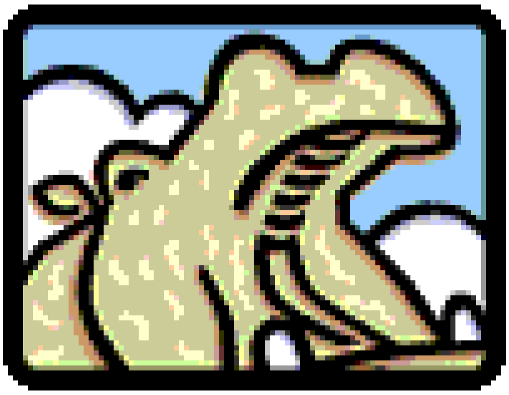
Revision April 2017

UPGRADE YOUR PHYSICS



NOTES FOR BRITISH SIXTH FORM STUDENTS WHO

- ARE PREPARING FOR THE INTERNATIONAL PHYSICS OLYMPIAD, OR
- WISH TO TAKE THEIR KNOWLEDGE OF PHYSICS BEYOND THE A-LEVEL SYLLABI.

A. C. MACHACEK

Introduction

The International Physics Olympiad is an annual international physics competition for pre-university students. Teams of five from each participating nation attend, and recently over 60 countries have taken part. Each nation has its own methods for selecting its team members. In Britain, this is by means of a series of written and practical exams. The question paper for the first round is circulated to all secondary schools.

Once the team has been chosen, it is necessary for its members to broaden their horizons. The syllabus for the International Physics Olympiad is larger than that of the British A2-level, and indeed forms a convenient stepping-stone to first year undergraduate work. For this reason, training is provided to help the British team bridge the gap.

The British Olympiad Committee recognizes the need for teaching material to help candidates prepare for the international competition. Furthermore, this material ought to have greater potential in the hands of students who wish to develop their physics, even if they have no desire to take part in the examinations.

It is my hope that these notes make a start in providing for this need.

A.C. Machacek, 2001

2006: Revision, update of rigid body dynamics, addition of Bernouilli’s equation and addition of questions.

2017: Revision to add a chapter of Ray Optics.

About the author: Anton Machacek competed in the International Physics Olympiad in 1993, and has been involved in training the British team in most of the subsequent years. He served on the academic committee for the International Physics Olympiad in Leicester in 2000. He is Assistant Head at Westcliff High School for Boys having previously led the Physics department at the Royal Grammar School in High Wycombe. He is a contributor to the Isaac Physics project at the University of Cambridge.

Contents

1 Linear Mechanics

1.1 Motion in a Line

1.1.1 The Fundamentals

1.1.1.1 Kinematics

Mechanics is all about motion. We start with the simplest kind of motion – the motion of small dots or particles. Such a particle is described completely by its mass (the amount of stuff it contains) and its position. There is no internal structure to worry about, and as for rotation, even if it tried it, no-one would see. The most convenient way of labelling the position is with a vector **r** showing its position with respect to some convenient agreed stationary point.

If the particle is moving, its position will change. If its speed and direction are steady, then we can write its position after time t as

$$\mathbf{r} = \mathbf{s} + \mathbf{u}t,$$

where **s** is the starting point (the position of the particle at t=0) and **u** as the change in position each second – otherwise known as the velocity. If the velocity is not constant, then we can't measure it by seeing how far the object goes in one second, since the velocity will have changed by then. Rather, we say that u how far the object would go in one second if the speed or direction remained unchanged that long. In practice, if the motion remains constant for some small time (called δt), and during this small time, the particle's position changes $\delta\mathbf{r}$, then the change in position if this were maintained for a whole second (otherwise known as the velocity) is

$$\mathbf{u} = \delta\mathbf{r} \times \text{number of } \delta t \text{ periods in one second} = \delta\mathbf{r} \div \delta\text{t}.$$

Similarly, if the velocity is changing, we define the acceleration as the change in velocity each second (if the rate of change of acceleration were constant. Accordingly, our equation for acceleration becomes

$$\mathbf{a} = \delta\mathbf{u} \div \delta t.$$

Hopefully, it is apparent that as the motion becomes more complex, and the δt periods need to be made shorter and shorter, we end up with the differential equations linking position, velocity and acceleration:

$$\begin{array}{ll} \mathbf{u} = \dfrac{d}{dt}\mathbf{r} & \mathbf{r} = \int \mathbf{u}\, dt \\ \mathbf{a} = \dfrac{d}{dt}\mathbf{u} & \mathbf{u} = \int \mathbf{a}\, dt \end{array}.$$

1.1.1.2 Dynamics

Now we have a way of describing motion, we need a way of predicting or explaining the motion which occurs – changing our question from 'what is happening?' to 'why?' and our explanation is going to involve the activity of forces. What do forces do to an object?

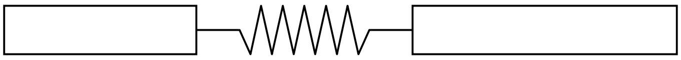
The first essential point is that forces are only needed to change (not maintain) motion. In other words – unless there is a change of velocity, no force is needed. But how much force is needed?

Newton made the assumption (which we find to be helpful and true) that the force causes a change in what he called the 'motion' –we now call it momentum. Suppose an object has mass m and velocity $\mathbf{u}$ (we shall clarify what we mean by mass later) – then its momentum is equal to $m\mathbf{u}$, and is frequently referred to by physicists by the letter $\mathbf{p}$. Newton's second law states that if a constant force $\mathbf{F}$ is applied to an object for a short time δt, then the change in the momentum is given by $\mathbf{F}\ \delta t$. In differential notation $\mathrm{d}(m\mathbf{u})/\mathrm{d}t = \mathbf{F}$.

In the case of a single object of constant mass it follows that

$$\mathbf{F} = \frac{d(m\mathbf{u})}{dt} = m\frac{d\mathbf{u}}{dt} = m\mathbf{a}\ .$$

His next assumption tells us more about forces and allows us to define 'mass' properly. Imagine two bricks are being pulled together by a strong spring. The brick on the left is being pulled to the right, the brick on the right is being pulled to the left.



Newton assumed that the force pulling the left brick rightwards is equal and opposite to the force pulling the right brick leftwards. To use more mathematical notation, if the force on block no.1 caused by block no.2 is called $\mathbf{f}_{12}$, then $\mathbf{f}_{12}=-\mathbf{f}_{21}$. If this were not the case, then if we looked at the bricks together as a whole object, the two internal forces would not cancel out, and there would be some 'left over' force which could accelerate the whole object.[1]

It makes sense that if the bricks are identical then they will accelerate together at the same rate. But what if they are not? This is where Newton's second law is helpful. If the resultant force on an object of

[1] If you want to prove that this is ridiculous, try lifting a large bucket while standing in it.

constant mass equals its mass times its acceleration, and if the two forces are equal and opposite, we say

$$\begin{aligned}\mathbf{f}_{12} &= -\mathbf{f}_{21} \\ m_1\mathbf{a}_1 &= -m_2\mathbf{a}_2\ , \\ \frac{a_1}{a_2} &= \frac{m_2}{m_1}\end{aligned}$$

and so the 'more massive' block accelerates less. This is the definition of mass. Using this equation, the mass of any object can be measured with respect to a standard kilogram. If a mystery mass experiences an acceleration of 2m/s^2 while pushing a standard kilogram in the absence of other forces, and at the same time the kilogram experiences an acceleration of 4m/s^2 the other way, then the mystery mass must be 2kg.

When we have a group of objects, we have the option of applying Newton's law to the objects individually or together. If we take a large group of objects, we find that the total force

$$\mathbf{F}_{\text{total}} = \sum_i \mathbf{F}_i = \frac{d}{dt}\sum_i m_i\mathbf{u}_i$$

changes the total momentum (just like the individual forces change the individual momenta). Note the simplification, though – there are no $\mathbf{f}_{ij}$ in the equation. This is because $\mathbf{f}_{ij} + \mathbf{f}_{ji} = 0$, so when we add up the forces, the internal forces sum to zero, and the total momentum is only affected by the external forces $\mathbf{F}_i$.

1.1.1.3 Energy and Power

Work is done (or energy is transferred) when a force moves something. The amount of work done (or amount of energy transformed) is given by the dot product of the force and the distance moved.

$$W = \mathbf{F} \bullet \mathbf{r} = F\,r\cos\theta \qquad (1)$$

where θ is the angle between the force vector **F** and the distance vector **r**. This means that if the force is perpendicular to the distance, there is no work done, no energy is transferred, and no fuel supply is needed.

If the force is constant in time, then equation (1) is all very well and good, however if the force is changing, we need to break the motion up into little parts, so that the force is more or less constant for each part. We may then write, more generally,

$$\delta W = \mathbf{F} \bullet \delta\mathbf{r} = F\,\delta r\cos\theta \qquad (1a)$$

Two useful differential equations can be formed from here.

1.1.1.4 Virtual Work

From equation (1a) it is clear that if the motion is in the direction of the force applied to the object (i.e. θ=0), then

$$\frac{\delta W}{\delta r} = F \text{ ,}$$

where W is the work done on the object. Accordingly, we can calculate the force on an object if we know the energy change involved in moving it. Let's give an example.

An electron (with charge *q*) is forced through a resistor (of length *L*) by a battery of voltage *V*. As it goes through, it must lose energy *qV*, since *V* is the energy loss per coulomb of charge passing through the resistor. Therefore, assuming that the force on the electron is constant (which we assume by the symmetry of the situation), then the force must be given by $\delta W / \delta d = qV / L$. If we define the electric field strength to be the force per coulomb of charge (*F/q*), then it follows that the electric field strength *E* = *V/L*.

So far, we have ignored the sign of *F*. It can not have escaped your attention that things generally fall downwards – in the direction of decreasing [gravitational] energy. In equations (1) and (1a), we used the vector **F** to represent the externally applied force we use to drag the object along. In the case of lifting a hodful of bricks to the top of a wall, this force will be directed upwards. If we are interested in the force of gravity **G** acting on the object (whether we drag it or not), this will be in the opposite direction. Therefore **F** = –**G**, and

$$\delta W = -\mathbf{G} \bullet \delta\mathbf{r}, \qquad (1b)$$

$$G = -\frac{\delta W}{\delta r}.$$

In other words, if an object can lose potential energy by moving from one place to another, there will always be a force trying to push it in this direction.

1.1.1.5 Power

Another useful equation can be derived if we differentiate (1a) with respect to time. The rate of 'working' is the power P, and so

$$P = \frac{\delta W}{\delta t} = \frac{\mathbf{F} \bullet \delta r}{\delta t} = \mathbf{F} \bullet \frac{\delta r}{\delta t}.$$

As we let the time period tend to zero, $\delta\mathbf{r}/\delta$t becomes the velocity, and so we have:

$$P = \mathbf{F} \bullet \mathbf{u} = F\,u \cos\theta \qquad (2)$$

where θ is now best thought of as the angle between force and direction of motion. Again we see that if the force is perpendicular to the direction of motion, no power is needed. This makes sense: think of a bike going round a corner at constant speed. A force is needed to turn the corner - that's why you lean into the bend, so that a component of your weight does the job. However no work is done – you don't need to pedal any harder, and your speed (and hence kinetic energy) does not change.

Equation (2) is also useful for working out the amount of fuel needed if a working force is to be maintained. Suppose a car engine is combating a friction force of 200N, and the car is travelling at a steady 30m/s. The engine power will be 200N × 30m/s = 6 kW.

Our equation can also be used to derive the kinetic energy. Think of starting the object from rest, and calculating the work needed to get it going at speed *U*. The force, causing the acceleration, will be **F**=m**a**. The work done is given by

$$\begin{aligned} W &= \int P\,dt = \int \mathbf{F} \bullet \mathbf{v}\,dt = \int m\frac{d\mathbf{v}}{dt} \bullet \mathbf{v}\,dt \\ &= \int m\mathbf{v} \bullet d\mathbf{v} = \left[\tfrac{1}{2}mv^2\right]_0^U = \tfrac{1}{2}mU^2 \end{aligned} \qquad (3)$$

although care needs to be taken justifying the integration stage in the multi-dimensional case.[2]

1.1.2 Changing Masses

The application of Newton's Laws to mechanics problems should pose you no trouble at all. However there are a couple of extra considerations which are worth thinking about, and which don't often get much attention at school.

The first situation we'll consider is when the mass of a moving object changes. In practice the mass of any self-propelling object will change as it uses up its fuel, and for accurate calculations we need to take this into account. There are two cases when this *must* be considered to get the answer even roughly right – jet aeroplanes and rockets. In the case of rockets, the fuel probably makes up 90% of the mass, so it must not be ignored.

[2] The proof is interesting. It turns out that $\mathbf{v} \bullet \mathbf{dv} \equiv |\mathbf{v}||\mathbf{dv}|\cos\theta = v\,dv$ since the change in *speed* dv is equal to |**dv**| cosθ where **dv** (note the bold type) is the vector giving the change in velocity.

Changing mass makes the physics interesting, because you need to think more carefully about Newton's second law. There are two ways of stating it – either

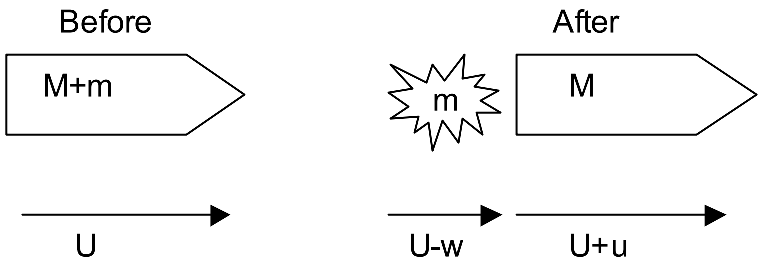
(i) Force on an object is equal to the rate of change of its momentum
(ii) Force on an object is equal to mass × acceleration

The first says $F = d(mu)/dt = m\dot{u} + \dot{m}u = ma + \dot{m}u$, whereas the second simply states $F=ma$. Clearly they can't both be correct, since they are different. Which is right? The first: which was actually the way Newton stated it in the first place! The good old $F=ma$ will still work – but you have to break the rocket into parts (say grams of fuel) – so that the rocket loses parts, but each part does not lose mass – and then apply $F=ma$ to each individual part. However if you want to apply a law of motion to the rocket as a whole, you have to use the more complicated form of equation.

This may be the first time that you encounter the fact that momentum is a more 'friendly' and fundamental quantity to work with mathematically than force. We shall see this in a more extreme form when looking at special relativity.

Let us now try and calculate how a rocket works. We'll ignore gravity and resistive forces to start with, and see how fast a rocket will go after it has burnt some fuel. To work this out we need to know two things – the exhaust speed of the combustion gas (*w*), which is always measured relative to the rocket; and the rate at which the motor burns fuel (in kg/s), which we shall call α.

We'll think about one part of the motion, when the rocket starts with mass (*M*+*m*), burns mass m of fuel, where m is very small, and in doing so increases its speed from *U* to *U*+*u*. This is shown below in the diagram.



Notice that the velocity of the burnt fuel is *U*-*w*, since *w* is the speed at which the combustion gas leaves the rocket (backwards), and we need to take the rocket speed *U* into account to find out how fast it is going relative to the ground.

Conservation of momentum tells us that

$$(M+m)\ U = m\ (U-w) + M\ (U+u)$$

so
$$m\,w = M\,u. \tag{4}$$

We can integrate this expression for *u* to evaluate the total change in speed after burning a large amount of fuel. We treat the *u* (change in *U*) as an infinitesimal calculus *dU*, and the *m* as a calculus *–dM*. Notice the minus sign – clearly the rocket must lose mass as fuel is burnt. Equation (4) now tells us

$$-w\frac{dM}{M} = dU \tag{5}$$

This can be integrated to give

$$\begin{aligned} -w\int\frac{1}{M}dM &= \int dU \\ -w[\ln M] &= [U] \\ U_{final} - U_{initial} &= w\ln\left(\frac{M_{initial}}{M_{final}}\right) \end{aligned} \tag{6}$$

This formula (6) is interesting because it tells us that in the absence of other forces, the gain in rocket speed depends *only* on the fraction of rocket mass that is fuel, and the exhaust speed.

In this calculation, we have ignored other forces. This is not a good idea if we want to work out the motion at blast off, since the Earth's gravity plays a major role! In order to take this, or other forces, into account, we need to calculate the thrust force of the rocket engine – a task we have avoided so far.

The thrust can be calculated by applying $F=ma$ to the (fixed mass) rocket M in our original calculation (4). The acceleration is given by $dU/t = u/t$, where t is the time taken to burn mass m of fuel. The thrust is

$$T = M\times\frac{u}{t} = M\times\frac{mw}{Mt} = \frac{mw}{t} = w\times\frac{m}{t} = w\alpha \tag{7}$$

given by the product of the exhaust speed and the rate of burning fuel. For a rocket of total mass M to take off vertically, T must be greater than the rocket's weight Mg. Therefore for lift off to occur at all we must have

$$w\alpha > Mg\,. \tag{8}$$

This explains why 'heavy' hydrocarbon fuels are nearly always used for the first stage of liquid fuel rockets. In the later stages, where absolute thrust is less important, hydrogen is used as it has a better ‘kick per kilogram’ because of its higher exhaust speed.

1.1.3 Fictitious Forces

Fictitious forces do not exist. So why do we need to give them a moment’s thought? Well, sometimes they make our life easier. Let’s have a couple of examples.

1.1.3.1 Centrifugal Force

You may have travelled in one of those fairground rides in which everyone stands against the inside of the curved wall of a cylinder, which then rotates about its axis. After a while, the floor drops out – and yet you don’t fall, because you’re “stuck to the side”. How does this work?

There are two ways of thinking about this. The first is to look at the situation from the stationary perspective of a friend on the ground. She sees you rotating, and knows that a centripetal force is needed to keep you going round – a force pointing towards the centre of the cylinder. This force is provided by the wall, and pushes you inwards. You feel this strongly if you’re the rider! And by Newton’s third law it is equally true that you are pushing outwards on the wall, and this is why you feel like you are being ‘thrown out’.

While this approach is correct, sometimes it makes the maths easier if you analyse the situation from the perspective of the rider. Then you don’t need to worry about the rotation! However in order to get the working right you have to include an outwards force – to balance the inward push of the wall. If this were not done, the force from the walls would throw you into the central space. The outward force is called the centrifugal force, and is our first example of a fictitious force. It doesn’t *really* exist, unless you are working in a rotating reference frame, and insist that you are at rest.

The difference between the two viewpoints is that in one case the inward push of the wall provides the centripetal acceleration. In the other it opposes the centrifugal force - giving zero resultant, and keeping the rider still. Therefore the formulae used to calculate centripetal force also give the correct magnitude for centrifugal force. The two differences are:

(i) Centrifugal force acts outwards, centripetal force acts inwards

(ii) Centrifugal force is *only* considered if you are assuming that the cylinder is at rest (in the cylinder’s reference frame). On the other hand, you *only* have centripetal accelerations if you do treat the cylinder as a moving object and work in the reference frame of a stationary observer.

This example also shows that fictitious forces generally act in the opposite direction to the acceleration that is being ‘ignored’. Here the acceleration is an inward centripetal acceleration, and the fictitious centrifugal force points outward.

1.1.3.2 Inertial Force

The second example we will look at is the motion of a lift (elevator) passenger. You know that you 'feel heavier' when the lift accelerates upwards, and 'feel lighter' when it accelerates downwards. Therefore if you want to simplify your maths by treating the lift car as a stationary box, you must include an extra downward force when the lift is actually accelerating upwards, and vice-versa. This fictitious force is called the inertial force. We see again that it acts in the opposite direction to the acceleration we are trying to ignore.

We shall look more closely at this situation, as it is much clearer mathematically.

Suppose we want to analyse the motion of a ball, say, thrown in the air in a lift car while it is accelerating upwards with acceleration $\mathbf{A}$. We use the vector $\mathbf{a}$ to represent the acceleration of the ball as a stationary observer would measure it, and $\boldsymbol{a'}$ to represent the acceleration as measured by someone in the lift. Therefore, $\boldsymbol{a} = \boldsymbol{A} + \boldsymbol{a}'$. Now this ball won't simply travel in a straight line, because forces act on it. Suppose the force on the ball is $\mathbf{F}$. We want to know what force $\mathbf{F'}$ is needed to get the right motion if we assume the lift to be at rest.

Newton's second law tells us that $\mathbf{F}$=m$\mathbf{a}$, if m is the mass of the ball. Therefore $\mathbf{F}$=m($\mathbf{A}$+$\mathbf{a'}$), and so $\mathbf{F}$-m$\mathbf{A}$ = m$\mathbf{a'}$. Now the force $\mathbf{F'}$ must be the force needed to give the ball acceleration $\mathbf{a'}$ (the motion relative to the lift car), and therefore $\mathbf{F'}$=m$\mathbf{a'}$. Combining these equations gives

$$\mathbf{F'} = \mathbf{F} - m\mathbf{A}. \qquad (9)$$

In other words, if working in the reference frame of the lift, you need to include not only the forces which are really acting on the ball (like gravity), but also an extra force –m$\mathbf{A}$. This extra force is the inertial force.

