4}_{2}\mathrm{He} \quad (\tau_{1/2} = 4{,}5\times10^{9}\ \mathrm{yr}).$$

Note that total *Z* and total *A* remain unchanged during the reaction; i.e. charge is conserved and the number of nucleons is conserved.

- These two important conservation laws hold for all nuclear transformations, as do the conservation of momentum and mass-energy.

Measurements show that during the decay the total mass does not remain constant – in fact it **decreases** by an amount

$$\Delta M = M_{\mathrm{P}} - M_{\mathrm{D}} - M_{\mathrm{He}}$$

where *M* represents an *atomic* mass. The lost mass is actually converted into energy, as determined by $E = mc^2$, so the amount of energy released in the α decay of the nucleus P is

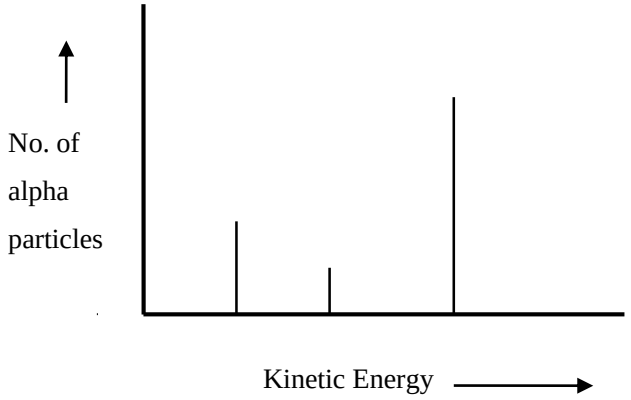
$$\boxed{(\Delta M)c^2 = (M_{\mathrm{P}} - M_{\mathrm{D}} - M_{\mathrm{He}})c^2}$$

This energy becomes:

- Kinetic energy of the α particle.
- Recoil energy of the daughter nucleus. The conservation of momentum requires that the daughter nucleus recoils with the same momentum as the alpha particles. Since it is much heavier than the α particle, its speed and therefore kinetic energy will be much less than that of the α particle, and can usually be ignored.
- Excitation energy of the daughter nucleus. Nuclei, like atoms, can have excited states. As a result of the decay, the daughter nucleus may be left in one of its excited states, rather than the ground state; this absorbs an amount of energy equal to the difference in energy between the excited state and the ground state.

As a result, the α particles emitted will have their maximum kinetic energy for decays in which the daughter nucleus is left in its ground state; this energy is characteristic of the **parent** nucleus. However, some α particles will be produced with kinetic energies reduced by amounts which are characteristic of the **daughter** nucleus.

Thus, the particles emitted in α decay form a **line spectrum** as illustrated schematically below.



- The spectral line of highest energy corresponds to the daughter nucleus being left in the ground state. If we ignore the recoil energy of the daughter nucleus, the kinetic energy of the α particle is $\left(M_{\mathrm{P}} - M_{\mathrm{D}} - M_{\mathrm{He}}\right)c^2$.
- The existence of the other lines indicates that the daughter nucleus has been left in an excited state. The excitation energies of the daughter nucleus can be deduced from the spacing of the lines.

β particles

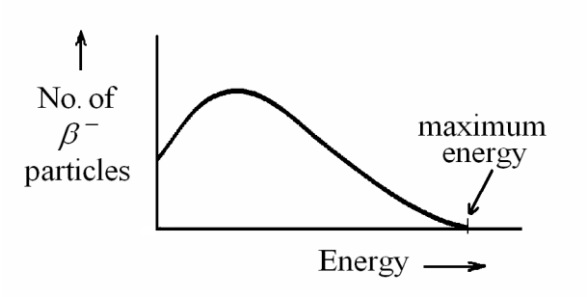
There are actually two types of β decay; by far the most common is β^- decay in which an electron is emitted. Less common is β^+ decay, in which the particle emitted is the antiparticle of the electron, the positron e^+; these are effectively positively-charged electrons.

Some properties of β particles are listed in the following table.

Charge:	Can be either negative ($-e$) or positive ($+e$)
Mass:	Very light particles.

Ionization:	Less ionizing than α particles.
Range:	About 1 m in air.
Velocity:	Up to 99,5% that of light.