Let us continue this line of thought a little further. Suppose the only force on the ball was gravity. Therefore $\mathbf{F}$=m$\mathbf{g}$. Notice that

$$\mathbf{F'} = \mathbf{F} - m\mathbf{A} = m\,(\mathbf{g}\text{-}\mathbf{A}) \qquad (10)$$

and therefore if $\mathbf{g}$=$\mathbf{A}$ (that is, the lift is falling like a stone, because some nasty person has cut the cable), $\mathbf{F'}$=$\mathbf{0}$. In other words, the ball behaves as if no force (not even gravity) were acting on it, at least from the point of view of the unfortunate lift passengers. This is why weightlessness is experienced in free fall.

A similar argument can be used to explain the weightlessness of astronauts in orbiting spacecraft. As stationary observer (or a physics teacher) would say that there is only one force on the astronauts – gravity, and that this is just the right size to provide the centripetal force. The astronaut's perspective is a little different. He (or she) experiences two forces – gravity, and the fictitious centrifugal force. These two are equal and opposite, and as a result they add to zero, and so the astronaut feels just as weightless as the doomed lift passengers in the last paragraph.

1.2 *Going Orbital*

1.2.1 We have the potential

We shall now spend a bit of time reviewing gravity. This is a frequent topic of Olympiad questions, and is another area in which you should be able to do well with your A-level knowledge.

Gravitation causes all objects to attract all other objects. To simplify matters, we start with two small compact masses. The size of the force of attraction is best described by the equation

$$F_r = -\frac{GMm}{R^2} \tag{11}$$

Here G is the Gravitational constant (6.673×10^{-11} Nm^2/kg^2), M is the mass of one object (at the origin of coordinates), and m is the mass of the other. The equation gives the force experienced by the mass m. Notice the 'r' subscript and the minus sign – the force is radial, and directed inwards toward the origin (where the mass M is).

It is possible to work out how much work is needed to get the mass m as far away from M as possible. We use integration

$$\int_R^\infty \mathbf{F}\bullet\mathbf{dx} = \int_R^\infty -F_r\,dr = \int_R^\infty \frac{GMm}{r^2}dr = -\left[\frac{GMm}{r}\right]_R^\infty = \frac{GMm}{R}.$$

Notice the use of $-F_r$ in the second stage. In order to separate the masses we use a force $\mathbf{F}$ which acts in opposition to the gravitational attraction F_r. The equation gives the amount of work done by this force as it pulls the masses apart.

We usually define the zero of potential energy to be when the masses have nothing to do with each other (because they are so far away). Accordingly, the potential energy of the masses m and M is given by

$$E(R) = -\frac{GMm}{R}. \tag{12}$$

That is, GMm/R joules below zero energy. Notice that

$$F_r(R) = -\frac{dE(R)}{dR}. \tag{13}$$

This is a consequence of the definition of work as $W = \int \mathbf{F} \bullet \mathbf{dx}$, and is generally true. It is useful because it tells us that a forces always point in the direction of decreasing energy.

The potential energy depends on the mass of both objects as well as the position. The gravitational potential *V(R)* is defined as the energy per unit mass of the second object, and is given by

$$V(R) \equiv \lim_{m \to 0} \frac{E(R,m)}{m} = -\frac{GM}{R}. \tag{14}$$

Accordingly, the potential is a function only of position. The zero limit on the mass m is needed (in theory) to prevent the small mass disturbing the field. In practice this will not happen if the masses are fixed in position. To see the consequences of breaking this rule, think about measuring the Earth's gravitational field close to the Moon. If we do this by measuring the force experienced by a 1kg mass, we will be fine. If we do it by measuring the force experienced by a 10^{28}kg planet put in place for the job, we will radically change the motion of Earth and Moon, and thus affect the measurement.

In a similar way, we evaluate the gravitational field strength as the force per kilogram of mass. Writing the field strength as g gives

$$g = -\frac{MG}{R^2} \tag{15}$$

and equation (13) may be rewritten in terms of field and potential as

$$g(R) = -\frac{dV}{dR}. \tag{16}$$

1.2.2 Orbital tricks

There is a useful shortcut when doing problems about orbits. Suppose that an object of mass m is orbiting the centre of co-ordinates, and experiences an attractive force $F_r = -Ar^n$, where A is some constant. Therefore *n=-2* for gravity, and we would have *n=+1* for motion of a particle attached to a spring (the other end fixed at the origin).

If the object is performing circular orbits, the centripetal acceleration will be u^2/R where R is the radius of the orbit. This is provided by the attractive force mentioned, and so:

$$\begin{aligned} \frac{mu^2}{R} &= -F_r = AR^n \\ \frac{mu^2}{2} &= \frac{AR^{n+1}}{2} \end{aligned} \tag{17}$$

Now the potential energy E(R) is such that $dE/dR = -F_r = AR^n$, so

$$E(R) = \frac{AR^{n+1}}{n+1} \tag{18}$$

if we take the usual convention that E(R) is zero when the force is zero. Combining equations (17) and (18) gives

$$\frac{mu^2}{2} = \frac{n+1}{2} \times E(R) \tag{19}$$

so that

Kinetic Energy × 2 = Potential Energy × ($n+1$). (20)

This tells us that for circular gravitational orbits (where n=–2), the potential energy is twice as large as the kinetic energy, and is negative. For elliptical orbits, the equation still holds: but now in terms of the *average*[3] kinetic and potential energies. Equation (20) will not hold instantaneously at all times for non-circular orbits.

1.2.3 Kepler's Laws

The motion of the planets in the Solar system was observed extensively and accurately during the Renaissance, and Kepler formulated three "laws" to describe what the astronomers saw. For the Olympiad, you won't need to be able to derive these laws from the equations of gravity, but you will need to know them, and use them (without proof).

1. All planets orbit the Sun in elliptical orbits, with the Sun at one focus.

[3] By average, we refer to the mean energy in time. In other words, if T is the orbital period, the average of A is given by $\frac{1}{T}\int_0^T A(t)dt$.

2. The area traced out by the radius of an orbit in one second is the same for a planet, whatever stage of its orbit it is in. This is another way of saying that its angular momentum is constant, and we shall be looking at this in Chapter 3.

3. The time period of the orbit is related to the [time mean] average radius of the orbit: $T \propto \langle R \rangle^{3/2}$. It is not too difficult to show that this is true for circular orbits, but it is also true for elliptic ones.

1.2.4 Large Masses

In our work so far, we have assumed that all masses are very small in comparison to the distances between them. However, this is not always the case, as you will often be working with planets, and they are large! However there are two useful facts about large spheres and spherical shells. A spherical shell is a shape, like the skin of a balloon, which is bounded by two concentric spheres of different radius.

1. The gravitational field experienced at a point outside a sphere or spherical shell is the same as if all the mass of the shape were concentrated at its centre.

2. A spherical shell has no gravitational effect on an object inside it.

These rules only hold if the sphere or shell is of uniform density (strictly – if the density has spherical symmetry).

Therefore let us work out the gravitational force experienced by a miner down a very very very deep hole, who is half way to the centre of the Earth. We can ignore the mass above him, and therefore only count the bit below him. This is half the radius of the Earth, and therefore has one eighth of its mass (assuming the Earth has uniform density – which it doesn't). Therefore the M in equation (11) has been reduced by a factor of eight. Also the miner is twice as close to the centre (R has halved), and therefore by the inverse-square law, we would expect each kilogram of Earth to attract him four times as strongly. Combining the factors of 1/8 and 4, we arrive at the conclusion that he experiences a gravitational field ½ that at the Earth's surface, that is 4.9 N/kg.

1.3 *Fluids – when things get sticky*

Questions about fluids are *really* classical mechanics questions. You can tackle them without any detailed knowledge of fluid mechanics. There are a few points you need to remember or learn, and that is what this section contains. Perfect gases are also fluids, but we will deal with them in chapter 5 – "Hot Physics".

1.3.1 Floating and ... the opposite

The most important thing to remember is Archimedes' Principle, which states that:

> *When an object is immersed in a fluid (liquid or gas), it will experience an upwards force equal to the weight of fluid displaced.*

By “weight of fluid displaced” we mean the weight of the fluid that *would have been there* if the object was not in position. This upward force (sometimes called the buoyant upthrust) will be equal to

$$\text{Force} = \text{Weight of fluid displaced}$$

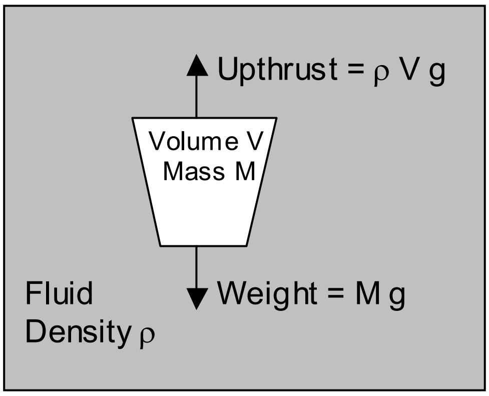
$$= g \times \text{Mass of fluid displaced}$$

$$= g \times \text{Density of fluid} \times \text{Volume of fluid displaced} \qquad (21)$$

For an object that is completely submerged, the “volume of fluid displaced” is the volume of the object.

For an object that is only partly submerged (like an iceberg or ship), the “volume of fluid displaced” is the volume of the object below the “waterline”.

This allows us to find out what will float, and what will sink. If an object is completely submerged, it will have two forces acting on it. Its weight, which pulls downwards, and the buoyant upthrust, which pulls upwards.



Object floats if:

$\rho V > M$
$\rho > M/V$

Therefore, things float if their overall density (total mass / total volume) is less than the density of the fluid. Notice that the overall density may not be equal to the actual density of the material. To give an example a ship is made of metal, but contains air, and is therefore able to float because its overall density is reduced by the air, and is therefore lower than the density of water. Puncture the hull, and the air is no longer held in place. Therefore the density of the ship = the density of the steel, and the ship sinks.

For an object that is floating on the surface of a fluid (like a ship on the ocean), the upthrust and weight must be equal – otherwise it would rise or fall. From Archimedes' principle, the weight of water displaced must equal the total weight of the object.

There is a "brain-teaser" question like this: A boat is floating in the middle of a lake, and the amount of water in the lake is fixed. The boat is carrying a large rock. The rock is lifted out of the boat, and dropped into the lake. Will the level of water in the lake go up or down?

Answer: Level goes down – while the rock was in the boat (and therefore floating) its *weight* of water was being displaced. When it was dropped into the depths, its *volume* of water was displaced. Now the density of rock is higher than that of water, so the water level in the lake was higher in the first case.

1.3.2 Under Pressure

What is the pressure in a fluid? This must depend on how deep you are, because the deeper you are, the greater weight of fluid you are supporting. We can think of the pressure (=Force/Area) as the weight of a square prism of fluid above a horizontal square metre marked out in the depths.

$$\begin{aligned}\text{Pressure} &= \text{Weight of fluid over } 1\text{m}^2 \text{ square}\\ &= g \times \text{Density} \times \text{Volume of fluid over } 1\text{m}^2\\ &= g \times \text{Density} \times \text{Depth} \times \text{Cross sectional area of fluid } (1\text{m}^2)\end{aligned}$$

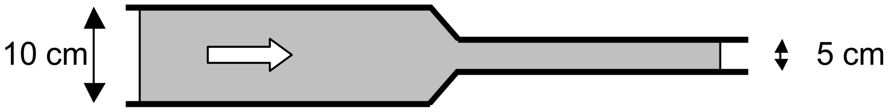
$$\text{Pressure} = g \times \text{Density} \times \text{Depth} \tag{22}$$

Of course, this equation assumes that there is nothing pushing down on the surface of the liquid! If there is, then this must be added in too. Therefore pressure 10m under the surface of the sea = atmospheric pressure + weight of a 10m high column of water.

It is wise to take a bit of caution, though, since pressures are often given relative to atmospheric pressure (i.e. 2MPa more than atmospheric) – and you need to keep your wits about you to spot whether pressures are relative (vacuum = -100 kPa) or absolute (vacuum = 0 Pa).

1.3.3 Continuity

Continuity means conservatism! Some things just don't change – like energy, momentum, and amount of stuff. This gives us a useful tool. Think about the diagram below, which shows water in a 10cm [diameter] pipe being forced into a 5cm pipe.



Water, like most liquids, doesn't compress much – so it can't form bottlenecks. The rate of water flow (cubic metres per second) in the big pipe must therefore be equal to the rate of water flow in the little pipe.

You might like to draw an analogy with the current in a series circuit. The light bulb has greater resistance than the wire but the current in both is the same, because the one feeds the other.

How can we express this mathematically? Let us assume that the pipe has a cross sectional area *A*, and the water is going at speed *u* m/s. How much water passes a point in 1 second? Let us put a marker in the water, which moves along with it. In one second it moves *u* metres. Therefore volume of water passing a point = volume of cylinder of length *u* and cross sectional area *A* = *u A*. Therefore

$$\text{Flow rate (m}^3\text{/s)} = \text{Speed (m/s)} \times \text{Cross sectional area (m}^2\text{)}. \quad (23)$$

Now we can go back to our original problem. The flow rate in both wide and narrow pipes must be the same. So if the larger one has twice the diameter, it has four times the cross sectional area; and so its water must be travelling *four times more slowly*.

1.3.4 Bernoulli's Equation

Something odd is going on in that pipe. As the water squeezes into the smaller radius, it speeds up. That means that its kinetic energy is increasing. Where is it getting the energy from? The answer is that it can only do so if the pressure in the narrower pipe is lower than in the wider pipe. That way there is an unbalanced force on the fluid in the cone-shaped part speeding it up. Lets follow a cubic metre of water through the system to work out how far the pressure drops.

The fluid in the larger pipe pushes the fluid in the cone to the right. The force = pressure × area = $P_L A_L$. A cubic metre of fluid occupies length *u* in the pipe, where *u* is the speed in m/s. Accordingly, the work done by the fluid in the wider pipe on the fluid in the cone in pushing the cubic metre through is $P_L A_L u_L = P_L$, since $u_L A_L = 1\text{m}^3$. However this cubic metre does work $P_R A_R u_R = P_R$ in getting out the other side. Thus the net energy gain of the cubic metre is $P_L - P_R$, and this must equal the change in the cubic metre's kinetic energy $\rho u_R^2/2 - \rho u_L^2/2$.

1.3.5 The Flow Equation

Equation (23) is also useful in the context of electric currents, and can be adapted into the so-called *flow equation*. Let us suppose that the fluid contains charged particles. Suppose that there are N of these particles per cubic metre of fluid, and each particle has a charge of q coulombs, then:

$$\begin{aligned}\text{Current} &= \text{Flow rate of charge (charge / second)}\\ &= \text{Charge per cubic metre (C/m}^3) \times \text{flow rate (m}^3\text{/s)}\\ &= N\,q \times \text{Area} \times \text{Speed}\ . \qquad (24)\end{aligned}$$

Among other things, this equation shows why the free electrons in a semiconducting material travel faster than those in a metal. If the semiconductor is in series with the metal, the current in both must be the same. However, the free charge density N is much smaller in the semiconductor, so the speed must be greater to compensate.

1.4 Questions

1. Calculate the work done in pedalling a bicycle 300m up a road inclined at 5° to the horizontal.

2. Calculate the power of engine when a locomotive pulls a train of 200 000kg up a 2° incline at a speed of 30m/s. Ignore the friction in the bearings. +

3. A trolley can move up and down a track. It's potential energy is given by V = $k\ x^4$, where x is the distance of the trolley from the centre of the track. Derive an expression for the force exerted on the trolley at any point. +

4. A ball bearing rests on a ramp fixed to the top of a car which is accelerating horizontally. The position of the ball bearing relative to the ramp is used as a measure of the acceleration of the car. Show that if the acceleration is to be proportional to the horizontal distance moved by the ball (measured relative to the ramp), then the ramp must be curved upwards in the shape of a parabola. ++

5. Use arguments similar to equation (3) to prove that the kinetic energy is still given by $\frac{1}{2}mu^2$ even when the force which has caused the acceleration from rest has not been applied uniformly in a constant direction. +

6. Calculate the final velocity of a rocket 60% of whose launch mass is propellant, where the exhaust velocity is 2000m/s. Repeat the calculation for a rocket where the propellant makes up 90% of the launch mass. In both cases neglect gravity.

7. Repeat question 6, now assuming that rockets need to move vertically in a uniform gravitational field of 9.8N/kg. Calculate the velocity at MECO (main engine cut-off) and the greatest height reached. Assume that both rockets have a mass of 10 000kg on the launch pad, and that the propellant is consumed evenly over one minute. ++

8. A 70kg woman stands on a set of bathroom scales in an elevator. Calculate the reading on the scales when the elevator starts accelerating upwards at 2m/s^2, when the elevator is going up at a steady speed, and when the elevator decelerates at 2m/s^2 before coming to a halt at the top floor of the building.

9. The woman in q8 is a juggler. Describe how she might have to adjust her throwing techniques in the elevator as it accelerates and decelerates.

10. Architectural models can not be properly tested for strength because they appear to be stronger than the real thing. To see why, consider a half-scale model of a building made out of the same materials. The weight is 1/8 of the real building, but the columns are ¼ the cross sectional area. Accordingly the stress on the columns is half of that in the full size building, and accordingly the model can withstand much more severe load before collapsing. To correct for this, a 1:300 architectural model is put on the end of a centrifuge arm of radius 10m which is spun around. The spinning ‘simulates’ an increased gravitational force which allows the model to be accurately tested. How many times will the centrifuge go round each minute?

11. Consider an incompressible fluid flowing from a 15cm diameter pipe into a 5cm diameter pipe. If the velocity and pressure before the constriction are 1m/s and 10 000 N/m^2, calculate the velocity and pressure in the constricted pipe. Neglect the effects of viscosity and turbulent flow. To work out the new pressure, remember that the increase in speed involves an increase of kinetic energy, and this energy must come from somewhere – so there will be a drop in pressure.

12. Calculate the orbital time period *T* of a satellite skimming the surface of a planet with radius *R* and made of a material with density ρ. Calculate the orbital speed for an astronaut skimming the surface of a comet with a 10km radius.

13. The alcohol percentage in wine can be determined from its density. A very light glass test tube (of cross sectional area 0.5cm^2) has 5g of lead pellets fixed to the bottom. You place the tube in the wine, lead first, and it floats with the open end of the tube above the surface of the wine. You can read the % alcohol from markings on the side of the tube. Calculate how far above the lead the 0% and 100% marks should be placed. The density of water is 1.00g/cm^3, while that of ethanol is

1.98g/cm^3. Where should the 50% line go? Remember that alcohol percentages are always volume percentages. +

2 Fast Physics

Imagine a summer's day. You are sunbathing by the side of a busy motorway while you wait for a pickup truck to rescue your car, which has broken down. All of a sudden, an irresponsible person throws a used drinks can out of their car window, and it heads in your direction. To make things worse, they were speeding at the time. Ouch.

The faster the car was going, the more it will hurt when the can hits you. This is because the can automatically takes up the speed the car was travelling at. Suppose the irresponsible person could throw the can at 10mph, and their car is going at 80mph. The speed of the can, as you see it, is 90mph if it was thrown forwards, and 70mph if it was thrown backwards.

To sum this up,

Velocity as measured by you = Velocity of car + Velocity of throwing

where we use velocities rather than speeds so that the directionality can be taken into account.

So far, this probably seems very obvious. However, let's extend the logic a bit further. Rather than a car, let us have a star, and in place of the drinks can, a beam of light. Many stars travel towards us at high speeds, and emit light as they do so. We can measure the speed of this light in a laboratory on Earth, and compare it with the speed of 'ordinary' light made in a stationary light bulb. And the worrying thing is that the two speeds are the same.

No matter how hard we try to change it, light always goes at the same speed.[4] This tells us that although our ideas of adding velocities are nice and straightforward, they are also wrong. In short, there is a problem with the Newtonian picture of motion. This problem is most obvious in the case of light, but it also occurs when anything else starts travelling very quickly.

While this is not the way Einstein approached the problem, it is our way into one of his early theories – the Special Theory of Relativity – and it is part of the Olympiad syllabus.

Before we go further and talk about what does happen when things go fast, please be aware of one thing. These observations will seem very

[4] Light does travel different speeds in different materials. However if the measurement is made in the same material (say, air or vacuum) the speed registered will always be the same, no matter what we do with the source.

weird if you haven't read them before. But don't dismiss relativity as nonsense just because it seems weird – it is a better description of Nature than classical mechanics – and as such it demands our respect and attention.

2.1 *The Principle of Relativity*

The theory of special relativity, like all theories, is founded on a premise or axiom. This axiom cannot be 'derived' – it is a guessed statement, which is the starting point for the maths and the philosophy. In the case of special relativity, the axiom must be helpful because its logical consequences agree well with experiments.

This principle, or axiom, can be stated in several ways, but they are effectively the same.

1. There is no method for measuring absolute (non-relative) velocity. The absolute speed of a car cannot be measured by any method at all. On the other hand, the speed of the car relative to a speed gun, the Earth, or the Sun can all be determined.

2. Since it can't be measured – there is no such thing as absolute velocity.

3. The 'laws of physics' hold in all non-accelerating laboratories[5], however 'fast' they may be going. This follows from statement 2, since if experiments only worked for one particular laboratory speed, that would somehow be a special speed, and absolute velocities could be determined relative to it.

4. Maxwell's theory of electromagnetism, which predicts the speed of light, counts as a law of physics. Therefore all laboratories will agree on the speed of light. It doesn't matter where or how the light was made, nor how fast the laboratory is moving.

2.2 *High Speed Observations*

In this section we are going to state what relativity predicts, as far as it affects simple observations. Please note that we are not deriving these statements from the principles in the last section, although this can be done. For the moment just try and understand what the statements mean. That is a hard enough job. Once you can use them, we shall then worry about where they come from.

[5] We say non-accelerating for a good reason. If the laboratory were accelerating, you would feel the 'inertial force', and thus you would be able to measure this acceleration, and indeed adjust the laboratory's motion until it were zero. However there is no equivalent way of measuring absolute speed.

2.2.1 Speeding objects look shortened in the direction of motion.

A metre stick comes hurtling towards you at high speed. With a clever arrangement of cameras and timers, you are able to measure its length as it passes you. If the stick's length is perpendicular to the direction of travel, you still measure the length as 1 metre.

However, if the stick is parallel to its motion, it will seem shorter to you. If we call the stick's actual length (as the stick sees it) as L_0, and the apparent length (as you measure it) L_a, we find

$$L_a = L_0\sqrt{1-\frac{u^2}{c^2}}\ , \tag{1}$$

where u is the speed of the metre stick relative to the observer. The object in the square root appears frequently in relativistic work, and to make our equations more concise, we write

$$\gamma \equiv \frac{1}{\sqrt{1-(u/c)^2}} \tag{2}$$

so that equation (1) appears in shorter form as

$$L_a = \frac{L_0}{\gamma}\,. \tag{3}$$

2.2.2 Speeding clocks tick slowly

A second observation is that if a clock whizzes past you, and you use another clever arrangement of timers and cameras to watch it, it will appear (to you) to be going slowly.

We may state this mathematically. Let T_0 be a time interval as measured by our (stationary) clock, and let T_a be the time interval as we see it measured by the whizzing clock.

$$T_a = \frac{T_0}{\gamma} \tag{4}$$

2.2.3 Slowing and shrinking go together

Equations (3) and (4) are consistent – you can't have one without the other. To see why this is the case, let us suppose that Andrew and Betty both have excellent clocks and metre sticks, and they wish to measure their relative speed as they pass each other. They must agree on the relative speed. Andrew times how long it takes Betty to travel along his metre stick, and Betty does the same.

The question is: how does Andrew settle his mind about Betty's calculation? As far as he is concerned, she has a short metre stick, and a slow clock – how can she possibly get the answer right! Very easily – providing that her clock runs 'slow' by the same amount that her metre stick is 'short'[6].

An experimental example may help clarify this. Muons are charged particles that are not stable, and decay with a half-life of 2μs. Because they are charged, you can accelerate them to high speeds using a large electric field in a particle accelerator. You can then measure how far they travel down a tube before decaying. Given that 'the laws of physics are the same in all reference frames', this must mean that muon and experimenter agree on the position in the tube at which the muon passes away.

The muon gets much further down the tube than a classical calculation would predict, however the reason for this can be explained in two ways:

- According to the experimenter, the muon is travelling fast, so it has a slow clock, and therefore lives longer – so it can get further.
- According to the muon, it still has a woefully short life, but the tube (which is whizzing past) is shorter so it appears to get further along in the 2μs.

For the two calculations to agree, the 'clock slowing' must be at the same rate as the 'tube shrinkage'.

2.2.4 Speeding adds weight to the argument

The most useful observation of them all, as far as the Olympiad syllabus is concerned is this: if someone throws a 1kg bag of sugar at you at high speed, and you (somehow) manage to measure its mass as it passes, you will register more than 1kg.

If the actual mass of the object is M_0, and the apparent mass is M_a, we find that

$$M_a = \gamma M_0 \,. \qquad (5)$$

The actual mass is usually called the 'rest mass' – in other words the mass as measured by an observer who is at rest with respect to the object.

[6] Note that 'slow' and 'short' are placed in quotation marks. Betty's clock and metre stick are not defective – however to Andrew they appear to be.

2.2.4.1 *The Universal Speed Limit*

This formula has important consequences. First of all, this is the origin of the 'universal speed limit', which is a well-known consequence of special relativity. This states that you will never measure the speed of an object (relative to you) as being greater than the speed of light.

Let us pause for a moment to see why. Suppose the object concerned is an electron in a particle accelerator (electrons currently hold the speed record on Earth for the fastest humanly accelerated objects). It starts at rest with a mass of about 10^{-30} kg. We turn on a large, constant electric field, and the electron starts to move relative to the accelerator. However, as it gets close to the speed of light, it starts to appear more massive. Therefore since our electric field (hence accelerating force) is constant, the electron's acceleration decreases. In fact, the acceleration tends to zero as time passes, although it never reaches zero exactly after a finite time. We are never able to persuade the electron to break the 'light-barrier', since when $u \to c$, $\gamma \to \infty$, and the apparent mass becomes very large (so the object becomes impossible to accelerate any further).

Please note that this does not mean that faster-than-light speeds can never be obtained. If we accelerate one electron to 0.6c Eastwards, and another to 0.6c Westwards, the approach speed of the two electrons is clearly superlumic (1.2c) as we measure it with Earth-bound speedometers. However, even in this case we find that the velocity of one of the electrons *as measured by the other* is still less than the speed of light. This is a consequence of our first observation – namely that relative velocities do not add in a simple way when the objects are moving quickly.

In fact the approach speed, as the electrons see it, is 0.882*c*. If you want to perform these calculations, the formula turns out to be

$$u_{AC} = \frac{u_{AB} + u_{BC}}{1 + u_{AB}u_{BC}/c^2}, \tag{6}$$

where u_{AB} means the velocity of B as measured by A. Equation (6) only applies when all three relative velocities are parallel (or antiparallel).

2.2.4.2 *Newton's Law of motion*

Our second consequence is that we need to take great care when using Newton's laws. We need to remember that the correct form of the second law is

$$F = \frac{d}{dt}\text{momentum} \tag{7}$$

Why is care needed? Look closely for the trap – if the object speeds up, its mass will increase. Therefore the time derivative of the mass needs to be included as well as the time derivative of the velocity. We shall postpone further discussion until we have had a better look at momentum.

2.3 Relativistic Quantities

Now that we have mentioned the business of relativistic mass increase, it is time to address the relativistic forms of other quantities.

2.3.1 Momentum

Momentum is conserved in relativistic collisions, providing we define it as the product of the apparent mass and the velocity.

$$\mathbf{p} = \gamma m_0 \mathbf{u} \tag{7}$$

Notice that when you use momentum conservation in collisions, you will have to watch the γ factors. Since these are functions of the speed *u*, they will change if the speed changes.

2.3.2 Force

The force on a particle is the time derivative of its momentum. Therefore

$$\mathbf{F} = \frac{d}{dt}\mathbf{p} = m_0\left(\gamma \frac{d}{dt}\mathbf{u} + \mathbf{u}\frac{d\gamma}{dt}\right). \tag{8}$$

In the case where the speed is not changing, γ will stay constant, and the equation reduces to the much more straightforward **F**=γm_0**a**. One example is the motion of an electron in a magnetic field.

2.3.3 Kinetic Energy

Now that we have an expression for force, we can integrate it with respect to distance to obtain the work done in accelerating a particle. As shown in section 1.1.1, this will give the kinetic energy of the particle. We obtain the result[7]

[7] If you wish to derive this yourself, here are the stages you need. Firstly, differentiate γ with respect to *u* to convince yourself that

$$\frac{d\gamma}{du} = \frac{u}{c^2\left(1-u^2/c^2\right)^{3/2}} = \frac{\gamma^3 u}{c^2} \quad \Rightarrow \quad \frac{du}{d\gamma} = \frac{c^2}{\gamma^3 u}.$$

Using this result, the derivation can be completed (see over the page):

$$K = (\gamma - 1)m_0 c^2 . \tag{9}$$

This states that the gain in energy of a particle when accelerated is equal to the gain in mass × c^2. From this we postulate that any increase in energy is accompanied by a change in mass. The argument works backwards too. When stationary, the particle had mass m_0. Surely therefore, it had energy m_0c^2 when at rest.

We therefore write the total energy of a particle as

$$E = K + m_0 c^2 = \gamma m_0 c^2 . \tag{10}$$

2.3.4 A Relativistic Toolkit

We can derive a very useful relationship from (10), (7) and the definition of γ:

$$\begin{aligned} E^2 - p^2c^2 &= \gamma^2 m_0^2 c^2 \left(c^2 - v^2\right) \\ &= \gamma^2 m_0^2 c^4 \left(1 - \left(\frac{v}{c}\right)^2\right). \\ &= m_0^2 c^4 \end{aligned} \tag{11}$$

This is useful, since it relates E and p without involving the nasty γ factor. Another equation which has no gammas in it can be derived by dividing momentum by total energy:

$$\frac{p}{E} = \frac{\gamma m_o u}{\gamma m_0 c^2} = \frac{u}{c^2} , \tag{12}$$

which is useful if you know the momentum and total energy, and wish to know the speed.

2.3.5 Tackling problems

If you have to solve a ‘collision’ type problem, avoid using speeds at all costs. If you insist on having speeds in your equations, you will also have gammas, and therefore headaches. So use the momenta and energies of the individual particles in your equations instead. Put more bluntly, you should write lots of ‘p’s, and ‘E’s, but no ‘u’s. Use the

$$\begin{aligned} K &= \int F dx = \int F u dt = \int m_0 u^2 \frac{d\gamma}{dt} dt + \int \gamma m_0 u \frac{du}{dt} dt \\ &= m_0 \int u \left(u + \gamma \frac{du}{d\gamma} \right) d\gamma = m_0 \int u \left(u + \frac{c^2}{\gamma^2 u} \right) d\gamma = m_0 \int u \frac{c^2}{u} d\gamma = \left[\gamma m_0 c^2\right] \end{aligned}$$

conservation laws to help you. In relativistic work, you can always use the conservation of *E* – even in non-elastic collisions. The interesting thing is that in an inelastic collision, you will find the rest masses greater after the collision.

To obtain the values you want, you need an equation which relates E and p, and this is provided by (11). Notice in particular that the quantity $\left(\sum E\right)^2 - \left(\sum p\right)^2 c^2$ when applied to a group of particles has two things to commend it.

- Firstly, it is only a function of total energy and momentum, and therefore will remain the same before and after the collision.

- Secondly, it is a function of the rest masses (see equation 11) and therefore will be the same in all reference frames.

Finally, if the question asks you for the final speeds, use (12) to calculate them from the momenta and energies.

2.4 The Lorentz Transforms

The facts outlined above (without the derivations) will give you all the information you need to tackle International Olympiad problems. However, you may be interested to find out how the observations of section 2.2 follow from the general assumptions of section 2.1. A full justification would require a whole book on relativity, however we can give a brief introduction to the method here.

We start by stating a general problem. Consider two frames of reference (or co-ordinate systems) – Andrew's perspective (t,x,y,z), and Betty's perspective (t',x',y',z'). We assume that Betty is shooting past Andrew in the +x direction at speed *v*. Suppose an 'event' happens, and Andrew measures its co-ordinates. How do we work out the co-ordinates Betty will measure?

The relationship between the two sets of co-ordinates is called the Lorentz transformation, and this can be derived as shown below:

2.4.1 Derivation of the Lorentz Transformation

We begin with the assumption that the co-ordinate transforms must be linear. The reason for this can be illustrated by considering length, although a similar argument works for time as well. Suppose that Andrew has two measuring sticks joined end to end, one of length L1 and one of length L2. He wants to work out how long Betty reckons they are. Suppose the transformation function is T. Therefore Betty measures the first rod as T(L1) and the second as T(L2). She therefore will see that the total length of the rods is T(L1) + T(L2). This must also be equal to T(L1+L2), since L1+L2 is the length of the whole rod according to Andrew. Since T(L1+L2) = T(L1) +T(L2), the transformation function is linear.

We can now get to work. Let us consider Betty's frame of reference to be moving in the +*x* direction at speed *v*, as measured by Andrew. Betty will therefore see Andrew moving in her –*x* direction at the same speed. To distinguish Betty's co-ordinates from Andrew's, we give hers dashes.

Given the linear nature of the transformation, we write

$$\begin{pmatrix} x' \\ t' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix}$$

where A, B, C and D are functions of the relative velocity +*v* (i.e. Betty's velocity as measured by Andrew).

There must also be an inverse transformation

$$\begin{pmatrix} x \\ t \end{pmatrix} = \frac{1}{d} \begin{pmatrix} D & -B \\ -C & A \end{pmatrix} \begin{pmatrix} x' \\ t' \end{pmatrix}$$

where *d* is the determinant of the first matrix.

Now this second matrix is in itself a transformation for a relative velocity –*v*, and therefore should be of a very similar form to the first matrix. We find that the only way we can ensure that there is symmetry between the two is to make the determinant equal to one (*d*=1). We shall therefore assume this from here on.

Next we consider what happens if *x'*=0. In other words we are tracing out Betty's motion as Andrew sees it. Therefore we must have *x*=*vt*. Using the first matrix, this tells us that *B*=-*vA*. A similar argument on the second matrix – where we must have *x'*=-*vt'* where Betty now watches Andrew's motion [*x*=*0*], gives –*Dv* = *B* = -*vA*. Therefore *A*=*D*.

We now have *B* and *D* expressed in terms of *A*, so the next job is to work out what *C* is. This can be done since we know that the determinant *AD* – *BC* = 1. Therefore we find that

$$C = \frac{1 - A^2}{vA}.$$

Summarizing, our matrix is now expressed totally in terms of the unknown variable *A*. We may calculate it by remembering that both Andrew and Betty will agree on the speed of travel, *c*, of a ray of light. Andrew will express this as *x*=*ct*, Betty would say *x'*=*ct'*, but both must be valid ways of describing the motion. Therefore

$$\begin{pmatrix} ct' \\ t' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} ct \\ t \end{pmatrix}$$

$$c = \frac{ct'}{t'} = \frac{Ac+B}{Cc+D}$$

$$(Cc+D)c = Ac+B$$

$$Cc^2 = B \text{ since A} = \text{D}$$

$$\left(\frac{1\text{ - A}^2}{vA}\right)c^2 = -vA$$

$$(1-A^2)c^2 = -v^2A^2$$

$$A = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$$

This concludes our reasoning, and gives the Lorentz transforms (after a little algebra to evaluate *C*) as:

$$\begin{aligned} x' &= \gamma(x - vt) \\ t' &= \gamma\left(t - \frac{xv}{c^2}\right) \\ x &= \gamma(x' + vt') \quad . \\ t &= \gamma\left(t' + \frac{x'v}{c^2}\right) \\ \gamma &\equiv \frac{1}{\sqrt{\left(1-\left[\frac{v}{c}\right]^2\right)}} \end{aligned}$$

We have not considered any other dimensions here, however the transformation here is easy since Andrew and Betty agree on all lengths in the *y* and *z* directions. In other words *y'*=*y*, *z'*=*z*. This is a necessary consequence of the principle of relativity: the distance between the ends of a rod held perpendicular to the direction of motion can be measured simultaneously in all frames of reference. If this agreed measurement was different to that of an identical rod in a different frame, the observers would be able to work out which of them was 'moving' and which of them was still.

2.4.2 Using the Lorentz Transforms

Having these transforms at our disposal, we can now derive the 'shrinking rod' and 'slowing clock' equations.

Suppose Betty is holding a stick (of length L) parallel to the x-axis. We want to know how long Andrew thinks it is. To measure it, he will measure where the ends of the rod are at a particular moment, and will then measure the distance between these points. Clearly the two positions need to be measured simultaneously in his frame of reference, and thus t is the same for both measurements. We know from that Betty thinks it has length L, and therefore Δx'=L. Using the first of the Lorentz equations (the one which links x', x and t), and remembering that t is the same for both measurements,

$$\Delta x' = \gamma \Delta x$$
$$L_{apparent} = \frac{L}{\gamma}.$$

Similarly we may show how a clock appears to slow down. Betty is carrying the clock, so it is stationary with respect to her, and x' (her measurement of the clock's position) will therefore be constant. The time interval shown on Betty's clock is Δt', while Andrew's own clock will measure time Δt. Here Δt' is the time Andrew sees elapsing on Betty's clock, and as such is equal to T_{apparent}. Using the fourth Lorentz equation (the one with x', t and t' in it), and remembering that x' remains constant, we have

$$\Delta t = \gamma \Delta t'$$
$$T_{apparent} = \frac{T}{\gamma}.$$

2.4.3 Four Vectors

The Lorentz transforms show you how to work out the relationships between the (t,x,y,z) co-ordinates measured in different frames of reference. We describe anything that transforms in the same way as a four vector, although strictly speaking we use (ct,x,y,z) so that all the components of the vector have the same units. Three other examples of four vectors are:

- (γc, γu_x, γu_y, γu_z) is called the *four velocity* of an object, and is the derivative of (ct, x,y,z) with respect to the proper time τ. Proper time is the time elapsed as measured in the rest frame of the object t=$\gamma\tau$.

- (mc,p_x,p_y,p_z) the momentum four vector. Here *m* is equal to γm_0. This must be a four vector since it is equal to the rest mass multiplied by the four velocity (which we already know to be a four vector).

- (ω/c,k$_x$,k$_y$,k$_z$) the wave four vector, where ω is the angular frequency of the wave (ω=2πf), and **k** is a vector which points in the direction the wave is going, and has magnitude 2π/λ. This can be derived from the momentum four vector in the case of a photon, since the momentum and total energy of a photon are related by E=pc, and the quantum theory states that E=hf=hω/2π and p=h/λ=hk/2π.

It also turns out that the dot product of any two four-vectors is 'frame-invariant' – in other words all observers will agree on its value. The dot product of two four-vectors is slightly different to the conventional dot product, as shown below:

$$(ct,x,y,z)\bullet(ct,x,y,z)\equiv x^2+y^2+z^2-(ct)^2 .$$

Notice that we subtract the product of the first elements.

The dot product of the position four vector with the wave four vector gives

$$(ct,x,y,z)\bullet(\omega/c,k_x,k_y,k_z)\equiv \mathbf{k}\bullet\mathbf{r}-\omega t .$$

Now this is the phase of the wave, and since all observers must agree whether a particular point is a peak, a trough or somewhere in between, then the phase must be an invariant quantity. Accordingly, since (ct,x,y,z) makes this invariant when 'dotted' with (ω/c,k$_x$,k$_y$,k$_z$), it follows that (ω/c,k$_x$,k$_y$,k$_z$) must be a four vector too.

2.5 Questions

1. Work out the relativistic γ factor for speeds of 1%, 50%, 90% and 99% of the speed of light.

2. Work out the speeds needed to give γ factors of 1.0, 1.1, 2.0, 10.0.

3. A muon travels at 90% of the speed of light down a pipe in a particle accelerator at a steady speed. How far would you expect it to travel in 2μs (a) without taking relativity into account, and (b) taking relativity into account?

4. A particle with rest mass *m* and momentum *p* collides with a stationary particle of mass *M*. The result is a single new particle of rest mass *R*. Calculate *R* in terms of *p* and *M*.

5. The principal runway at the spaceport on Arcturus-3 has white squares of side length 10m painted on it. A set of light sensors on the base of a spacecraft can take a 'picture' of the whole runway at the same time. What will the squares look like in the image if the spacecraft is passing the runway at a very high speed? Each sensor takes a picture of the runway directly underneath it, so you do not need to take into account the different times taken by light to reach the sensors from different parts of the runway.

6. When an electron is accelerated through a voltage V, its kinetic energy is given by eV where e is the size of the charge on the electron and is equal to 1.6×10^{-19}C. Taking the mass of the electron to be 9.1×10^{-31}kg, work out (a) the kinetic energy and speed of the electron when V=511kV (b) the kinetic energy and speed when V=20kV (c) the percentage error in the kinetic energy for V=20kV when calculated using the non-relativistic equation ½ mu^2.

7. Prove that the kinetic energy of a particle of rest mass m and speed u is given by ½ mu^2 if the speed is small enough in comparison to the speed of light. Work out the speed at which the non-relativistic calculation would be in error by 1%.

8. Suppose a spacecraft accelerates with constant acceleration a (as measured by the spacecraft's onboard accelerometers). At t=0 it is at rest with respect to a planet. Work out its speed relative to the planet as a function of time (a) as measured by clocks on the spacecraft, and (b) as measured by clocks on the planet. Note that the instantaneous speed of the craft relative to the planet will be agreed upon by spacecraft and planet.

3 Rotation

Rotational motion is all around us [groan] – from the acts of subatomic particles, to the motion of galaxies. Calculations involving rotations are no harder than linear mechanics; however the quantities we shall be talking about will be unfamiliar at first. Having already studied linear mechanics, you will be at a tremendous advantage, since we shall find that each ‘rule’ in linear mechanics has its rotational equivalent.

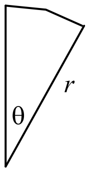
3.1 *Angle*

In linear mechanics, the most fundamental measurement is the position of the particle. The equivalent base of all rotational analysis is angle: the question “How far has the car moved?” being exchanged for “How far has the wheel gone round?” – a question which can only be answered by giving an angle. In mechanics, the radian is used for measuring angles. While you may be more familiar with the degree, the radian has many advantages.