The **spectrum** of β particles is quite unlike that of α particles; it is a **continuous spectrum** rather than a line spectrum. This is because the β particle is emitted together with another particle, the neutrino ν, which carries off some of the energy released in the reaction. The β particles are therefore emitted with a large range in energies, from zero up to a maximum which is characteristic of the emitter, with the neutrinos taking the rest of the energy.



The existence of the neutrino was predicted by Wolfgang Pauli in 1930, based on the apparent violation of the conservation of energy and momentum in β decay. It is uncharged, has zero rest mass (or a very small mass) and interacts extremely little with matter; neutrinos were first detected experimentally in 1956.

There are two kinds of neutrino associated with β decay, the neutrino itself, and the antineutrino $\bar{\nu}$. In β^- decay a neutron is turned into a proton:

$$n \rightarrow p + e^- + \bar{\nu} \qquad {}^{A}_{Z}\mathrm{P} \rightarrow {}^{A}_{Z+1}\mathrm{D} + e^- + \bar{\nu}$$

while in β^+ decay a proton is turned into a neutron:

$$p \rightarrow n + e^+ + \nu \qquad {}^{A}_{Z}\mathrm{P} \rightarrow {}^{A}_{Z-1}\mathrm{D} + e^+ + \nu$$

β^- decay leads to an increase of atomic number Z by one, with the mass number A unchanged. In β^+ decay, Z decreases by one. Overall, the charge and the number of nucleons are conserved.

Examples of β^- decay:

$${}^{234}_{90}\mathrm{Th} \rightarrow {}^{234}_{91}\mathrm{Pa} + e^- + \bar{\nu} \qquad (\tau_{1/2} = 24{,}5 \text{ days})$$
$${}^{234}_{91}\mathrm{Pa} \rightarrow {}^{234}_{92}\mathrm{U} + e^- + \bar{\nu} \qquad (\tau_{1/2} = 68 \text{ s})$$

Note that:

- The electrons emitted in β^- decay are **not** orbital electrons, but originate through the process $n \rightarrow p + e^- + \bar{\nu}$ inside the nucleus.
- A free neutron (not bound within a nucleus) undergoes β decay: $n \rightarrow p + e^- + \bar{\nu}$ ($\tau_{1/2}$ = 14 minutes)

γ rays

These are electromagnetic waves of very short wavelength (less than about 0,1 nm).

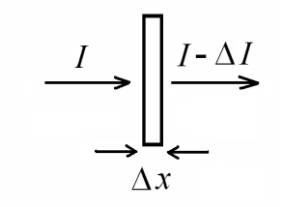
As discussed in connection with α decay, when some nuclei decay by α or β emission the daughter nucleus may be left in an excited state. The daughter nucleus subsequently returns to its ground state by emission of energy in the form of γ radiation.

This process is similar to the emission of X rays during atomic transitions, except that the energies involved are usually much larger and the photon wavelengths consequently much smaller.

Since they are uncharged and cannot cause ionization, γ rays can penetrate several mm of lead. Their intensity is gradually attenuated in their passage through matter.

Absorption of γ rays and X rays

The intensity of a beam of γ rays or X rays passing through an absorbing material decreases exponentially with increasing thickness of material.

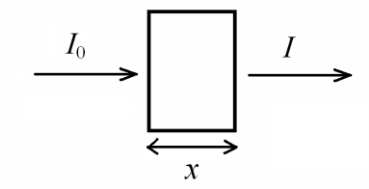


For a **thin** layer of absorber, the fractional decrease in intensity is proportional to absorber thickness.

$$\frac{\Delta I}{I} \propto \Delta x \text{ or } \frac{dI}{dx} = -\mu I$$

where μ is called the **linear absorption coefficient**.

The minus sign is inserted since the intensity decreases with increasing thickness.



For a slab of absorbing material of thickness x, we must integrate this equation, giving

$$\boxed{I = I_0 \exp(-\mu x)}$$

I_0: incident intensity, I: transmitted intensity

The coefficient μ is a constant for a particular material for γ or X rays of a given energy. It has SI unit m^{-1}.

The absorption can be described by other parameters, such as the **half value layer** $x_{1/2}$. This is the thickness of absorber necessary to reduce the incident intensity to half its original value.