We shall start, then by defining what we mean by a radian. Consider a sector of a circle, as in the diagram; and let the circle have a radius *r*. The length of the arc, that is the curved line in the sector, is clearly related to the angle. If the angle were made twice as large, the arc length would also double.

Can we use arc length to measure the angle? Not as it stands, since we haven’t taken into account the radius of the circle. Even for a fixed angle (say 30°), the arc will be longer on a larger circle. We therefore define the angle (in radians) as the arc length divided by the circle radius. Alternatively you might say that the angle in radians is equal to the length of the arc of a unit circle (that is a circle of 1m radius) that is cut by the angle.

Arc length = $r\theta$ if θ is measured in radians

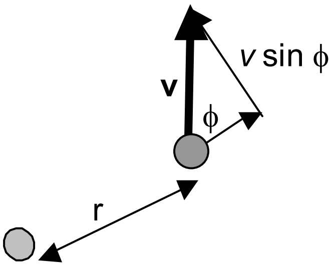


Notice one simplification that this brings. If a wheel, of radius *R*, rolls a distance *d* along a road, the angle the wheel has turned through is given by *d/R* in radians. Were you to calculate the angle in degrees, there would be nasty factors of 180 and π in the answer.

Before getting too involved with radians, however, we must work out a conversion factor so that angles in degrees can be expressed in radians. To do this, remember that a full circle (360°) has a circumference or arc length of $2\pi r$. So 360°=2π rad. Therefore, 1 radian is equivalent to (360/2π)° = (180/π)°.

3.2 Angular Velocity

Having discussed angle as the rotational equivalent of position, we now turn our attention to speed. In linear work, speeds are given in metres per second – the distance moved in unit time. For rotation, we speak of 'angular velocity', which tells us how fast something is spinning: how many radians it turns through in one second. The angular velocity can also be thought of as the derivative of angle with respect to time, and as such is sometimes written as $\dot{\theta}$, however more commonly the Greek letter ω is used, and the dot is avoided. To check your understanding of this, try and show that 1 rpm (revolution per minute) is equivalent to $\pi/30$ rad/s, while one cycle per second is equivalent to 2π rad/s.



Now remember the definition of angle in radians, and that the distance moved by a point on the rim of a wheel will move a distance $s = r\theta$ when the wheel rotates by an angle θ. The speed of the point will therefore be given by $u = \mathrm{d}s/\mathrm{d}t = r\ \mathrm{d}\theta/\mathrm{d}t = r\omega$.

For a point that is not fixed to the wheel, the situation is a little more complex. Suppose that the point has a velocity **v**, which makes an angle ϕ to the radius (as in the figure above). We then separate **v** into two components, one radial ($v \cos \phi$) and one rotational ($v \sin \phi$). Clearly the latter is the only one that contributes to the angular velocity, and therefore in this more general case, $v \sin \phi = r\omega$.

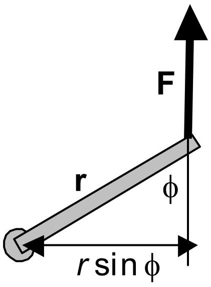
3.3 Angular acceleration

It should come as no surprise that the angular acceleration is the time derivative of ω, and represents the change in angular velocity (in rad/s) divided by the time taken for the change (in s). It is measured in rad/s^2, and denoted by α or $\dot{\omega}$ or $\ddot{\theta}$. For an object fastened to the rim of a wheel, the 'actual' acceleration round the rim (a) will be given by $a = \mathrm{d}u/\mathrm{d}t = r\ \mathrm{d}\omega/\mathrm{d}t = r\alpha$, while for an object not fastened, we have $a \sin \phi = r\alpha$.[8]

3.4 Torque – Angular Force

Before we can start 'doing mechanics' with angles, we need to consider the rotational equivalent of force – the amount of twist. Often a twist can be applied to a system by a linear force, and this gives us a 'way in' to the analysis. We say that the strength of the twist is called the 'moment' of the force, and is equal to the size of the force multiplied by the distance from pivot to the point where the force acts. A complication

[8] Here we are not including the centripetal acceleration which is directed towards the centre of the rotation.



arises if the force is not tangential – clearly a force acting along the radius of a wheel will not turn it – and so our simple 'moment' equation needs modifying.[9]

There are two ways of proceeding, and they yield the same answer. Suppose the force F makes an angle ϕ with the radius. We can break this down into two components – one of magnitude F cos ϕ, which is radial and does no turning; and the other, tangential component (which does contribute to the turning) of magnitude F sin ϕ. The moment or torque only includes the relevant component, and so the torque is given by $C = Fr \sin \phi$.

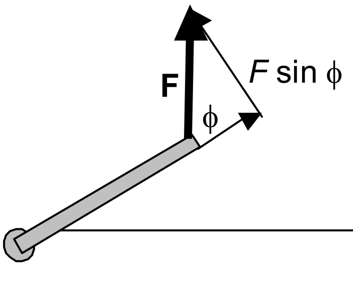
The alternative way of viewing the situation is not to measure the distance from the centre to the *point at which the force is applied*. Instead, we draw the force as a long line, and to take the distance as the *perpendicular* distance from force line to centre. The diagram shows that this new distance is given by r sin ϕ, and since the force here is completely tangential, we may write the moment or torque as the product of the full force and this perpendicular distance – i.e. $C = F\, r \sin \phi$, as before.

3.5 Moment of Inertia – Angular Mass

Of the three base quantities of motion, namely distance, mass and time, only time may be used with impunity in rotational problems. We now have an angular equivalent for distance (namely angle), so the next task is to determine an angular equivalent for mass.

This can be done by analogy with linear mechanics, where the mass of an object in kilograms can be determined by pushing an object, and calculating the ratio of the applied force to the acceleration it caused: $m = F/a$. Given that we now have angular equivalents for force and acceleration, we can use these to find out the 'angular mass'.

Think about a ball of mass m fixed to the rim of a wheel that is accelerating with angular acceleration α. We shall ignore the mass of the wheel itself for now. Now let us push the mass round the wheel with a force F. Therefore we calculate the 'angular mass' I by



[9] Why force × radius? We can use a virtual work argument (as in section 1.1.1.4) to help us. Suppose a tangential force F is applied at radius r. When the object moves round by angle θ, it moves a distance $d = \theta r$, and the work done by the force = $Fd = F\theta r = Fr \times \theta =$ angular force × angular distance. Now since energy must be the same sort of thing with rotational motion as linear, the rotational equivalent of force must be Fr.

$$I = \frac{C}{\alpha} = \frac{rF\sin\phi}{(a\sin\phi/r)} = r^2\,\frac{F}{a} = r^2\,m$$

where we have used the fact that the mass *m* will be the ratio of the force *F* to the linear acceleration *a*, as dictated by Newton's Second Law. This formula can also be used for solid objects, however in this case, the radius *r* will be the perpendicular distance from the mass to the axis. The total 'angular mass' of the object is calculated by adding up the $I = m\,r^2$ from each of the points it is made from.

Usually this 'angular mass' is called the **moment of inertia** of the object. Notice that it doesn't just depend on the mass, but also on the distance from the point to the centre. Therefore the moment of inertia of an object *depends on the axis* it is spun round.

An object may have a high angular inertia, therefore, for two reasons. Either it is heavy in its own right; or for a lighter object, the mass is a long way from the axis.

3.6 *Angular Momentum*

In linear motion, we make frequent use of the 'momentum' of objects. The momentum is given by mass × velocity, and changes when a force is applied to the object. The force applied, is in fact the time derivative of the momentum (provided that the mass doesn't change). Frequent use is made of the fact that total momentum is conserved in collisions, provided that there is no *external* force acting.

It would be useful to find a similar 'thing' for angular motion. The most sensible starting guess is to try 'angular mass' × angular velocity. We shall call this the angular momentum, and give it the symbol $L = I\,\omega$. Let us now investigate how the angular momentum changes when a torque is applied. For the moment, assume that *I* remains constant.

$$\frac{dL}{dt} = \frac{d}{dt}I\omega = I\frac{d\omega}{dt} = I\alpha = C$$

Thus we see that, like in linear motion, the time derivative of angular momentum is 'angular force' or torque. Two of the important facts that stem from this statement are:

1. If there is no torque *C*, the angular momentum will not change. Notice that radial forces have *C* = 0, and therefore will not change the angular momentum. This result may seem unimportant – but think of the planets in their orbits round the Sun. The tremendous force exerted on them by the Sun's gravity is *radial*, and therefore does not change their angular momenta even a smidgen. We can therefore calculate the velocity of planets at different parts of their orbits using the fact that the angular momentum will remain the same. This principle also holds when

scientists calculate the path of space probes sent out to investigate the Solar System.

2. The calculation above assumes that the moment of inertia *I* of the object remains the same. This seems sensible, after all, in a linear collision, the instantaneous change of a single object's mass would be bizarre[10], and therefore we don't need to guard against the possibility of a change in mass when we write $F = \mathrm{d}p/\mathrm{d}t$.

In the case of angular motion, this situation is different. The moment of inertia can be changed, simply by rearranging the mass of the object closer to the axis. Clearly there is no *external* torque in doing this, so we should expect the angular momentum to stay the same. But if the mass has been moved closer to the axis, *I* will have got smaller. Therefore ω must have got bigger. The object will now be spinning faster! This is what happens when a spinning ice dancer brings in her/his arms – and the corresponding increase in revs. per minute is well known to ice enthusiasts and TV viewers alike.

To take an example, suppose that all the masses were moved twice as close to the axis. The value of *r* would halve, so *I* would be quartered. We should therefore expect ω to get four times larger. This is in fact what happens.

3.7 Angular momentum of a single mass moving in a straight line

If we wished to calculate the angular momentum of a planet in its orbit round the Sun, we need to know how *L* is related to the linear speed *v*. This is what we will now work out.

Using the same ideas as in figure 2, the velocity **v** will have both radial and 'rotational' components. The rotational component will be equal to $v \sin\phi$, while the radial component cannot contribute to the angular momentum. It is the rotational component that corresponds to the speed of a mass fixed to the rim of a wheel, and as such is equal to radius × angular velocity. Thus $v \sin\phi = r\,\omega$. So the angular momentum

$$L = I\omega = mr^2 \times \frac{v\sin\phi}{r} = mvr\sin\phi$$

[10] Two cautions. Firstly, in a rocket, the mass of the rocket does decrease as the burnt fuel is chucked out the back, however the total mass does not change. Therefore $F=\mathrm{d}p/\mathrm{d}t=ma$ still works, we just need to be careful that the force *F* acts on (and only on) the stuff included in the mass *m*. A complication *does* arise when objects start travelling at a good fraction of the speed of light – but this is dealt with in the section on Special Relativity.

is given by the product of the mass, the radius and the rotational (or tangential) component of the velocity.

For an object on a *straight line path*, this can also be stated (using figure 3) as the mass × speed × distance of closest approach to centre.

3.8 Rotational Kinetic Energy

Lastly, we come to the calculation of the rotational kinetic energy. We may calculate this by adding up the linear kinetic energies of the parts of the object as the spin round the axis. Notice that in this calculation, as the objects are purely rotating, we shall assume $\phi = \pi/2$ – i.e. there is no radial motion.

$$K = \tfrac{1}{2}mv^2 = \tfrac{1}{2}m(r\omega)^2 = \tfrac{1}{2}mr^2\omega^2 = \tfrac{1}{2}I\omega^2$$

We see that the kinetic energy is given by half the angular mass × angular velocity squared – which is a direct equivalent with the half mass × speed^2 of linear motion.

3.9 Summary of Quantities

Quantity	Symbol	Unit	Definition	Other equations
Angular velocity	ω	rad/s	$\omega = d\theta/dt$	$r\,\omega = v \sin \phi$
Angular acceleration	α	rad/s^2	$\alpha = d\omega/dt$	$r\,\alpha = a \sin \phi$
Torque	C	N m	$C = F\,r \sin \phi$	
Moment of inertia	I	kg m^2	$I = C / \alpha$	$I = m\,r^2$
Angular momentum	L	kg m^2 /s	$L = I\,\omega$	$L = m\,v\,r \sin \phi$
Rot. Kinetic Energy	K	J	$K = I\,\omega^2 / 2$	$K = ½m\,(v \sin \phi)^2$

3.10 *Rotational mechanics with vectors*

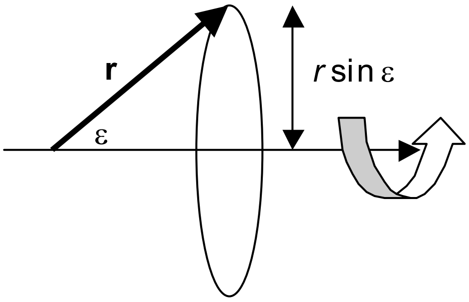
This section involves much more advanced mathematics, and you will be able to get by in Olympiad problems perfectly well without it. However, if you like vectors and matrices, read on...

So far we have just considered rotations in one plane – that of the paper. In general, of course, rotations can occur about any axis, and to describe this three dimensional situation, we use vectors. With velocity **v**, momentum **p** and force **F**, there is an obvious direction – the direction of motion, or the direction of the 'push'. With rotation, the 'direction' is less clear.

Imagine a clock face on this paper, with the minute hand rotating clockwise. What direction do we associate with this motion? Up towards 12 o'clock because the hand sometimes points that way? Towards 3 o'clock because the hand sometimes points that way? Both are equally ridiculous. In fact the only way of choosing a direction that will always apply is to assign the rotation 'direction' *perpendicular* to the clock face – the direction in which the hands *never* point.

This has not resolved our difficulty completely. Should the arrow point upwards out of the paper, or down into it? After thought we realise that one should be used for clockwise and one for anti-clockwise motion, but which way? There is no way of proceeding based on logic – we just have to accept a convention. The custom is to say that for a clockwise rotation, the 'direction' is down away from us, and for anticlockwise rotation, the direction is up towards us.

Various aides-memoire have been presented for this – my favourite is to consider a screw. When turned clockwise it moves away from you: when turned anticlockwise it moves towards you. For this reason the convention is sometimes called the 'right hand screw rule'.



With this convention established, we can now use vectors for angular velocity ω, angular momentum **L**, and torque **C**. Kinetic energy, like in linear motion, is a scalar and therefore needs no further attention. The moment of inertia *I* is more complex, and we shall come to that later.

Let us consider the angular velocity first. If we already know ω and **r**, what is **v**, assuming that only rotational velocities are allowed? Remembering that w must point along the axis of the rotation, we may draw the diagram above, which shows that the radius of the circle that our particle actually traces out is $r \sin \varepsilon$ where ε is the angle between **r** and ω.[11] This factor of $\sin \varepsilon$ did not arise before in this way, since our motion was restricted to the plane which contained the centre point, and thus $\varepsilon = \pi/2$ for all our 2-dimensional work. Therefore the velocity is equal to w multiplied by the radius of the circle traced out, i.e. $v = \omega\, r \sin \varepsilon$. This may be put on a solid mathematical foundation using the vector cross product namely $\mathbf{v} = \omega\times\mathbf{r}$. This is our first vector identity for rotational motion.

By a similar method, we may analyse the acceleration. We come to the corresponding conclusion $\mathbf{a} = \alpha \times \mathbf{r}$.[12]

Next we tackle torque. Noting our direction convention, and our earlier equation $C = F\, r \sin \phi$, we set $\mathbf{C} = \mathbf{r} \times \mathbf{F}$. Similarly, from $L = (mv)\, r \sin \phi = p\, r \sin \phi$, we set $\mathbf{L} = \mathbf{r} \times \mathbf{p}$.

With these three vector equations we may get to work. Firstly, notice:

$$\frac{d}{dt}\mathbf{L} = \frac{d}{dt}(\mathbf{r} \times \mathbf{p}) = \mathbf{v} \times \mathbf{p} + \mathbf{r} \times \frac{d}{dt}\mathbf{p} = \mathbf{0} + \mathbf{r} \times \frac{d}{dt}(m\mathbf{v}) = \mathbf{r} \times m\mathbf{a} = \mathbf{r} \times \mathbf{F} = \mathbf{C}$$

The time derivative of angular momentum is the torque, as before. Notice too that the (**v×p**) term disappears since **p** has the same direction as **v**, and the vector cross product of two parallel vectors is zero.

3.10.1.1 General Moment of Inertia

Our next task is to work out the moment of inertia. This can be more complex, since it is not a vector. Previously we defined I by the relationships $C = I\,\alpha$, and also used the expression $L = I\,\omega$. Now that **C**, α, **L** and ω are vectors, we conclude that I must be a matrix, since a vector is made when I is multiplied by the vectors α or ω. Our aim is to find the matrix that does the job.

For this, we use our vector equations $\mathbf{v} = \omega \times \mathbf{r}$ and $\mathbf{C} = \mathbf{r} \times \mathbf{F}$, we let the components of **r** be (x,y,z), and we also use the mathematical result that for any three vectors **A**, **B** and **C**, $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \bullet \mathbf{C})\mathbf{B} - (\mathbf{A} \bullet \mathbf{B})\mathbf{C}$.

[11] We use ε to represent the angle between **r** and ω, to distinguish it from the angle ϕ between **r** and **v**, which is of course a right angle for a strict rotation.

[12] This intentionally does not include the centripetal acceleration, as before. If you aim to calculate this **a** from the former equation $\mathbf{v} = \omega \times \mathbf{r}$, then you get $\mathbf{a} = d\mathbf{v}/dt = d(\omega\times\mathbf{r})/dt = \alpha\times\mathbf{r} + \omega\times\mathbf{v} = \alpha\times\mathbf{r} + \omega\times(\omega\times\mathbf{r}) = \alpha\times\mathbf{r} + \omega\,(\mathbf{r}.\omega) - \mathbf{r}\,\omega^2$. The final two terms in this equation deal with the centripetal acceleration. However in real situations, the centripetal force is usually provided by internal or reaction forces, so often problems are simplified by not including it.

$$\begin{aligned}\mathbf{C} &= \mathbf{r} \times \mathbf{F} = m(\mathbf{r} \times \mathbf{a}) \\ &= m[\mathbf{r} \times (\boldsymbol{\alpha} \times \mathbf{r})] \\ &= mr^2\boldsymbol{\alpha} - m(\mathbf{r} \bullet \boldsymbol{\alpha})\mathbf{r} \\ &= m\begin{pmatrix} r^2 - x^2 & -xy & -xz \\ -xy & r^2 - y^2 & -yz \\ -xz & -yz & r^2 - z^2 \end{pmatrix}\boldsymbol{\alpha} \\ &= m\begin{pmatrix} y^2 + z^2 & -xy & -xz \\ -xy & x^2 + z^2 & -yz \\ -xz & -yz & x^2 + y^2 \end{pmatrix}\boldsymbol{\alpha}\end{aligned}$$

This result looks horrible. However let us simplify matters by aligning our axes so that the *z* axis is the axis of the acceleration α. In other words α = (0,0,α). We now have

$$\mathbf{C} = m\begin{pmatrix} -xz \\ -yz \\ x^2 + y^2 \end{pmatrix}\alpha$$

which is a little better. Notice that it is still pretty nasty in that the torque required to cause this *z*-rotation acceleration is not necessarily in the *z*-direction! Another consequence of this is that the angular momentum **L** is not necessarily parallel to the angular velocity ω. However for many objects, we rotate them about an axis of symmetry. In this case the *xz* and *yz* terms become zero when summed for all the masses in the object, and what we are left with is the mass multiplied by the distance from the axis to the masses (that is $x^2 + y^2$). Alternatively, for a flat object (called a lamina) which has no thickness in the *z* direction, the *xz* and *yz* terms are zero anyway, because *z*=0.

At this point, you are perfectly justified in saying ‘yuk’ and sticking to two-dimensional problems. However this result we have just looked at has interesting consequences. When a 3-d object has little symmetry, it can roll around in some very odd ways. Some of the asteroids and planetary moons in our Solar System are cases in point.

The moment of inertia can also be obtained from the rotational momentum, however, the form is identical to that worked out above from Newton’s second law, as shown here.

$$\begin{aligned}\mathbf{L} &= \mathbf{r} \times \mathbf{p} = m(\mathbf{r} \times \mathbf{v}) \\ &= m[\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r})] \\ &= mr^2\boldsymbol{\omega} - m(\mathbf{r} \bullet \boldsymbol{\omega})\mathbf{r}\end{aligned}$$

The calculation then proceeds as before.

3.10.1.2 General Kinetic Energy

Our final detail is kinetic energy. This can be calculated using **v** = ω × **r**, and the vector rule that $\mathbf{A}\bullet(\mathbf{B}\times\mathbf{C})=\mathbf{B}\bullet(\mathbf{C}\times\mathbf{A})$.

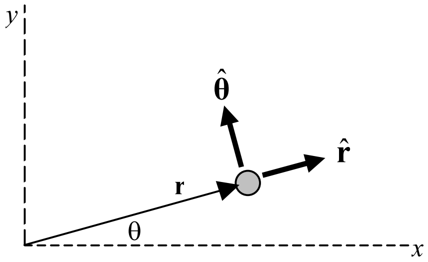
$$\begin{aligned}K &= \tfrac{1}{2}mv^2 = \tfrac{1}{2}m(\mathbf{v}\bullet\mathbf{v})\\ &= \tfrac{1}{2}m[\mathbf{v}\bullet(\boldsymbol{\omega}\times\mathbf{r})]\\ &= \tfrac{1}{2}m[\boldsymbol{\omega}\bullet(\mathbf{r}\times\mathbf{v})]\\ &= \tfrac{1}{2}\boldsymbol{\omega}\bullet(\mathbf{r}\times m\mathbf{v})\\ &= \tfrac{1}{2}\boldsymbol{\omega}\bullet(\mathbf{r}\times\mathbf{p})\\ &= \tfrac{1}{2}\boldsymbol{\omega}\bullet\mathbf{L} = \tfrac{1}{2}\boldsymbol{\omega}\bullet I\boldsymbol{\omega}\end{aligned}$$

For the cases where I can be simplified, this reduces to the familiar form $K = I\,\omega^2/2$.

3.11 Motion in Polar Co-ordinates

When a system is rotating, it often makes sense to use polar co-ordinates. In other words, we characterise position by its distance from the centre of rotation (i.e. the radius *r*) and by the angle θ it has turned through. Conversion between these co-ordinates and our usual Cartesian (*x,y*) form are given by simple trigonometry:

$$\begin{aligned}x &= r\cos\theta\\ y &= r\sin\theta\end{aligned} \tag{1}$$



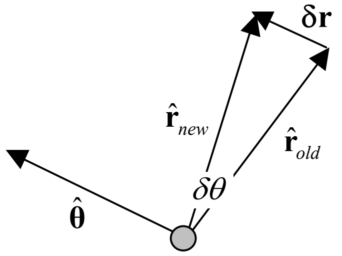
When analysing motion problems, though, there are complications if polar co-ordinates are used. These stem from the fact that the ‘increasing *r*’ and ‘increasing θ’ directions themselves depend on the value of θ, as we shall see. Let us start by defining the vector **r** to be the position of a particle relative to some convenient origin. The length of this vector *r* gives the distance from particle to origin. We define $\hat{\mathbf{r}}$ to be a unit vector parallel to **r**. Similarly, we define the vector $\hat{\boldsymbol{\theta}}$ to be a unit vector pointing in the direction the particle would have to go in order to increase θ while keeping *r* constant. Let us now evaluate the time derivative of **r** – in other words, let’s find the velocity of the particle:

$$\frac{d}{dt}\mathbf{r} = \frac{d(r\,\hat{\mathbf{r}})}{dt} = \frac{dr}{dt}\hat{\mathbf{r}} + r\frac{d\,\hat{\mathbf{r}}}{dt} = \dot{r}\,\hat{\mathbf{r}} + r\frac{d\,\hat{\mathbf{r}}}{dt}, \qquad (2)$$

where we have used the dot above a letter to mean 'time derivative of'. Now if the particle does not change its θ, then the direction $\hat{\mathbf{r}}$ will not change either, and we have a velocity given simply by $\dot{r}\,\hat{\mathbf{r}}$. We next consider the case when *r* doesn't change, and the particle goes in a circle around the origin. In this case, our formula would say that the velocity was $r\frac{d\,\hat{\mathbf{r}}}{dt}$. We know from section 3.2 that in this case, the speed is given by *r*ω, that is $r\dot{\theta}$, so the velocity will be $r\dot{\theta}\,\hat{\boldsymbol{\theta}}$. In order to make this agree with our equation for d**r**/dt, we would need to say that

$$\frac{d}{dt}\hat{\mathbf{r}} = \dot{\theta}\,\hat{\boldsymbol{\theta}}. \qquad (3)$$

Does this make sense? If you think about it for a moment, you should find that it does. Look at the diagram below. Here the angle θ has changed a small amount δθ. The old and new $\hat{\mathbf{r}}$ vectors are shown, and form two sides of an isosceles triangle, the angle between them being δθ. Given that the sides $\hat{\mathbf{r}}$ have length 1, the length of the third side is going to be approximately equal to δθ (with the approximation getting better the smaller δθ is). Notice also that the third side – the vector corresponding to $\hat{\mathbf{r}}_{new} - \hat{\mathbf{r}}_{old}$ is pointing in the direction of $\hat{\boldsymbol{\theta}}$. This allows us to justify statement (1).



In a similar way, we may show that

$$\frac{d}{dt}\hat{\boldsymbol{\theta}} = -\dot{\theta}\,\hat{\mathbf{r}}. \qquad (4)$$

Remembering that our velocity is given by

$$\mathbf{v} = \dot{\mathbf{r}} = \dot{r}\,\hat{\mathbf{r}} + r\dot{\theta}\,\hat{\boldsymbol{\theta}},$$

we may calculate the acceleration as

$$\begin{aligned}\mathbf{a} = \dot{\mathbf{v}} &= \ddot{r}\,\hat{\mathbf{r}} + \dot{r}\dot{\theta}\,\hat{\boldsymbol{\theta}} + r\ddot{\theta}\,\hat{\boldsymbol{\theta}} + \dot{r}\frac{d\,\hat{\mathbf{r}}}{dt} + r\dot{\theta}\frac{d\,\hat{\boldsymbol{\theta}}}{dt} \\ &= \ddot{r}\,\hat{\mathbf{r}} + \dot{r}\dot{\theta}\,\hat{\boldsymbol{\theta}} + r\ddot{\theta}\,\hat{\boldsymbol{\theta}} + \dot{r}\,\dot{\theta}\,\hat{\boldsymbol{\theta}} - r\dot{\theta}^2\hat{\mathbf{r}} \qquad . \\ &= \left(\ddot{r} - r\dot{\theta}^2\right)\hat{\mathbf{r}} + \left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)\hat{\boldsymbol{\theta}}\end{aligned} \tag{5}$$

Now suppose that a force acting on the particle (with mass m), had a radial component F_r, and a tangential component F_θ. We could then write

$$\begin{aligned}F_r &= m\left(\ddot{r} - r\dot{\theta}^2\right) \\ F_\theta &= m\left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)\end{aligned}. \tag{6}$$

There are many consequences of these equations for rotational motion. Here are three:

1. For an object to go round in a circle (that is *r* staying constant, so that $\dot{r} = \ddot{r} = 0$), we require a non-zero radial force $F_r = -mr\dot{\theta}^2$. The minus sign indicates that the force is to be in the opposite direction to **r**, in other words pointing towards the centre. This, of course, is the *centripetal* force needed to keep an object going around in a circle at constant speed.

2. If the force is purely radial (we call this a central force), like gravitational attraction, then $F_\theta = 0$. It follows that

$$\begin{aligned}0 &= mr\ddot{\theta} + 2m\dot{r}\dot{\theta} \\ &= mr^2\ddot{\theta} + 2mr\dot{r}\dot{\theta}\,, \\ &= \frac{d}{dt}\left(mr^2\,\dot{\theta}\right)\end{aligned} \tag{7}$$

   and accordingly the angular momentum $mr^2\dot{\theta} = mr^2\omega$ does not change. This ought to be no surprise, since we found in section 3.6 that angular momenta are only changed if there is a torque, and a radial force has zero torque.

3. One consequence of the conservation of angular momentum is the apparently odd behaviour of an object coming obliquely towards the centre (that is, it gets closer to the origin, but is not aimed to hit it). Since *r* decreases, ω must increase, and this is what happens – in fact the square term causes ω to quadruple when *r* halves.

   We can analyse this in terms of forces using (6): $F_\theta = m\left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)$ when $F_\theta = 0$. Since *r* is decreasing, while θ increases, the non-zero value of the $2m\dot{r}\dot{\theta}$ term gives rise to a non-zero $\ddot{\theta}$, and hence an acceleration of rotation. If you were sitting next to the particle at the time, you would wonder what caused it to speed up, and you would think that there must have been a force acting upon it.

This is another example of a fictitious force (see section 1.1.3), and is called the Coriolis force. It is used, among other things, to explain why the air rushing in to fill a low pressure area of the atmosphere begins to rotate – thus setting up a ‘cyclone’. Some people have attempted to use the equation to explain the direction of rotation of the whirlpool you get above the plughole in a bath.

Put very bluntly – the Coriolis force is the force needed to ‘keep’ the object going in a true straight line. Of course, a stationary observer would see no force – after all things go in straight lines when there are *no* sideways forces acting on them. The perspective of a rotating observer is not as clear – and this Coriolis force will be felt to be as real as the centrifugal force discussed in section 1.1.3.1.

3.12 *Motion of a rigid body*

When you are dealing with a rigid body, things are simplified in that it can only do two things – move in a line and rotate. If forces $\mathbf{F}_i$ are applied to positions $\mathbf{r}_i$ on a solid object free to move, its motion is completely described by

- a linear acceleration given by $\mathbf{a} = \sum \mathbf{F}_i \big/ M$, where *M* is the total mass of the body, and
- a rotational acceleration given by $\boldsymbol{\alpha} = \sum \mathbf{r}_i' \times \mathbf{F}_i \big/ I$ about a point called the centre of mass, where $\mathbf{r}_i$’ is the position of point *i* relative to the centre of mass and *I* is the moment of inertia of the object about the axis of rotation.[13]

This means, among other things, that the centre of mass itself moves as if it were a point particle of mass *M*. In turn, if a force is applied to the object at the centre of mass, it will cause the body to move with a linear acceleration, without any rotational acceleration at all.

The proof goes as follows. Suppose the object is made up of lots of points $\mathbf{r}_i$ (of mass m_i) fixed together. It follows that Newton’s second law states (as in section 1.1.1.2)

[13] This assumes that the angular acceleration is a simple speeding up or slowing down of an existing rotation. If α and ω are not parallel, the situation is more complex.

$$\sum \frac{d(m_i \mathbf{u}_i)}{dt} = \sum \mathbf{F}_i$$
$$\sum \frac{d^2(m_i \mathbf{r}_i)}{dt^2} = \sum \mathbf{F}_i$$
$$\frac{d^2 \sum m_i \mathbf{r}_i}{dt^2} = \sum \mathbf{F}_i$$

Now suppose we define the position **R** such that $M\mathbf{R} = \Sigma\, m_i\, \mathbf{r}_i$, then it follows that

$$M \frac{d^2\mathbf{R}}{dt^2} = \mathbf{F}_{\text{total}}$$

and the point **R** moves as if it were a single point of mass M being acted on by the total force. This position **R** is called the centre of mass.

Given that we already know that **R** does not have any rotational motion, this must be the centre of rotation, and we can use the equation from section 3.10 to show that the rate of change of angular momentum of the object about this point, $\mathrm{d}(I\omega)/\mathrm{d}t$, is equal to the total torque $\Sigma\ (\mathbf{r}_i - \mathbf{R})\times\mathbf{F}_i$ acting on the body about the point **R**. Given that the masses don't change, we may write

$$\begin{aligned}
\frac{d}{dt} m_i \mathbf{u}_i &= \mathbf{F}_i + \sum_j \mathbf{f}_{ij} \\
m_i \mathbf{a}_i &= \mathbf{F}_i + \sum_j \mathbf{f}_{ij} \\
m_i \mathbf{r}_i \times \mathbf{a}_i &= \mathbf{r}_i \times \mathbf{F}_i + \mathbf{r}_i \times \sum_j \mathbf{f}_{ij} \\
\sum_i m_i \mathbf{r}_i \times \mathbf{a}_i &= \sum_i \mathbf{r}_i \times \mathbf{F}_i + \sum_{ij} \mathbf{r}_i \times \mathbf{f}_{ij}
\end{aligned}.$$

The final term sums to zero since $\mathbf{f}_{ij}+\mathbf{f}_{ji}=\mathbf{0}$, and the internal forces between two particles must either constitute a repulsion, an attraction or the two forces must occur at the same place. In any of these cases $\mathbf{f}_{ij} \times (\mathbf{r}_i - \mathbf{r}j) = \mathbf{0}$.

If we now express the positions $\mathbf{r}_i$ in terms of the centre of mass position **R** and a relative position $\mathbf{r}_i'$, where $\mathbf{r}_i = \mathbf{R} + \mathbf{r}_i'$ (so $\mathbf{a}_i = \mathbf{A}+\mathbf{a}_i'$), then

$$\sum m_i\left(\mathbf{R}+\mathbf{r}_i{}'\right)\times\left(\mathbf{A}+\mathbf{a}_i{}'\right)=\sum\left(\mathbf{R}+\mathbf{r}_i{}'\right)\times\mathbf{F}_i$$

$$\sum m_i\mathbf{R}\times\mathbf{a}_i+\sum m_i\mathbf{r}_i{}'\times\mathbf{A}+\sum m_i\mathbf{r}_i{}'\times\mathbf{a}_i{}'=\sum\mathbf{R}\times\mathbf{F}_i+\sum\mathbf{r}_i{}'\times\mathbf{F}_i$$

$$\sum\mathbf{R}\times m_i\mathbf{a}_i+0+\sum m_i\mathbf{r}_i{}'\times\mathbf{a}_i{}'=\sum\mathbf{R}\times\mathbf{F}_i+\sum\mathbf{r}_i{}'\times\mathbf{F}_i$$

$$\sum m_i\mathbf{r}_i{}'\times\mathbf{a}_i{}'=\sum\mathbf{r}_i{}'\times\mathbf{F}_i$$

since $\Sigma m_i\mathbf{r}_i'$ = $\Sigma m_i(\mathbf{r}_i–\mathbf{R})$ = $M\mathbf{R}$ – $M\mathbf{R}$ = $\mathbf{0}$. Now, as shown earlier, $d(m_i\mathbf{r}_i\times\mathbf{u}_i)/dt=m_i\mathbf{u}_i\times\mathbf{u}_i+m_i\mathbf{r}_i\times\mathbf{a}_i=0+m_i\mathbf{r}_i\times\mathbf{a}_i$, and so

$$\frac{d}{dt}\sum\mathbf{r}_i{}'\times\mathbf{u}_i=\sum\mathbf{r}_i{}'\times\mathbf{F}_i$$

$$\frac{d}{dt}\mathbf{L}'=\frac{d}{dt}I\boldsymbol{\omega}=\sum\mathbf{r}_i{}'\times\mathbf{F}_i=\mathbf{C}'$$

and so the rate of change of angular momentum about the centre of mass is given by the total moment of the external forces about the centre of mass.

3.13 Questions

1. A car has wheels with radius 30cm. The car travels 42km. By what angle have the wheels rotated during the journey? Make sure that you give your answer in radians and in degrees.

2. Why does the gravitational attraction to the Sun not change the angular momentum of the Earth?

3. Calculate the speed of a satellite orbiting the Earth at a distance of 42 000km from the Earth's centre.

4. A space agency plans to build a spacecraft in the form of a cylinder 50m in radius. The cylinder will be spun so that astronauts inside can walk on the inside of the curved surface as if in a gravitational field of 9.8 N/kg. Calculate the angular velocity needed to achieve this.

5. A television company wants to put a satellite into a 42 000km radius orbit round the Earth. The satellite is launched into a circular low-Earth orbit 200km above the Earth's surface, and a rocket motor then speeds it up. It then coasts until it is in the 42000km orbit with the correct speed. How fast does it need to be going in the low-Earth orbit in order to coast up to the correct position and speed?

6. Estimate the gain in angular velocity when an ice-skater draws her hands in towards her body.

7. One theory of planet formation says that the Earth was once a liquid globule which gradually solidified, and its rotation as a liquid caused it to bulge outwards in the middle – a situation which remains to this day: the equatorial radius of the Earth is about 20km larger than the polar radius. If the theory were correct, what would the rotation rate of the Earth have been just before the crust solidified? Assume that the liquid globule was sufficiently viscous that it was all rotating at the same angular velocity.

4 Vibes and Wiggles

4.1 Oscillation

Any system in stable equilibrium can be persuaded to oscillate. If it is removed from the equilibrium, there will be a force (or other influence) that attempts to maintain the status quo. The size of the force will depend on the amount of the disturbance.

Suppose that the disturbance is called *x*. The restoring force can be written

$$F = -\left(Ax + Bx^2 + Cx^3 \ldots\right), \tag{1}$$

where the minus sign indicates that the force acts in the opposite direction to the disturbance. If *x* is small enough, x^2 and x^3 will be so small that they can be neglected. We then have a restoring force proportional to the displacement *x*.

Just because the system has a force acting to restore the equilibrium, this does not mean that it will return to *x*=0 immediately. All systems have some inertia, or reluctance to act quickly. For a literal force, this inertia is the mass of the system – and we know that the acceleration caused by a force (*F*) is given by *F/m*, where *m* is the mass. We can therefore work on equation (1) to find out more:

$$\begin{gathered} F = m\frac{d^2x}{dt^2} = -Ax \\ \frac{d^2x}{dt^2} = -\frac{A}{m}x \end{gathered}\quad . \tag{2}$$

This differential equation has the solution:

$$\begin{gathered} x = x_0 \cos(\omega t + \phi) \\ \omega = \sqrt{A/m} \end{gathered}\quad , \tag{3}$$

which is indeed an oscillation. We are using x_0 to denote the amplitude. Notice that, as we are working in radians, the cosine function needs to advance 2π to go through a whole cycle. Therefore we can work out the time period (T) and frequency (f):

$$\begin{gathered} \omega T = 2\pi \\ T = \frac{2\pi}{\omega} \\ f = \frac{1}{T} = \frac{\omega}{2\pi} \end{gathered}\quad . \tag{4}$$

Seeing that $\omega=2\pi f$, we notice that ω is none other than the angular frequency of the oscillation, as defined in chapter 3.

These equations are perfectly general, and so whenever you come across a system with a differential equation like (2), you know the system will oscillate, and furthermore you can calculate the frequency.

4.1.1 Non-linearity

Equation (1) has left an unanswered question. What happens if *x* is big enough that x^2 and x^3 can't be neglected? Clearly solution (3) will no longer work. In fact the equation probably won't have a simple solution, and the system will start doing some really outrageous things. Given that it has quadratic terms in it, we say it is non-linear; and a non-linear equation will send most physics students running away, screaming for mercy.

Let me give you an example. There are very nice materials that look harmlessly transparent. However they are designed so that the non-linear terms are *very* important when light passes through them. The result – you put red laser light in, and it comes out blue (at twice the frequency). They are called 'doubling crystals' and are the sort of thing that might freak out an unsuspecting GCSE examiner.

Our world would be much less wonderful if it were purely linear – no swirls in smoke, no wave-breaking (and hence surfing), and extremely boring weather – not to mention rigid population dynamics. While the non-linear terms add to the spice of life, I for one am grateful that many phenomena can be well described using linear equations. Otherwise physics would be much more frustrating, and bridge design would be just as hard as predicting the weather.

4.1.2 Energy

Before we move from oscillations to waves, let us make one further observation. The energy involved in the oscillation is proportional to the square of the amplitude. We shall show this in two ways.

First: If the displacement is given by equation (3), we notice that the velocity is given by

$$u \equiv \dot{x} = -x_0 \omega \sin(\omega t + \phi). \qquad (5)$$

At the moment when the system passes through its equilibrium (x=0) point, all the energy is in kinetic form. Therefore the total energy is

$$E = K(x = x_0) = \tfrac{1}{2} m u^2 = \tfrac{1}{2} m \omega^2 x_0^2 \qquad (6)$$

which is indeed proportional to the amplitude squared.

Second: When the displacement is at its maximum, there is no kinetic energy. The energy will all be in potential form. We can work out the potential energy in the system at displacement *x*, by evaluating the work done to get it there:

$$E_{pot} = \int F dx = \int A x dx = \left[\tfrac{1}{2} A x^2\right] . \qquad (7)$$

Notice that we did not include the minus sign on the force. This is because when we work out the 'work done' the force involved is the force of us pulling the system. This is equal and opposite to the restoring force of the system, and as such is positive (directed in the same direction as *x*).

The total energy is given by the potential energy at the moment when *x* has its maximum (i.e. x=x_0). Therefore

$$E = E_{pot}(x = x_0) = \tfrac{1}{2} A x_0^2 . \qquad (8)$$

Equations (8) and (6) are in agreement. This can be shown by inserting the value of ω from equation (3) into (6).

While we have only demonstrated that energy is proportional to amplitude squared for an oscillation, it turns out that the same is true for linked oscillators – and hence for waves. The intensity of a wave (joules of energy transmitted per second) is proportional to the amplitude squared in exactly the same way.

Intensity of a wave is also related to another wave property – its speed. The intensity is equal to the amount of energy stored on a length *u* of wave, where *u* is the speed. This is because this is the energy that will pass a point in one second (a length *u* of wave will pass in this time).

4.2 *Waves & Interference*

The most wonderful property of waves is that they can interfere. You can add three and four and get six, or one, or 4.567, depending on the phase relationship between the two waves. You can visualize this using either trigonometry or vectors (phasors). However, before we look at interference in detail, we analyse a general wave.

4.2.1 Wave number

Firstly, we define a useful parameter called the *wave number*. This is usually given the letter *k*, and is defined as

$$k = \frac{2\pi}{\lambda} , \qquad (13)$$

where λ is the wavelength. If we write the shape of a 'paused' wave as $y=A\cos(\phi)$, the phase ϕ of a wave is given by

$$\phi = kD. \tag{14}$$

We can see that this makes sense by combining equations (13) and (14):

$$\phi = kD = \frac{2\pi D}{\lambda}. \tag{15}$$

If the distance *D* is equal to a whole wavelength, we expect the wave to be doing the same thing as it was at $\phi=0$. And since $\cos(2\pi)=\cos(0)$, this is indeed the case.