When $x = x_{1/2}$, then $I = I_0/2$. From $I = I_0 \exp[-\mu x]$ we see that $\frac{1}{2} = \exp[-\mu x_{1/2}]$, giving

$\ln 2 = \mu x_{1/2}$ or

$$\boxed{x_{1/2} = \frac{\ln 2}{\mu}}$$

- A good absorber has a small value of $x_{1/2}$ and a large value of μ.
- Materials of large Z are generally good absorbers of γ and X radiation (e.g. lead shielding).

Radioactive decay laws

Radioactive decay is a random process and hence can only be described statistically. For example, we can predict what fraction of the atoms in a sample of radioactive material is going to decay in some time interval, but we have no way of knowing which atoms will actually decay.

The rate of disintegration of radioactive atoms in a sample is found by experiment to be proportional to the number of radioactive atoms present. This is on average; the actual value will fluctuate about this average. Thus for a small time interval Δt:

$$\frac{\Delta N}{\Delta t} \propto N \quad \rightarrow \quad \frac{dN}{dt} = -\lambda N$$

where the negative sign is inserted because the number is decreasing with time. The quantity λ is called the decay constant (it is constant for a given radioactive material); it determines how rapidly the radio-isotope decays.

Using integral calculus, we get an expression for the number N of radioactive atoms remaining in a sample after time t:

$$\boxed{N = N_0 \exp(-\lambda t)}$$

N_0 is the number of atoms at $t = 0$.

The **activity** R of a sample of radioactive material is defined as the rate at which its atoms are decaying.

$$\boxed{R = -\frac{dN}{dt}} \qquad \text{(activity defined)}$$

Combining this with $\frac{dN}{dt} = -\lambda N$, we see that activity is proportional to the number of radioactive atoms remaining in the sample, i.e.

$$\boxed{R = \lambda N}$$

giving

$$\boxed{R(t) = R_0 \exp(-\lambda t)} \qquad \text{(radioactive decay law)}$$

where R_0 is the activity at time $t = 0$.

Activity is easily measured with, for example, a Geiger counter.

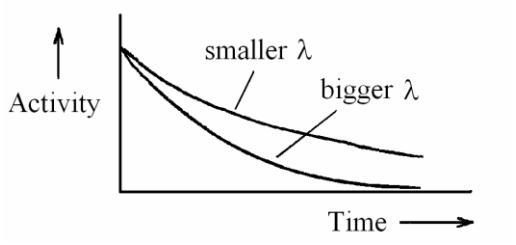
- The S.I. unit of activity is the becquerel (Bq): 1 Bq = 1 disintegration per second.
- An older unit frequently used is the curie (Ci): by definition 1 Ci ≡ $3{,}7 \times 10^{10}$ disintegrations/s. This unit came originally from work on radium: 1g of radium has an activity very close to 1 Ci.

To describe the activity of the sample, the **half-life** $\tau_{1/2}$ is normally used. This is the time required for the number of atoms originally present (or the original activity) to decay to half that number.

Therefore, at time $t = \tau_{1/2}$, we have $N = N_0/2$. Substituting this into $N = N_0 \exp(-\lambda t)$ gives $\frac{1}{2} = \exp(-\lambda \tau_{1/2})$ or

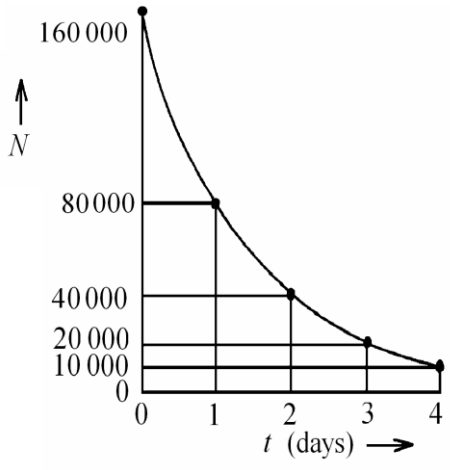
$$\boxed{\tau_{1/2} = \frac{\ln 2}{\lambda}}$$

Thus, a large value for the decay constant means a short half-life, and vice versa.



The diagram on the left shows how the rate of decay depends on the magnitude of the decay constant.

The diagram below shows a plot of the number of atoms remaining as a function of time for a particular radioactive sample.



The sample initially contains 160000 atoms with a half-life of 1 day.

After each half-life the number remaining has halved.

After n half-lives, the number remaining has been reduced by a factor 2^n.