A variation on the theme is possible. You may also see *wave vectors* **k**: these have magnitude as defined in (13), and point in the direction of energy transfer.

4.2.2 Wave equations

We are now in a position to write a general equation for the motion of a wave with angular frequency ω and wave number *k*:

$$y = A\cos(\omega t - kx + \phi).$$

We can check that this is correct, since

- if we look at a particular point (value of *x*), and watch as time passes, we will pass from one peak to the next when $\omega t = 2\pi f\, t$ has got bigger by 2π (i.e. $t=1/f$ as it should).

- if we look at a particular moment in time (value of *t*), and look at the position of adjacent peaks, they should be separated by one wavelength = $2\pi/k$. Now for adjacent peaks, the values of *kx* will differ by 2π according to the formula above, and so this is correct.

- if we follow a particular peak on the wave – say the one where $\omega t - kx + f = 0$, we notice that $x = (\omega t + f)/k = \omega t/k$ + constant, and hence the position moves to increasing *x* at a speed equal to ω/k, as indeed it should since $\omega/k = 2\pi f/(2\pi/\lambda) = f\lambda = v$.

It follows that a leftwards-travelling wave has a function which looks like

$$y = A\cos(\omega t + kx + \phi).$$

4.2.3 Standing waves

Imagine we have two waves of equal amplitude passing along a string in the two different directions. The total effect of both waves is given by adding them up:

$$\begin{aligned} y &= A\cos(\omega t - kx + \phi_1) + A\cos(\omega t + kx + \phi_2) \\ &= 2\cos\left(\omega t + \tfrac{1}{2}(\phi_1 + \phi_2)\right) \times \cos\left(kx + \tfrac{1}{2}(\phi_2 - \phi_1)\right) \end{aligned}$$

At any time, the peaks and troughs will only occur at the places where the second cosine is +1 or –1, and so the positions of the peaks and troughs do not change. This is why this kind of situation is called a standing wave. While there is motion, described by the first cosine, the positions of constructive interference between the two counter-propagating waves remain fixed (these are called antinodes), as to the positions of destructive interference (the nodes).

While there are many situations which involve counter-propagating waves, this usually is caused by the reflection of waves at boundaries (like the ends of a guitar string). Accordingly, there is nothing keeping the phase constants ϕ_1 and ϕ_2 the same, and so the standing wave doesn't develop. However if the frequency is just right, then it works, as indicated in section 4.2.7.5.

4.2.4 Trigonometric Interference

We are now in a position to look at the fundamental property of waves – namely interference. Our first method of analysis uses trigonometry. Suppose two waves arrive at the same point, and are described by $x_1 = A\cos(\omega t)$ and $x_2 = B\cos(\omega t + \phi)$ respectively. To find out the resulting sum, we add the two disturbances together.

$$\begin{aligned} X &\equiv x_1 + x_2 \\ &= A\cos\omega t + B\cos(\omega t + \phi) \\ &= A\cos\omega t + B\cos\omega t\cos\phi - B\sin\omega t\sin\phi \\ &= (A + B\cos\phi)\cos\omega t - B\sin\phi\sin\omega t \\ &= X_0(\cos\delta\cos\omega t - \sin\delta\sin\omega t) \\ &= X_0\cos(\omega_0 t + \delta) \end{aligned} \tag{9}$$

where we define

$$\begin{aligned} X_0 &= \sqrt{(A + B\cos\phi)^2 + (B\sin\phi)^2} \\ &= \sqrt{A^2 + B^2 + 2AB\cos\phi} \quad . \\ \cos\delta &= \frac{A + B\cos\phi}{X_0} \quad \sin\delta = \frac{B\sin\phi}{X_0} \end{aligned} \tag{10}$$

The amplitude of the resultant is given by X_0. Notice that if A=B, the expression simplifies:

$$\begin{aligned} X_0 &= A\sqrt{2+2\cos\phi} \\ &= A\sqrt{2\left(1+\cos\phi\right)} \\ &= A\sqrt{4\cos^2\left(\tfrac{1}{2}\phi\right)} \\ &= 2A\left|\cos\left(\tfrac{1}{2}\phi\right)\right| \end{aligned}, \qquad (11)$$

and we obtain the familiar result that if the waves are 'in phase' (ϕ=0), the amplitude doubles, and if the waves are π radians (half a cycle) 'out of phase', we have complete destructive interference.

Equation (10) can be used to provide a more general form of this statement – the minimum resultant amplitude possible is $|A-B|$, while the maximum amplitude possible is $A+B$.

This statement is reminiscent of the 'triangle inequality', where the length of one side of a triangle is limited by a similar constraint on the lengths of the other two sides. This brings us to our second method of working out interferences: by a graphical method.

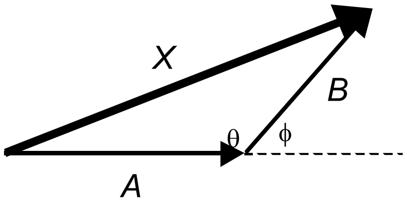
4.2.5 Graphic Interference

In the graphic method a vector represents each wave. The length of the vector gives the amplitude, and the relative orientation of two vectors indicates their phase relationship. If the phase relationship is zero, the two vectors are parallel, and the total length is equal to the sum of the individual lengths. If the two waves are π out of phase, the vectors will be antiparallel, and so will partly (or if $A=B$, completely) cancel each other out.

The diagram below shows the addition of two waves, as in the situation above. Notice that since $\theta + \phi = \pi$, $\cos\theta = -\cos\phi$. One application of the cosine rule gives

$$X_0 = \sqrt{A^2 + B^2 + 2AB\cos\phi} \qquad (12)$$

in agreement with equation (10).



4.2.6 Summary of Interference Principles

The results of the last section allow us to determine the amplitude once we know the phase difference between the two waves. Usually the two waves have come from a common source, but have travelled different distances to reach the point. Let us suppose that the difference in distances is D – this is sometimes called the *path difference*. What will ϕ be?

To find out, we use the wave number *k*. The phase difference ϕ is given by

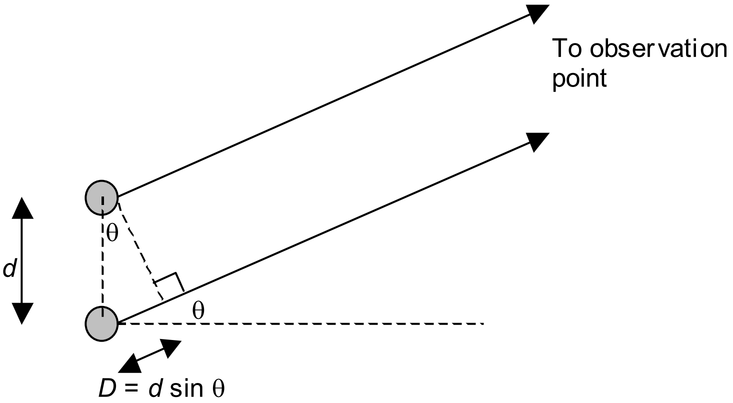
$$\phi = kD = \frac{2\pi D}{\lambda}. \qquad (14)$$

If the distance *D* is equal to a whole wavelength, we expect the two waves to interfere constructively, since peak will meet peak, and trough will meet trough. In equation (15), if D=λ, then ϕ=2π, and constructive interference is indeed obtained, as can be seen from equation (12). Similarly, we find that if D is equal to λ/2, then ϕ=π, and equation (12) gives destructive interference.

4.2.7 Instances of two-wave interference

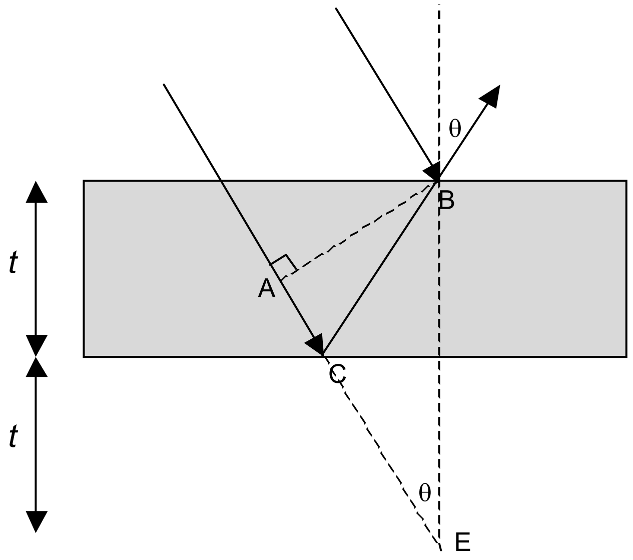
4.2.7.1 Young's "Two Slit Experiment"

Two cases need to be dealt with. The first is known as the two-slit experiment, and concerns two sources in phase, which are a distance *d* apart, as shown in the diagram below. The path difference is given by $D = d\sin\theta$, in the case that *d* is much smaller than the distance from sources to observer. Using the conditions in the last section, we see that interference will be constructive if $D = d\sin\theta = n\lambda$ where *n* is an integer.



4.2.7.2 Thin films and colours on soap bubbles

The second case is known as thin film interference, and concerns the situation in the diagram below. Here the light can take one of two routes.



The path difference is calculated:

$$\begin{aligned} D &= \mathrm{AC} + \mathrm{CB} \\ &= \mathrm{AC} + \mathrm{CE} \\ &= 2t\cos\theta \end{aligned}$$

Before we can work out the conditions required for constructive or destructive interference, there is an extra caution to be borne in mind – the reflections.

4.2.7.3 Hard & Soft Reflections

The reflection of a wave from a surface (or more accurately, the boundary between two materials) can be hard or soft.

- Hard reflections occur when, at the boundary, the wave passes into a 'sterner' material. At these reflections, a peak (before the reflection) becomes a trough (afterwards) and vice-versa. This is usually stated as "a π phase difference is added to the wave by the reflection." These mean the same thing since $\cos(\theta + \pi) = -\cos\theta$.

  To visualize this – imagine that you are holding one end of a rope, and a friend sends a wave down the rope towards you. You keep your hand still. At your hand, the incoming and outgoing waves interfere, but must sum to zero (after all, your hand is not moving, so neither can the end of the rope). Therefore if the incoming wave is above the rope, the outgoing wave must be below. In this way, peak becomes trough and vice-versa.

- Soft reflections, on the other hand, are where the boundary is *from* the 'sterner' material. At these reflections, a peak remains a peak, and there is no phase difference to be added.

What do we mean by 'sterner'? Technically, this is a measurement of the restoring forces in the oscillations which link to produce the wave – the *A* coefficients of (1). However, the following table will help you to get a feel for 'sternness'.

Wave	**From**	**To**	**Reflection**
Light	Reflection off mirror		Hard
Light	Air	Water / Glass	Hard
Light	Water / Glass	Air	Soft
Light	Lower refractive index	Higher refractive index	Hard
Sound	Solid / liquid	Air	Soft
Sound	Air	Solid / liquid	Hard
Wave on string	Reflection off fastened end of string		Hard
Wave on string	Reflection off unsecured end of string		Soft

4.2.7.4 Film Interference Revisited

Going back to our thin film interference: sometimes both reflections will be hard; sometimes one will be hard, and the other soft.

The formulae for constructive interference are:

Both reflections hard, or both soft: $D = 2t\cos\theta = n\lambda$ (16)

One reflection hard, one soft: $D = 2t\cos\theta + \frac{1}{2}\lambda = n\lambda$ (17)

The difference comes about because of the phase change on reflection at a hard boundary.

4.2.7.5 Standing Waves

Equations (16) and (17) with θ=0 can be used to work out the wavelengths allowed for standing waves. For a standing wave, we must have constructive interference between a wave and itself (having bounced once back and forth along the length of the device). The conditions for constructive interference in a pipe, or on a string of length L (round trip total path = 2L) are

$2L = n\lambda$ soft reflections at both ends
$2L = \left(n + \tfrac{1}{2}\right)\lambda$ soft reflection at one end
$2L = (n+1)\lambda$ hard reflections at both ends.

4.2.7.6 Two waves, different frequencies

All the instances given so far have involved two waves of identical frequencies (and hence constant phase difference). What if the frequencies are different? Let us suppose that our two waves are described by $x_1 = A\cos\omega_1 t$ and $x_2 = B\cos\omega_2 t$, where we shall write $\delta \equiv \omega_2 - \omega_1$. When we add them, we get:

$$\begin{aligned} X &= x_1 + x_2 \\ &= A\cos\omega_1 t + B\cos\delta t\,\cos\omega_1 t - B\sin\delta t\,\sin\omega_1 t \\ &= \left(A + B\cos\delta t\right)\cos\omega_1 t - B\sin\delta t\sin\omega_1 t \\ &= X_0\cos(\omega_1 t + \alpha) \end{aligned}, \qquad (18)$$

where

$$\begin{aligned} X_0 &= \sqrt{\left(A + B\cos\delta t\right)^2 + \left(B\sin\delta t\right)^2} \\ &= \sqrt{A^2 + B^2 + 2AB\cos\delta t} \\ \cos\alpha &= \frac{A + B\cos\delta t}{X_0} \end{aligned}. \qquad (19)$$

We see that the effective amplitude fluctuates, with angular frequency δ. On the other hand, if the two original waves had very different frequencies, then this fluctuation may be too quick to be picked up by the detector. In this case, the resultant amplitude is the root of the sum of the squares of the original amplitudes. Put more briefly – if the frequencies are very different, the total intensity is simply given by the sum of the two constituent intensities.

These fluctuations are known as 'beats', and the difference f_2-f_1 is known as the beat frequency. To give an illustration: While tuning a violin, if the tuning is slightly off-key, you will hear the note pulse: loud-soft-loud-soft and so on. As you get closer to the correct note, the pulsing slows down until, when the instrument is in tune, no pulsing is heard at all because f_2-$f_{1=0}$.

4.2.8 Adding more than two waves

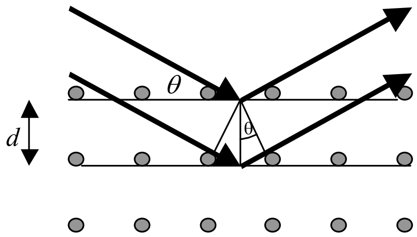
4.2.8.1 Diffraction Grating

The first case we come to with more than two waves is the diffraction grating. This is a plate with many narrow transparent regions. The light can only get through these regions. If the distance between adjacent 'slits' is *d*, we obtain constructive interference, as in section 4.2.4.1, when $d\sin\theta = n\lambda$ - in other words when the light from all slits is in phase.

The difference between this arrangement and the double-slit is that when $d\sin\theta \neq n\lambda$ we find that interference is more or less destructive. Therefore a given colour (or wavelength) only gets sent in particular directions. We can use the device for splitting light into its constituent colours.

4.2.8.2 Bragg Reflection

A variation on the theme of the diffraction grating allows us to measure the size of the atom.



The diagram shows a section of a crystal. Light (in this case, X-rays) is bouncing off the layers of atoms. There are certain special angles for which all the reflections are in phase, and interfere constructively.

Looking at the small triangle in the diagram, we see that the extra path travelled by the wave bouncing off the second layer of atoms is

$$D = 2d\sin\theta\,. \qquad (20)$$

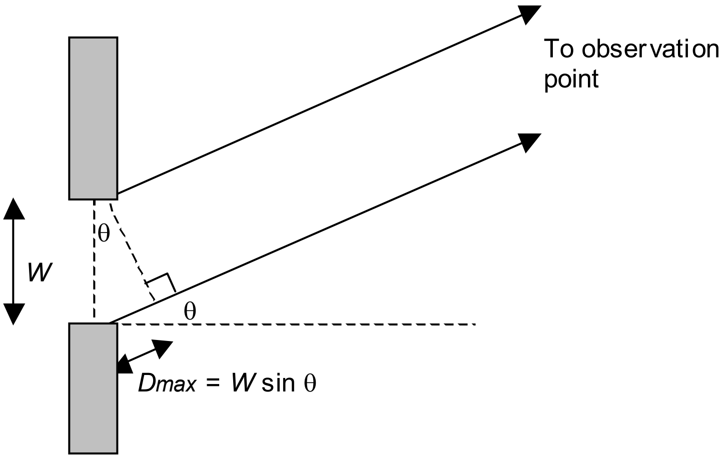
When $D=n\lambda$, we have constructive interference, and a strong reflection. There is one thing that takes great care – notice the definition of q in the diagram. It is not the angle of incidence, nor is it the angle by which the ray is deflected – it is the angle between surface and ray. This is equal to half the angle of deflection, also equal to $\pi/2 - i$.

Using this method, the spacing of atomic layers can be calculated – and this is the best measurement we have for the 'size' of the atom in a crystal.

4.2.8.3 Diffraction

What happens when we add a lot of waves together? There is one case we need to watch out for – when all possible phases are represented with equal strength. In this case, for each wave $\cos(\omega t + \phi)$, there will be an equally strong wave $\cos(\omega t + \phi + \pi) = -\cos(\omega t + \phi)$, which will cancel it out.

How does this happen in practice? Look at the diagram below. Compared with the wave from the top of the gap, the path differences of the waves coming from the other parts of the gap go from zero to $D_{\max} = W\sin\theta$, where *W* is the width of the gap.



If $kD_{\max}$ is a multiple of 2π, then we will have all possible phases represented with equal strength, and overall destructive interference will result.

To summarize, destructive interference is seen for angles θ, where

$$W\sin\theta = n\lambda\,. \qquad (21)$$

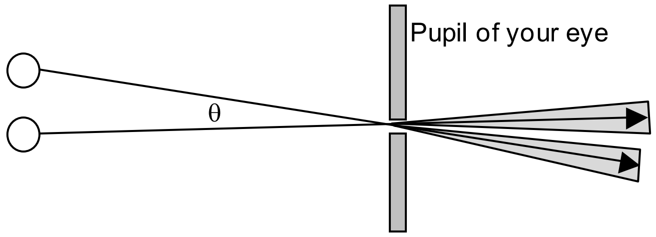
Make sure you remember that *W* is the width of the gap, and that this formula is for **destructive** interference.

This formula is only valid (as in the diagram) when the observer is so far away that the two rays drawn are effectively parallel. Alternatively the formula works perfectly when it is applied to an optical system that is focused correctly, for then the *image* is at infinity.

4.2.8.4 Resolution of two objects

How far away do you have to get from your best friend before they look like Cyclops? No offence – but how far away do you have to be before you can't tell that they've got two eyes rather than one? The results of diffraction can help us work this out. Let's call this critical distance *L*.

The rays from both eyes come into your eyeball. Let us suppose that the angle between these rays is θ, where θ is small, and that your friend's eyes are a distance *s* apart. Therefore $\tan\theta \approx \sin\theta \approx s/L$. These two rays enter your eye, and spread out (diffract) as a result of passing through the gap called your pupil. They can only just be 'resolved' – that is noticed as separate – when the first minimum of one's diffraction pattern lines up with the maximum of the other. Therefore $W\sin\theta = \lambda$ where *W* is the width of your pupil.



Putting the two formulae together gives:

$$\begin{aligned}\sin\theta &= \frac{s}{L} = \frac{\lambda}{W} \\ L &= \frac{sW}{\lambda}\end{aligned} . \qquad (22)$$

For normal light (average wavelength about 500nm), a 5mm pupil, and a 10cm distance between the eyes: you friend looks like Cyclops if you are more than 1km away! If you used a telescope instead, and the telescope had a diameter of 10cm, then your friend's two eyes can be distinguished at a distances up to 20km.

4.2.8.5 The Bandwidth Theorem

In the last section, we asked the question, "What happens when you add lots of waves together?" However we cheated in that we only considered waves of the same frequency. What happens if the waves have different frequencies?

Suppose that we have a large number of waves, with frequencies evenly spread between f and $f+\delta f$. The angular frequencies will be spread from ω to $\omega+\delta\omega$, where $\omega=2\pi f$ as in equation (4). Furthermore, imagine that we set them up so that they all agree in phase at time t=0. They will never agree again, because they all have different frequencies.

The phases of the waves at some later time t will range from ωt to $(\omega+\delta\omega)t$.

Initially we have complete constructive interference. After a short time δt, however, we have destructive interference. This will happen when (as stated in the last section) all phases are equally represented – when the range of phases is a whole multiple of 2π. This happens when $\delta t\times\delta\omega=2\pi$. After this, the signal will stay small, with occasional complete destructive interference.

From this you can reason (if you're imaginative or trusting) that if you need to give a time signal, which has a duration smaller than δt, you must use a collection of frequencies at least $\delta\omega=2\pi/\delta t$. This is called the bandwidth theorem. This can be stated a little differently:

$$\begin{aligned}\delta\omega\,\delta t &= 2\pi\\ \delta(2\pi f)\delta t &= 2\pi\,.\\ \delta f\,\delta t &= 1\end{aligned}\qquad(23)$$

A similar relationship between wavelength and length can be obtained, if we allow the wave to have speed *c*:

$$\begin{aligned}\delta f\,\delta\left(\frac{x}{c}\right) &= 1\\ \delta\left(\frac{f}{c}\right)\delta x &= 1\,.\\ \delta\left(\frac{1}{\lambda}\right)\delta x &= 1\end{aligned}\qquad(24)$$

Expressed in terms of the wave number *k*, this becomes:

$$\begin{aligned}\delta\left(\frac{k}{2\pi}\right)\delta x &= 1\\ \delta k\,\delta x &= 2\pi\end{aligned}\,.\qquad(25)$$

In other words, if you want a wave to have a pulse of length *x* at most, you must have a range of *k* values of at least 2π/*x*.

4.2.8.6 Resolution of spectra

A spectrometer is a device that measures wavelengths. Equation (25) can be used to work out the accuracy (or resolution) of the measurement.

If you want a minimum error δ*k* in the wavenumber, you must have a distance of at least δ*x*=2π/δ*k*. But what does this distance mean? It transpires that this is the maximum path difference between two rays in going through the device – and as such is proportional to the size of the spectrometer. So, the bigger the spectrometer, the better its measurements are.

4.2.9 Doppler Effect

4.2.9.1 Classical Doppler Effect

Suppose a bassoonist is playing a beautiful pure note with frequency *f*. Now imagine that he is practising while driving along a road. A fellow motorist hears the lugubrious sound. What frequency does the listener hear? Let us suppose that the player is moving at velocity *u*, and the listener is moving at velocity *v*. For simplicity we only consider the problem in one dimension, however velocities can still be positive or negative.

Furthermore, imagine that the distance between player and listener is L_0 at time zero, when the first wave-peak is broadcast from the bassoon. We assume that the waves travel at speed *c* with respect to the ground.

This peak is received at time t_1, where

$$\begin{aligned} L_0 + vt_1 &= ct_1 \\ L_0 &= (c-v)t_1 \end{aligned} \tag{26}$$

The first line is constructed like this: The travellers start a distance L_0 apart, so by the time the signal is received, the distance between them is $L_0 + vt_1$. This distance is covered by waves of speed *c* in time t_1 – hence the right hand side.

The next wave peak will be broadcast at time 1/f – one wave cycle later. At this time, the distance between the two musicians will be $L_0 + (v-u)T = L_0 + (v-u)/f$. This second peak will be received at time t_2, where

$$\begin{aligned} L_0 + \frac{v-u}{f} + v\left(t_2 - \frac{1}{f}\right) &= c\left(t_2 - \frac{1}{f}\right). \\ L_0 + \frac{c-u}{f} &= (c-v)t_2 \end{aligned} \tag{27}$$

Finally we can work out the time interval elapsed between our listener hearing the two peaks, and from this the apparent frequency is easy to determine.

$$\begin{aligned} (t_2 - t_1)(c-v) &= L_0 + \frac{c-u}{f} - L_0 \\ \frac{1}{f'} = (t_2 - t_1) &= \frac{c-u}{f(c-v)} \\ f' &= f \times \frac{c-v}{c-u} \end{aligned} \tag{28}$$

From this we see that if *v*=*u*, no change is observed. If the two are approaching, the apparent frequency is high (blue-shifted). If the two are receding, the apparent frequency is low (red-shifted).

4.2.9.2 Relativistic Doppler Effect

Please note that if either *u* or *v* are appreciable fractions of the speed of light, this formula will give errors, and the relativistic calculation must be used.

For light *only*, the relativistic formula is

$$f' = f\sqrt{\frac{c+u}{c-u}}\,, \tag{29}$$

where u is the approach velocity as measured by the observer (u is negative if the source and observer are receding). The relativistic form for other waves is more complicated, and will be left for another day.

4.3 Questions

1. A hole is drilled through the Earth from the U.K. to the centre of the Earth and out of the other side. All the material is sucked out of it, and a 1kg mass is dropped in at the British end. How much time passes before it momentarily comes to rest at the Australian end? (NB You may need some hints from section 1.2.4) +

2. Repeat q1 where a straight hole is drilled between any two places on Earth. Assume that the contact of the mass with the sides of the hole is frictionless. ++

3. In an interferometer, a beam of coherent monochromatic light (with wavelength λ) is split into two parts. Both parts travel for a distance L parallel to each other. One travels in vacuum, the other in air. The beam is then re-combined. If destructive interference results, what can you say about L, λ, and n_{air}?

4. Your wind band is about to play on a pick-up truck going down a motorway at 30m/s. You want people on the bridges overhead to hear you playing 'in tune' (such that treble A is 440Hz) when you are coming directly towards them. What frequency should you tune your instruments to?

5. A police 'speed gun' uses microwaves with a wavelength of about 3cm. The 'gun' consists of a transmitter and receiver, with a small mirror which sends part of the transmission directly into the receiver. Here it interferes with the main beam which has reflected off a vehicle. The received signal strength pulsates (or beats). What will be the frequency of this pulsation if the vehicle is travelling towards you at 30mph? +

6. How far away do you need to hold a ruler from your left eye before you can no longer resolve the millimetre markings? Keep your right eye covered up during this experiment. Use equation (22) to make an estimate for the wavelength of light based on your measurement. Remember that W is the width of your pupil.

7. A signal from a distant galaxy has one third of the frequency you would expect from a stationary galaxy. Calculate the galaxy's recession velocity using equation (29), and comment on your answer. (NB red-shifts this big *are* measured with very distant astronomical objects.)

5 Optics

5.1 Principles

This chapter is concerned with light. Given that light can be treated in most practical situations as a wave, you might think that having covered waves in Chapter 4, there is nothing more to be said. In many cases, however, given that the wavelength of light is so much shorter than the apparatus we use, the effects of diffraction and interference can be ignored, and we can simply assume that light travels in straight lines.

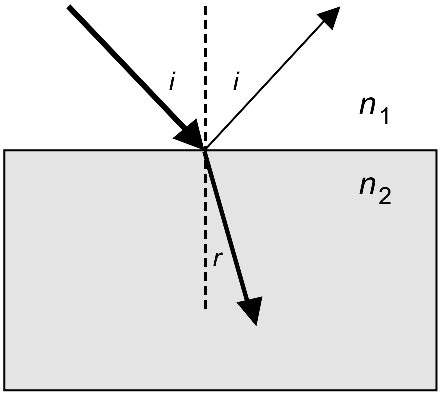
This simplifies our working, which is highly necessary if we wish to study the inner workings of lenses and optical systems. We begin with the fundamentals of reflection and refraction. We then introduce you to Ray Optics which allows the calculation of the route of light through complex systems.

5.1.1 Reflection and Refraction

The diagram below shows both reflection and refraction. We refer to a refractive index of a material, which is defined as

$$\text{Refractive Index }(n) = \frac{\text{Speed of light in vacuum}}{\text{Speed of light in the material}}\,. \qquad (1)$$

Air has a refractive index of about 1.0003[14], glass has a refractive index of about 1.5, and water about 1.3.



First of all, the angle of reflection is equal to the angle of incidence (both were labelled *i* in the diagram).

[14] The refractive index also gives a measure of pressure, since *n*-1 is proportional to pressure.

Secondly, the angle of incidence is related to the angle of refraction *r* by the formula:

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1} \qquad (2)$$

Notice that since the sine of an angle can be no larger than one, if $n_2 < n_1$, then refraction becomes impossible if $i > \sin^{-1}(n_2/n_1)$. This limiting angle is called the critical angle. For greater angles of incidence, the entire wave is reflected, and this is called total internal reflection.

When a wave passes from one material into another, the frequency remains the same (subject to the linearity provisos of section 4.1.1). Given that the speed changes, the wavelength will change too. The wavelength of light in a particular material can be evaluated:

$$\lambda = \frac{c}{f} = \frac{c_0}{nf} = \frac{\lambda_0}{n} \qquad (3)$$

where c_0 (and λ_0) represent speed of light (and wavelength) in vacuum.

5.1.2 Fermat's Principle

Fermat's principle gives us a method of working out the route light will take in an optical system. It states that light will take the route that takes the least time. Given that the time taken in a single material is equal to

$$T = \frac{D}{c} = \frac{nD}{c_0} \qquad (4)$$

where c_0 is the speed of light in vacuum; minimizing the time is the same as minimizing the product of distance and refractive index. This latter quantity (*nD*) is called the optical path. It is possible to prove the laws of reflection and refraction using this principle.[15]

5.2 Ray Optics

5.2.1 The Paraxial Ray approximation and Apparent Depth

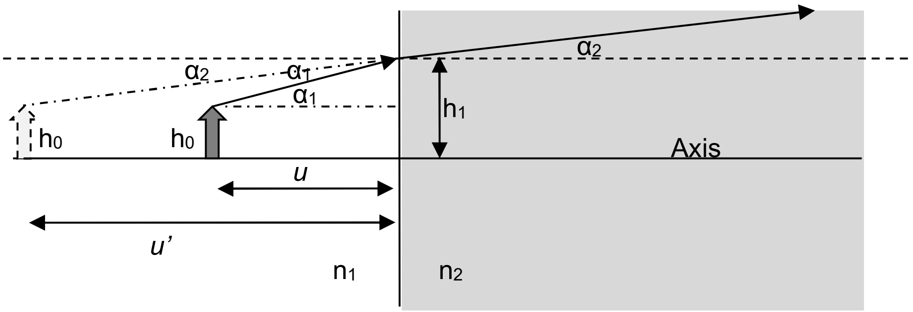
When studying an optical system, we obtain much clearer, linear equations (which fit with the vast majority of practical situations) by

[15] To do this, imagine the plane as a sheet of graph paper, with the boundary along the x-axis. Suppose that the light starts at point (0,Y), and needs to get to (X,-Y). Now assume that the light crosses the x-axis at point (x,0). Work out the total optical path travelled along the route, and then minimize it with respect to x. You should then be able to identify sin *i* and sin *r* in the algebraic soup, and from this, you should be able to finish the proof.

making the assumptions that our rays make small angles to the optical axis. Our diagrams will always show the axis as a horizontal line, with the rays passing from left to right. When tracing a ray through a system, we use two parameters. The displacement h records how far above the axis the ray is at that particular point, while the angle α measure the angle between the ray and the axis (where upward slopes are considered positive).

The use of the small angle approximation enables us to assume that $\sin(\alpha) = \alpha$, where of course we measure all of our angles in radians. Please see section 3.1 if you need further information about radians.

The simplest situation is where a ray strikes a boundary into a material of different refractive index. Here light from an object of height h_0 hits a boundary after travelling distance u in a material with refractive index n_1.



Our equation of refraction tells us that $\sin(\alpha_1)/\sin(\alpha_2) = n_2/n_1$. It follows that $n_1 \sin \alpha_1 = n_2 \sin \alpha_2$, and if we use the paraxial ray approximation, this simplifies to $n_1 \alpha_1 = n_2 \alpha_2$, so $\alpha_2 = n_1 \alpha_1 / n_2$.

If we are interested in the height of the ray as it hits the boundary (h_1), this is equal to $h_0 + u \tan \alpha_1$, which by the small angle approximations means that $h_1 = h_0 + u \alpha_1$.

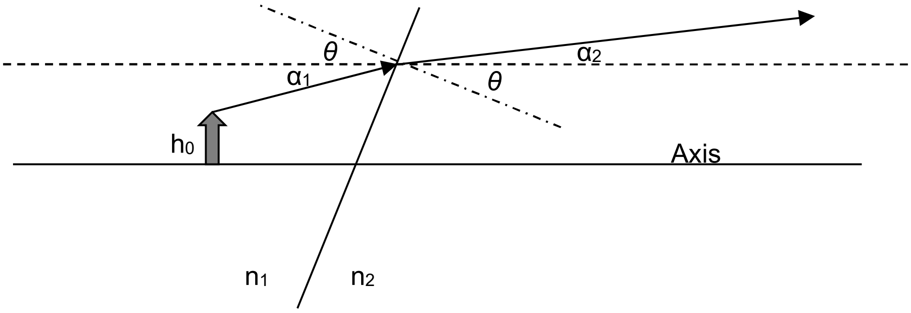
From the perspective of a viewer in the material on the right (of refractive index n_2), the light appears to come from a virtual image a distance u' from the boundary. Given that $h_0 + u' \alpha_2 = h_1$, it follows that $u' \alpha_2 = u \alpha_1$, and thus that $u' n_1 \alpha_1 / n_2 = u \alpha_1$. From this you can see that $u' = u\, n_2/n_1$. The fact that the α_1 cancelled out of the equation is significant. This means that all light coming from the object appears to come from depth u' within the n_1 material – not just the light emitted at the particular angle α_1.

The distance u' is sometimes called the apparent depth, and can therefore be calculated by the formula $u' = u\, n_2/n_1$. To visualize this, imagine a coin in a mug of water, where the water is of depth u. In this situation, $n_1 = n_{water} \approx 1.3$. When viewed from above (from the air where $n_2 = n_{air} = 1$), the coin seems to be at depth $u' = u\, n_2/n_1 \approx u / 1.3 \approx 0.8\, u$.

This is why objects underwater look closer to the surface than they really are.

5.2.2 Oblique Boundaries and Curvature

We next consider the situation when the boundary is not perpendicular to the axis. This is a vital intermediate step before we can deal with lenses.

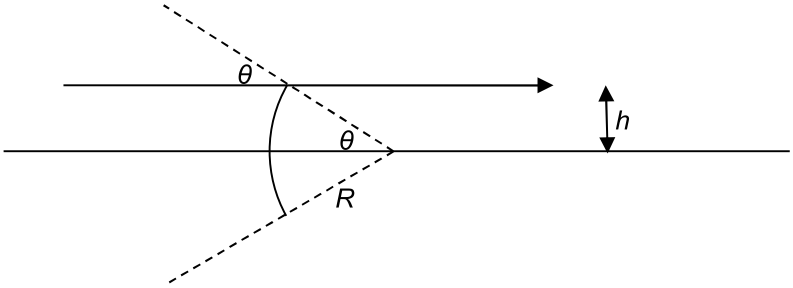


In this situation, the boundary makes an angle θ to the vertical. The incident angle $i = \theta + \alpha_1$, while the angle of refraction is $r = \theta + \alpha_2$.

The Law of Refraction in this case gives us $n_1 \times i = n_2 \times r$, so

$$n_1(\theta + \alpha_1) = n_2(\theta + \alpha_2). \qquad (5)$$

We next consider a curved boundary. We assume that the curve is the arc of a circle of radius *R*, with the centre of the circle on the axis. *R* is regarded as positive if it is to the right of the arc. We can analyse its effect on the light if we calculate the value of θ for each possible height *h* at which a ray of light might strike the boundary.



Here light strikes a curved boundary at height above the axis. This height $h = R \sin \theta$, or in the small angle approximation, $h = R\,\theta$. It follows that $\theta = h/R$.

Substituting this into equation (5) gives us

$$n_1\left(\frac{h}{R}+\alpha_1\right)=n_2\left(\frac{h}{R}+\alpha_2\right)$$
$$n_2\alpha_2=n_1\alpha_1+h\left(\frac{n_1-n_2}{R}\right). \qquad (6)$$

5.2.3 Thin Lenses

We are now in a position to apply our knowledge to a simple practical situation – viewing objects in lenses. Initially we consider the simplest kind. These are 'thin' lenses where we assume that the value of *h* will not change as the light progresses through the lens.

Let us consider light arriving at the lens at a distance h_1 above the axis, with an initial angle to the axis of α_1. The light arrives in air ($n_1 = 1$). It first meets a surface of radius R_1, and goes into the material of the lens, which has refractive index n_2. It then meets a second curved surface of radius R_2, and passes back out into air again ($n_3 = 1$). We take the angles made by the ray to the axis as α_2 in the lens, and α_3 after it has left the lens.

For the first surface, equation (6) becomes

$$n_2\alpha_2=\alpha_1+h_1\left(\frac{1-n_2}{R_1}\right),$$

and for the second surface we have

$$\alpha_3=n_2\alpha_2+h_1\left(\frac{n_2-1}{R_2}\right).$$

Putting the two equations together gives

$$\alpha_3=\alpha_1+h_1(n_2-1)\left(\frac{1}{R_2}-\frac{1}{R_1}\right). \qquad (7)$$

We usually group together the parts which are entirely dependent on the lens and call it the power of the lens:

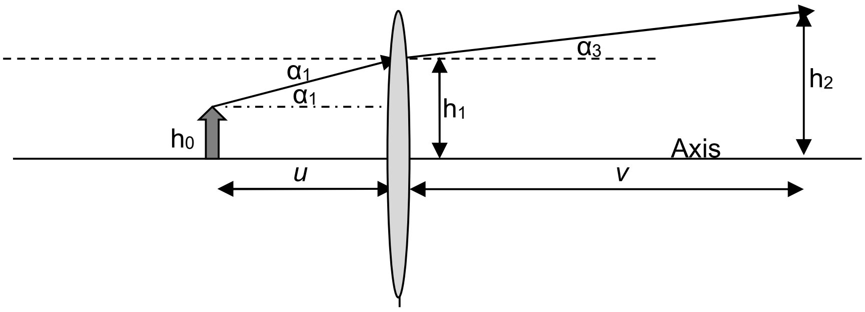
$$P=-(n_2-1)\left(\frac{1}{R_2}-\frac{1}{R_1}\right)=(n_2-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right). \qquad (8)$$

This simplifies equation (7) to

$$\alpha_3=\alpha_1-Ph_1 \qquad (9)$$

where our choice of sign convention means that positive powers of lens direct the ray back toward the axis (reducing the magnitude of α).

Let us know consider what this lens will do to our ray of light from the object of height h_0 a distance u to the left of the lens.



We already know that $h_1 = h_0 + \alpha_1 u$. It is also true that $h_2 = h_1 + \alpha_3 v$. Putting these facts together with equation (9) we have

$$\begin{aligned} h_2 &= h_1 + (\alpha_1 - Ph_1)v \\ &= h_1(1-Pv) + \alpha_1 v \\ &= (h_0 + \alpha_1 u)(1-Pv) + \alpha_1 v \\ &= h_0(1-Pv) + \alpha_1(u + v - Puv) \\ &= h_0(1-Pv) + \alpha_1 uv\left(\frac{1}{v} + \frac{1}{u} - P\right) \end{aligned} \quad . \qquad (10)$$

Something particularly amazing happens if we choose the value of v so that the final term in brackets is zero. In this case, the value of h_2 does not depend upon α_1, but only on h_0. In other words, all of the light from the height h_0 is brought to the same height h_2. Furthermore, h_2 is proportional to h_0, and so an image of the original object has been formed, with magnification $M = 1 - Pv$.

In order for this to be true we require

$$P = \frac{1}{v} + \frac{1}{u} . \qquad (11)$$

Optical folk frequently refer to the focal length of lenses. The focal length is the distance f required to bring a parallel beam of light to one point. To find an expression for the focal length we look once more at equation (10). If we arrange matters so that $1 = Pv$, then the height h_2 is independent of the original height, and only dependent on the angle. This means that if the light was parallel to begin with (all rays had the same value of α_1), it would all be brought to the same place. By definition in this case, $v=f$. Accordingly, $1 = Pf$, and

$$P = \frac{1}{f}. \qquad (12)$$

Lens powers are measured in dioptres (D), so an *f*=5cm lens has a power of *P*=1 / 0.05m = 20D. Using this notation, equation (11) can be rewritten

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u}. \qquad (13)$$

This is sometimes known as the lens equation, as it tells you where to expect images for particular strengths of lenses given the object distances *u*. Equation (8) is known as the lens-maker's formula as it tells you the radii of curvature needed to make a lens with a given power or focal length. You can see from this equation that the larger the refractive index of the material, the larger the radii, and accordingly the thinner the lens can be. Opticians take great pleasure in selling these lenses in spectacles as they are less heavy, many find them more elegant.

The magnification, defined as h_2/h_0 has already been shown to be equal to 1-*Pv*. By similar triangles, it is also equal to -*v/u*.

5.2.3.1 Convex Lenses

A convex lens is one which is thicker in the middle than at the edges. Remembering our convention that curvature of a surface is positive if the centre of the circle is to the right of the curve itself, then we will definitely have a convex lens if R_1 is positive and R_2 is negative, however as long as $R_1>R_2$ then the lens will be thicker in the middle, equation (8) tells us that it will have a positive power and it will have the effect of directing rays towards the axis.

Using equation (13),

$$v = \left(\frac{1}{f} - \frac{1}{u}\right)^{-1} = \left(\frac{u-f}{fu}\right)^{-1} = \frac{fu}{u-f}. \qquad (14)$$

If *u*<*f*, then *v* becomes negative, and a virtual image is formed behind the lens. This is what is happening in a magnifying glass. If *u*>*f*, then a real image will be formed. The magnification is given by

$$M = 1 - Pv = 1 - \frac{v}{f} = 1 - \frac{u}{u-f} = \frac{f}{f-u} = \frac{1}{1-\frac{u}{f}} \qquad (15)$$

Given that for the image to be real (*v*>0), we must have u>f, all real images have negative magnification. This means that upward pointing objects for downward pointing images. The image will be the same size as the object when *M*= –1, which equation (15) tells us will occur when *u*=2*f*. If the object distance is greater than 2*f*, then |*M*|<1 and the image is diminished, whereas if *f*<*u*<2*f*, then |*M*|>1 and the image is magnified.

Note that convex lenses can be used to shorten the distance needed to bring rays to a focus. This is how they are used in spectacles for patients with long sight. Rays which were heading to focus behind the retina rather than on it because the eyeball was too short or the eye's lens was too weak can be focused clearly on the retina instead.