The half-lives of naturally occurring radioisotopes for α and β emission vary over a wide range: for example $6{,}5 \times 10^{18}$ yr (for the α decay of ^{186}W) and $3{,}0 \times 10^{-7}$ s (for the α decay of ^{212}Po). Half-lives for γ emission can be much smaller.

6.3.3. Nuclear Reactions

Rutherford produced the first artificial nuclear reaction in 1919. He aimed α particles (from a natural radioactive source) at a target of nitrogen gas, producing an isotope of oxygen and protons.

$$^{4}_{2}\mathrm{He} + {}^{14}_{7}\mathrm{N} \rightarrow {}^{17}_{8}\mathrm{O} + {}^{1}_{1}p$$

Cockroft and Walton (1932) were the first to use an accelerator. They accelerated protons through a very high voltage and fired them at a lithium target. Two α particles were produced in each reaction.

$$^{1}_{1}\mathrm{H} + {}^{7}_{3}\mathrm{Li} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{4}_{2}\mathrm{He}$$

A general nuclear reaction can be represented as

$$a + \mathrm{A} \rightarrow \mathrm{B} + b$$

where A is the target nucleus, a is the accelerated particle, b is the emitted particle and B is called the residual nucleus.

The kinetics of these artificial reactions can be studied (e.g. in a cloud chamber) and the energy of the emitted particles can be found. From such studies an important result emerged: if the masses of the reactants (LHS) are added and the masses of the products (RHS) are added, these totals are different; i.e.

$$M_a + M_A \neq M_B + M_b$$

As discussed in connection with α decay, during a nuclear reaction mass is converted into energy, or vice versa, according to $E = mc^2$.

The conservation of mass-energy must be applied to all transmutation reactions, i.e.

If mass on LHS < mass on RHS (endothermic reaction)	Energy must be supplied to the system if the reaction is to take place (to create the additional mass). This is done by giving sufficient kinetic energy to the incident particles.
If mass on LHS > mass on RHS (exothermic reaction)	Excess energy released in the reaction appears as kinetic energy of the particles produced.

Fission

In a fission reaction, a heavy isotope, e.g. ^{235}U or ^{239}Pu, splits into two fragments, each of mass number around 100-120. As can be seen from the plot of binding energy against *A* shown earlier, in such a process the value of *B*/*A* increases. This means that energy is released in a fission event, around 200 MeV per reaction.

During fission two or three neutrons are emitted by each nucleus undergoing fission. Each neutron can cause a further fission, thereby producing a chain reaction. This is the basis of nuclear reactors and atomic bombs.

The fission process was first observed in 1939. The first nuclear reactor was built in 1942 by an international team led by Enrico Fermi (1901-1954) and the first atomic bombs were exploded in 1945.

Many of the fission fragments produced by a bomb are serious health hazards. For example ^{90}Sr, which is a β emitter with a half life of 29 years, is chemically similar to calcium and can ultimately find its way into bone, where it irradiates the bone marrow.

The large numbers of neutrons produced in a reactor can be used to produce nuclear transformations in suitable elements, producing radioisotopes for therapy, tracing etc.

Fusion

If two light nuclei (e.g. H, D or He) fuse, much more energy is released per unit mass than in fission (since the left-hand part of the *B*/*A* curve is steeper than the right-hand part).

For a fusion reaction to take place, the reacting particles must be given high energies in order to overcome the electrostatic force of repulsion between the fusing nuclei. This translates into very high temperatures, about 10^7 K for hydrogen.

Fusion processes of this kind occur naturally in stars (fusion is in fact the main source of energy in the Universe) and in the hydrogen bomb, which was developed in 1952. So far, fusion reactors have not been developed, mainly because of the difficulties in maintaining the required high temperatures for sufficiently long that a chain reaction can occur.

MODERN PHYSICS

LECTURE EXAMPLES

University of the Witwatersrand, Johannesburg

School of Physics

Physics I D (PHYS1001/6)

Lecture Example 6 – [2018] **Modern Physics**

1. The longest wavelength that will cause photoelectrons to be emitted from a sodium surface is 583 nm.
   If the surface is illuminated with light of wavelength 450 nm what is the maximum speed of the photoelectrons emitted? [$4{,}71\times10^5$ m.s^{-1}]

2. The energy levels in hydrogen can be expressed in the form $E = -\frac{13{,}60}{n^2}$ eV.
   (i) To what level *n* will a hydrogen atom be excited after absorbing a photon of energy 12,0 eV, assuming that it is initially in its ground (lowest-energy) state? [3]
   (ii) Calculate the energies and wavelengths of the photons which could be emitted as the atom returns to its ground state.
   [1,89 eV, 10,20 eV, 12,09 eV; 658 nm, 122 nm, 103 nm]

3. The wavelength of the K_α line of molybdenum is 0,071 nm.
   (i) What is the energy (in keV) of the photon emitted when this line is produced?
   [17,5 keV]
   (ii) Explain whether the K_α line of a molybdenum target will be observed with an X-ray tube run at 25 kV. [It will]

4. Calculate the binding energy per nucleon (in MeV/nucleon) of: (i) the $^{3}_{2}$He atom, and (ii) the $^{4}_{2}$He atom.
   [2,57 MeV/nucleon; 7,07 MeV/nucleon]

5. The rubidium isotope $^{87}_{37}$Rb, a β^- emitter with a half-life of $4{,}75 \times 10^{10}$ years, is used to determine the age of rocks and fossils.
   Rocks containing fossils of early animals are found to contain a ratio of $^{87}_{38}$Sr to $^{87}_{37}$Rb of 0,0160.
   Assuming that there was no $^{87}_{38}$Sr present when the rocks were formed, calculate the age of these fossils. [1,11 Gyr]

6. Freshly-bottled wine contains radioactive tritium which decays giving a count rate of 10 min^{-1} per kg of wine. Calculate the age of wine which gives a count rate of 8,3 min^{-1}.kg^{-1}.
   The half-life of tritium is 12,3 years. [3,3 yr]

7. ^{238}U decays into an isotope of thorium (Th) by emitting an α particle with energy 4,19 MeV and a γ ray with energy 0,048 MeV.
   (i) Write down the complete equation for the reaction.
   (ii) Calculate the atomic mass of the thorium isotope, assuming it to be at rest after the reaction. Atomic mass of ^{238}U = 238,0508 u. [234,0436 u]

8. Explain, using the data below, why $^{8}_{4}\text{Be}$ decays spontaneously into two α particles but $^{16}_{8}\text{O}$ does not decay spontaneously into four α particles.
Mass of: α particle = 4,00260 u, $^{8}_{4}\text{Be}$ = 8,00531 u, $^{16}_{8}\text{O}$ = 15,99491 u.

9. Lithium is a potential nuclear fuel on the basis of the following reaction:

$$^{6}_{3}\text{Li} + ^{2}_{1}\text{H} \rightarrow 2\ ^{4}_{2}\text{He}$$

Calculate the expected energy production from the consumption of 1,00 g of lithium, assuming 100% efficiency in the process.
Mass of ^{6}Li = 6,0151 u. [$3{,}59\times10^{11}$ J]

Hint: first calculate the mass decrease in the consumption of one atom of lithium.

PHYS 1001/1006 TUTORIALS
YEAR 2018
4th BLOCK

TUTORIALS TO PREPARE

WEEK: 10 September:	**No Tutorials converted to lecture**
WEEK: 17 September:	**Geometrical Optics I**
WEEK: 24 September:	**Geometrical Optics II**
WEEK: 01 October:	**Physical Optics/Modern Physics**
WEEK: 15 October:	**Physical Optics/Modern Physics**

- All students are expected to prepare solutions to these questions listed above prior to attending the tutorial session. Tutors will do a quick check. Preparation of tutorials will be recorded in the class registers as Not Prepared (NP), Partially Prepared (PP) and Well Prepared (WP).

- Students should form mini-groups of 3 and discussion within the groups should commence immediately on arrival at the tutorial venue. This discussion should be for a minimum of 20 minutes. Tutors will float around groups and assist students and will only give feedback on questions to the entire class that he/she might find as problematic. Note tutors are expected to give an overview of the solution and not complete solutions. Any problems not addressed will be carried over to the next week for discussion.

- A tutorial test will then be given at the end of each session (approximately 10 minutes). Tutors are expected to give feedback on the tutorial test at the beginning of the next session.

PHYS1001/1006 Course Co-ordinator

University of the Witwatersrand, Johannesburg

School of Physics

PHYS1001/6 (Physics I D)

Tutorial 6.1-3 – [2018] **Modern Physics**

Question 1

The rate of emission of energy of a certain kind of glow-worm is 1 μW.

If the average wavelength of the light radiated is 570 nm (i.e. in the visible spectrum), what is the rate of emission of photons by the glow-worm?

[2.9×10^{12} s^{-1}]

Question 2

Sodium has a work function of 2,3 eV.

(i) What is the maximum wavelength of light that will cause photoelectrons to be emitted from sodium?

(ii) What will the maximum kinetic energy of the photoelectrons be if light of wavelength 200 nm falls on a sodium surface.

[(i) 540 nm (ii) 6.26×10^{-19} J = 3.92 eV]

Question 3

Which one of the following statements about the photoelectric effect is true?

(a) The electrons are always emitted with zero energy.

(b) For fixed frequency above the threshold, the electron current is independent of the intensity of the incident light.

(c) For fixed frequency above the threshold, the maximum kinetic energy of the emitted electrons depends on the intensity of the incident light.

(d) For any intensity of light, the maximum kinetic energy of the emitted electrons depends on the frequency of the incident light.

(e) The cut-off frequency depends on the intensity of the incident light.

[d]

Question 4

The wavelength of the yellow sodium line is 589,6 nm.

What is the difference in energy between the two energy levels involved in the transition?

[3.37×10^{-19} J = 2.11 eV]

Questions 5

What accelerating voltage applied to a beam of electrons will just cause them to excite the characteristic radiation, with wavelength 0,063 nm, in a molybdenum target?

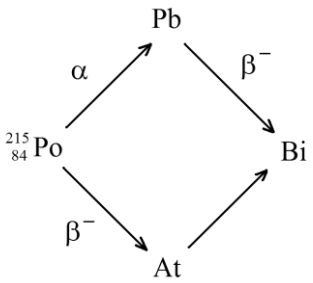
[19.7 kV]

Question 6

The binding energy per nucleon for the nucleus $^{7}_{3}\mathrm{Li}$ is 5,606 MeV. If the masses of the neutron and hydrogen atom are 1,00866 u and 1,00783 u respectively, what is the atomic mass of $^{7}_{3}\mathrm{Li}$?

[7.01598 u]

Question 7



Consider the decays shown in the diagram.

The symbols next to the arrows indicate the modes of decay.

(i) Give the atomic numbers and mass numbers of Pb, At and Bi in this series.

(ii) What is the mode of decay for the stage At to Bi? Explain your reasoning.

[(i) Pb: Z = 82, A = 211; At: Z = 85, A = 215; Bi: Z = 83, A =211 (ii) α]

Question 8

The intensity of a given X-ray beam is reduced by a factor of 8 by 12 mm of aluminium.

(i) Calculate the linear absorption coefficient of aluminium for this X-ray beam.

(ii) What thickness of aluminium would be required to reduce the intensity to 1% of its initial value?

[(i) 173 m^{-1} , (ii) 26.6 mm]

Question 9

A small volume of solution containing a radioactive isotope, of half-life 15 hours, had an activity of 185 Bq when injected into the bloodstream of a patient. After 30 hours the activity of 1 ml of blood was $9{,}25 \times 10^{-3}$ Bq.

What is the volume of blood in the patient?

[5 *l*]

Question 10

The equation ${}^{14}_{6}\mathrm{C} \rightarrow {}^{x}_{7}\mathrm{N} + y + z$ represents the β-decay reaction involved in radioactive carbon dating.

(i) What are *x*, *y* and *z* in the equation?

(ii) A burial site yields a wooden artefact which gives 11,6 counts per minute from ^{14}C per gram of carbon present. The corresponding count rate from wood from living trees is 15,3 per minute. Calculate the age of the artefact, given that the half-life of ^{14}C is 5730 years.

[(i) 14, β, ν (ii) 2290 yr]

Question 11

The sun radiates energy at the rate $6{,}46 \times 10^{7}$ W per square meter of surface area.

(i) If the energy emitted by the sun has its origin in nuclear fusion processes, calculate the rate at which the mass of the sun is decreasing. Diameter of sun = $1{,}39 \times 10^{6}$ km.

(ii) If all the energy comes from reactions in which two deuterium atoms fuse to form one helium atom, calculate the mass of helium produced per second.

[(i) 4.36×10^{9} kg.s^{-1} (ii) 6.81×10^{11} kg.s^{-1}]