5.2.3.2 Concave Lenses

A concave lens is thinner in the middle that at the edges. This will definitely be the case if R_1 is negative and R_2 is positive, but will be true as long as $R_1<R_2$. Concave lenses accordingly have a negative power, and in equations such as (13) we take the focal length *f* as negative too. This means that when a parallel beam of light strikes the lens, it is spread out to look as if it came from a single point a distance *f* behind the lens. For rays coming from an object (the situation where *u* is positive), the concave lens will always form a virtual image, as can be seen from the fact that if you put a negative value for *f* into equation (14), *v* will be less than zero. Equation (15) also makes it clear that such images will always be diminished as *M* will be less than 1 (but positive) if *u* is positive but *f* is negative.

Concave (negative power) lenses can form real images if the light entering them is sufficiently converging. This situation can be modelled by setting *u* as negative (e.g. rays which would have converged to a point 5cm to the right of the lens would be said to have *u*= – 0.05m). The lens in this case will cause the focus to move to the right.

Therefore concave lenses can be used to lengthen the distance before rays reach a focus. Patients with short sight tend to have an eyeball which is too large or an eye lens which is too strong, and accordingly the sharp image is formed in front of the retina rather than on it. The concave lens moves it back onto the retina.

5.2.3.3 Multiple Lenses acting as One

If two lenses are placed one-behind-the-other, their effects are combined. Let us initially assume that there is no gap. In this case, we assume that the value of *h* will not change between the lenses. This enables us to write two versions of equation (7) for the two lenses.

First lens (light enters at α_1, and leaves at α_3)

$$\alpha_3 = \alpha_1 + h_1(n_2 - 1)\left(\frac{1}{R_2} - \frac{1}{R_1}\right) = \alpha_1 - h_1 P_1$$

Second lens (light enters at α_3, and leaves at α_5). We assume that the refractive index of the material in this lens is n_4, and that the radii of curvature of the surfaces are R_3 and R_4.

$$\alpha_5 = \alpha_3 + h_1(n_4 - 1)\left(\frac{1}{R_4} - \frac{1}{R_3}\right) = \alpha_3 - h_1 P_2$$

So therefore,

$$\alpha_5 = \alpha_3 - h_1 P_2 = \alpha_1 - h_1 P_1 - h_2 P_2 = \alpha_1 - h_1(P_1 + P_2). \qquad (16)$$

In short, the combination acts as a single lens of power $P_1 + P_2$.

There is a complication if there is a gap between the lenses. If the gap width is d, then the distance from the axis at the second lens will be $h_2 = h_1 + \alpha_3 d$. In this case

$$\begin{aligned}\alpha_5 &= \alpha_3 - h_2 P_2 = \alpha_3 - (h_1 + \alpha_3 d)P_2 = \alpha_3(1 - dP_2) - h_1 P_2 \\ &= (\alpha_1 - P_1 h_1)(1 - dP_2) - h_1 P_2 = \alpha_1 - h_1(P_1 + P_2 - dP_1 P_2) - \alpha_1 dP_2\end{aligned}. \qquad (17)$$

This is a complicated expression, but two assumptions simplify it. Firstly, remember that under the small angle approximation $\alpha \ll 1$. Secondly d must be much smaller than the focal lengths of the lenses, or we would not be able to regard this as a single lens – the image could well be in between the two lenses, for example. If this is the case, the final term ($\alpha_1 dP_2$) can be assumed to be significantly smaller than the other terms (it contains two small numbers α_1 and d, and only one big one P_2, whereas other terms have only one small number, or if they have two {e.g. $h_1 dP_1 P_2$} this is counterbalanced by two large ones).

In the cases where these assumptions are valid the power of the two lenses is

$$P = P_1 + P_2 + dP_1 P_2 . \qquad (18)$$

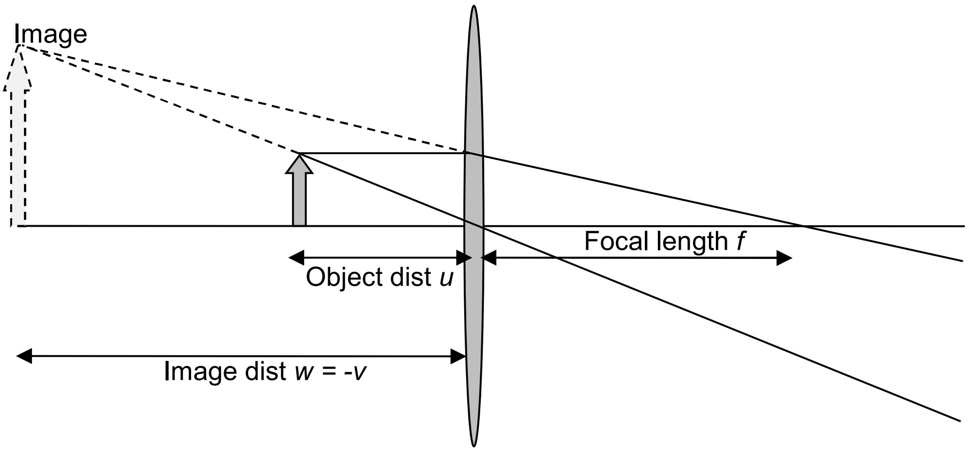
Where the assumptions are not valid, it is unwise to look for a simple solution, and the overall combination can not be treated simply as a lens with a particular power. A method for analysing such systems in terms of matrices is introduced later in this chapter.

5.3 Optical Systems

5.3.1 Magnifying lens

The usual method of enabling finer detail to be seen is to move the object nearer to your eye, so it occupies a larger fraction of your total vision. When you do this, you do not make the object larger, but you do make its angular size increase (when the angle is measured at the eye). Unfortunately, this process has limits, as the human eye can only comfortably focus on objects more than 25cm away. This distance is called the 'distance of distinct vision' d_0. The role of the magnifier is not so much to make the image bigger than a closely held object in angular terms, but to make it appear with the same angular size, but at a distance where the eye can focus on it without strain.

The ray diagram for a magnifying lens is shown below. This is a convex lens being used with an object distance *u* less than its focal length. The image is therefore to the left of the lens, and in our equations (10-15) the image distance *v* is accordingly negative.

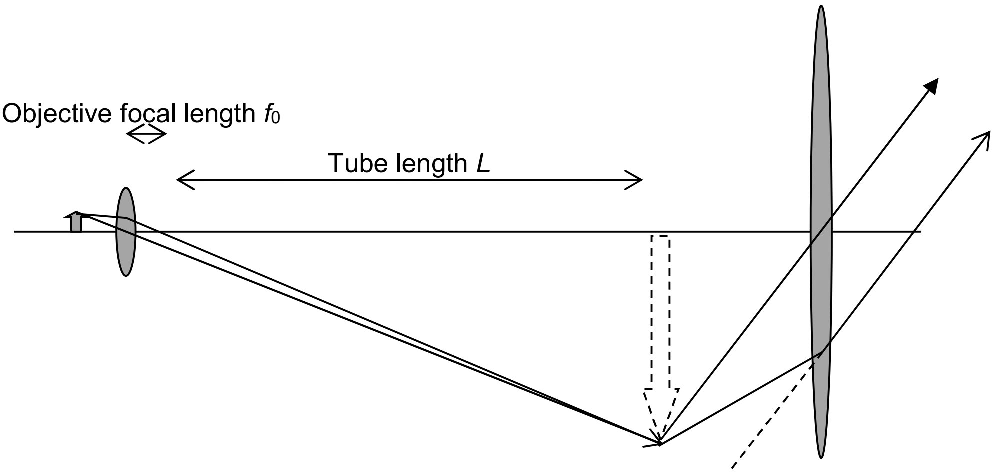


The magnification of this device is defined as the angular size of the image at the lens (which we assume to be very close to your eye) divided by the angular size the image would have if it were placed 25cm from your eye (where you could focus comfortably). We shall simplify the situation and design our lens to put the image exactly 25cm from the eye. In other words, we have set *w* so that it equals d_0. In this case the magnification defined above will be the same as the 'ordinary' magnification (the ratio of the image size to object size).

The reasoning following equation (10) shows that the magnification of this system is $M = 1 - Pv = 1 + Pw$ where *P* is the power of the lens. Therefore if the object had height h_0, the image would have height $(1+Pw)\ h_0$. This image is formed at distance d_0. If the object were actually placed this far away, it would still only have height h_0. Accordingly the magnification is $1+Pw = 1+Pd_0$.

5.3.2 Standard microscope

The job of a microscope is to make small objects look large. A simple microscope will have two lenses – objective and eyepiece. The objective is close to the sample and forms a real, magnified (though inverted) image relatively close to the eyepiece. The eyepiece then acts as a magnifier enabling this image to be seen at the distance of distinct vision.



By similar triangles, the ratio between the first image height and the object height (the magnification M_0) will be equal to the tube length divided by the focal length of the objective lens. The focal length of the objective is equal to $1/P_0$ where P_0 is the power of the objective. Thus $M_0 = L/f_0 = LP_0$. The eyepiece then magnifies this by a factor $(1+P_ed_0)$ where P_e is the power of the eyepiece and d_0 is the distance of distinct vision, as in the section on magnifiers (5.3.1). The overall magnification is therefore $LP_0(1+P_ed_0)$, although we ought to write this as a negative number as the image is inverted.

The action of enlarging the image spreads the light out, which makes it dimmer. It is accordingly vital that the object is very well illuminated by a bright lamp.

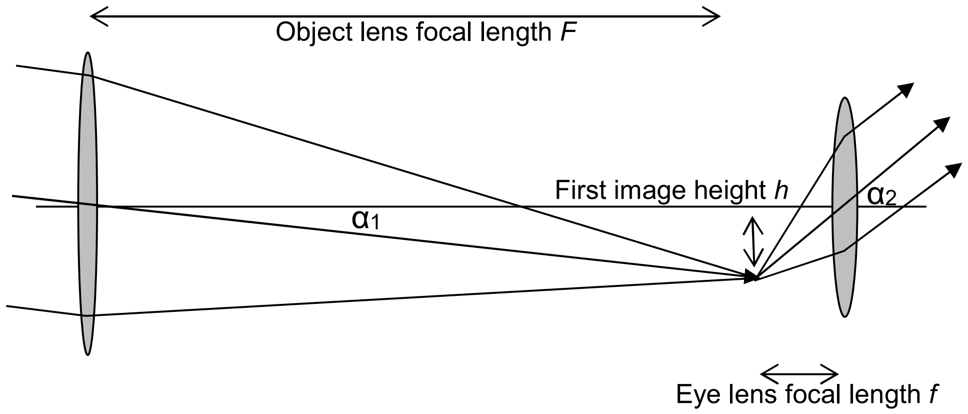
Ray optics are not the only consideration when designing a microscope. Light does also behave as a wave, and diffraction effects need to be kept to a minimum to ensure a clear image. This is achieved by ensuring that the objective lens is significantly wider than the object. Re-arranging equation (23) in chapter 4, the lens must have a diameter of at least $W = \lambda L/s$ where L is the distance of the specimen from the objective and s is the size of the detail you wish to see in that specimen.

5.3.3 Refracting telescope

Unlike the magnifier and microscope, the principal aim of a telescope is not to make things appear larger. If you were to look at the sky with an optical system with a magnification of 10 000, you could take very good pictures of the surface of the Moon, however when looking at the stars, even this magnification would not make them seem any larger (they are a very very long way away), but by magnifying you would make the images dimmer and might lose the ability to navigate the night sky. After all, it is the patterns of the stars which help us find our way about and identify particular stellar objects – and a pattern requires more than one star to be visible at a given moment.

The main object is to make the stars brighter, and this is done by using wide lenses to catch more light than possible with the naked eye. A wide lens has a second benefit – it reduces the angle by which light diffracts on passing into the telescope and enables crisper images of distant objects to be formed than is possible with the naked eye. Clarity is aided by a dose of magnification, but only by a factor of tens not thousands.

A simple refracting telescope contains two lenses. The first (the object lens) makes real images of the distant stars. As the stars are so far away from the telescope, the light from each star is more or less parallel when it arrives at the lens. Accordingly it will be focused to a point one focal length from the lens. An eye lens then enables the astronomer to view this light comfortably.



The magnification here is defined as α_2/α_1. Now $\alpha_1 = h/F$, while $\alpha_2 = h/f$, given that the initial and final beams of light are parallel. It follows that the magnification will be *F/f*.

5.3.4 General method for analysing optical systems

In the section on thin lenses acting as one (5.2.3.3) it was stated that the approximations used there only worked if the distance between the lenses was small, and a promise was made to show a more general system of analysing optical systems which did not suffer this limitation. Now is the time to introduce a general method for analysing an optical system.

We represent the ray at any point in the system with a vector

$$\mathbf{r} = \begin{pmatrix} h \\ \alpha \end{pmatrix}$$

where *h* is the distance of the ray above the axis, and α is the angle (in radians) made by the ray to the axis. When a ray passes some optical component (or even a region of empty space), the component acts on the vector to make a new one. Each component is represented as a matrix C.

5.3.4.1 Light travelling a distance u in a uniform medium

Our simplest component is no component at all – just passing through space. This doesn't change the angle, but the height above the axis does change – we have already used the equation $h_1 = h_0 + \alpha u$. In other words

$$h \rightarrow h' = h + \alpha u$$
$$\alpha \rightarrow \alpha' = \alpha$$

where we use *h'* and *α'* to represent the new values of *h* and α. This can be written in matrix form as

$$\begin{pmatrix} h' \\ \alpha' \end{pmatrix} = \begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix} \begin{pmatrix} h \\ \alpha \end{pmatrix} \quad \text{or} \quad \mathbf{r'} = C_S(u)\mathbf{r}\,.$$

5.3.4.2 Boundaries between media

A surface perpendicular to the axis joining a material of refractive index n_1 to one of n_2 causes changes no change to *h*, but α' = n_1 α / n_2 as proved in section 5.2.1. The matrix for this boundary is therefore

$$\begin{pmatrix} h' \\ \alpha' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & n_1/n_2 \end{pmatrix} \begin{pmatrix} h \\ \alpha \end{pmatrix} \quad \text{or} \quad \mathbf{r'} = C_{PB}(n_1, n_2)\mathbf{r}\,.$$

A curved boundary of radius *R* can also be treated, where we start with equation (6). Again, there is no change to *h*, but the angle is changed by a function which this time depends on *α* and also *h*. Written in matrix form, we have:

$$\begin{pmatrix} h' \\ \alpha' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \dfrac{n_1 - n_2}{Rn_2} & \dfrac{n_1}{n_2} \end{pmatrix} \begin{pmatrix} h \\ \alpha \end{pmatrix} \quad \text{or} \quad \mathbf{r'} = C_{CB}(R, n_1, n_2)\mathbf{r}\,.$$

5.3.4.3 A complete system

We now apply the matrix method to a complete system. Our first system is a thin lens. The light starts in air, goes through a boundary of radius R_1 into a material of refractive index *n*, then goes through a boundary of radius R_2 into air again. So we begin with vector **r**. First to act upon it is the boundary into the material $C_{CB}(R_1,1,n)$, and second to act is the boundary out of the material $C_{CB}(R_2,n,1)$. The whole effect is given by

$$
\begin{aligned}
\mathbf{r'} &= C_{CB}(R_2,n,1)C_{CB}(R_1,1,n)\mathbf{r} \\
&= \begin{pmatrix} 1 & 0 \\ \dfrac{n-1}{R_2} & n \end{pmatrix}\begin{pmatrix} 1 & 0 \\ \dfrac{1-n}{R_1 n} & \dfrac{1}{n} \end{pmatrix}\begin{pmatrix} h \\ \alpha \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ \dfrac{n-1}{R_2}+\dfrac{1-n}{R_1} & 1 \end{pmatrix}\begin{pmatrix} h \\ \alpha \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ (n-1)\left(\dfrac{1}{R_2}-\dfrac{1}{R_1}\right) & 1 \end{pmatrix}\begin{pmatrix} h \\ \alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -P & 1 \end{pmatrix}\begin{pmatrix} h \\ \alpha \end{pmatrix}
\end{aligned}
$$

where *P* is defined in equation (8), and we have agreement with equations (7) and (9).

The effect of a lens of thickness *t* could also be calculated:

$$
\begin{aligned}
\mathbf{r'} &= C_{CB}(R_2,n,1)C_S(t)C_{CB}(R_1,1,n)\mathbf{r} \\
&= \begin{pmatrix} 1 & 0 \\ \dfrac{n-1}{R_2} & n \end{pmatrix}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ \dfrac{1-n}{R_1 n} & \dfrac{1}{n} \end{pmatrix}\begin{pmatrix} h \\ \alpha \end{pmatrix}.
\end{aligned}
$$

By multiplying the matrices for each part of the system, we can form a matrix for any optical system by multiplying the matrices corresponding to each component.

5.4 Questions

1. Show that Fermat's principle allows you to 'derive' the Law of Reflection. Assume that you have a mirror along the *x*-axis. Let light start at point (X_1,Y_1) and end at point (X_2,Y_2). Show that the least-time reflected route is the one which bounces off the mirror where angles of incidence and reflection will be equal. +

2. Show that Fermat's principle allows you to 'derive' Snell's Law. Assume that you have a material with refractive index 1 for *y*>0 (that is, above the *x*-axis), and refractive index *n* where *y*<0. Show that the shortest time route from point (X_1,Y_1) to (X_2,Y_2), where Y_1>0 and Y_2<0, crosses the boundary at the point where sin *i* / sin *r* = *n*. +

3. You are the navigator for a hiking expedition in rough ground. Your company is very thirsty and tired, and your supplies have run out. There is a river running East-West which is 4km South of your current position. Your objective is to reach the base camp (which is 2km South and 6km West of your current position), stopping off at the river on the way. What is the quickest route to the camp via the river? +

4. You are the officer in charge of a food convoy attempting to reach a remote village in a famine-stricken country. On your map, you see that 50km to your East is a straight border (running North-South) between scrub land (over which you can travel at 15km/hr) and marsh (over which you can only travel at 5km/hr). The village is 141km South-East of your current position. What is the fastest route to reach the village? +

5. Prove equation (13) by a graphical method. Draw an axis, a convex lens, and an object. Mark the object distance *u*, and a focal distance *f* to the right of the lens (but make sure that $f<u$). Now draw two rays. The first should start at the object, and pass straight through the centre of the lens and continue without bending (rays passing the centre of the lens pass a section with $\theta=0$, and accordingly are not bent). The second ray should start at the object and travel parallel to the axis until it reaches the lens. At this point it should bend so it passes the focal point on the axis and then keeps going. The image is where the two rays meet. Form two different equations for the ratio (image height)/(object height) using two sets of similar triangles, and eliminate the height ratio to form equation (13).

6. People with ‘normal’ eyes can bring objects into focus on their retina providing the object is further away than 25cm. Assuming that their retina is 2.0cm behind the lens, calculate the range of powers which the eye’s lens can adopt.

7. A patient has eyesight which is perfectly corrected by contact lenses with a power of -1D. Using the information from q6, calculate the range of object distances which can be viewed crisply by the patient without spectacles.

8. Would your answer to q7 be significantly different if the patient preferred wearing spectacles whose lenses sat 2cm in front of the eye?

9. Assume that a convex lens has been set up to form a sharp image of an object, where the object distance *u*, image distance *v* and lens power *P* are known. What will be the effect on the image distance of placing a glass block (of refractive index *n*) of thickness *t* in between the object and the lens? Calculate the distance by which the image moves if originally $u = v = 20$cm, the glass has refractive index 1.50 and the block is 2.0cm thick. +

10. Explain why the matrix for any optical system which produces an image must have a zero in its top right hand corner.

11. Work out the matrix which is equivalent to a glass block of refractive index *n* and thickness *t*. Don’t forget to include the boundaries with the air.

6 Hot Physics

This section gives an introduction to the areas of physics known as thermodynamics and statistical mechanics. These deal with the questions “What happens when things heat up or cool down?” and “Why?” respectively.

We start with a statement that will be very familiar – but then find that it leads us into new territory when explored further.

6.1 The Conservation of Energy

You will be used to the idea that energy can neither be used up nor created – only transferred from one object to another, perhaps in different forms.

For our purposes, this is stated mathematically as

$$dQ + dW = dU \, , \qquad (1)$$

where ‘dX’ refers to ‘a small change in X’. Put into words, this states: “Heat entering object + Work done on the object = the change in its internal energy.” Internal energy means any form of stored energy in the object. Usually this will mean the heat it has, and will be measured by temperature. However if magnetic or electric fields are involved, U can also refer to electrical or magnetic potential energy.

Given that the conservation of energy must be the starting point for a study of heat, it is called the First Law of Thermodynamics.

Equation (1) can be applied to any object or substance. The most straightforward material to think about is a perfect gas, and so we shall start there. It is possible to generalize our observations to other materials afterwards.

Imagine some gas in a cylinder with a piston of cross-sectional area A. The gas will have a volume V, and a pressure p. Let us now do some work on the gas by pushing the piston in by a small distance dL. The force required to push the piston $F = p\,A$, and so the work done on the gas is $dW = F\,dL = p\,A\,dL$. Notice that AdL is also the amount by which the volume of the gas has been decreased. If dV represents the change in volume, $dV = -A\,dL$. Therefore $dW = -p\,dV$.

For a perfect gas in a cylinder (or in fact in any other situation), the First Law can be written a bit differently as:

$$dQ = dU + p\,dV \qquad (2)$$

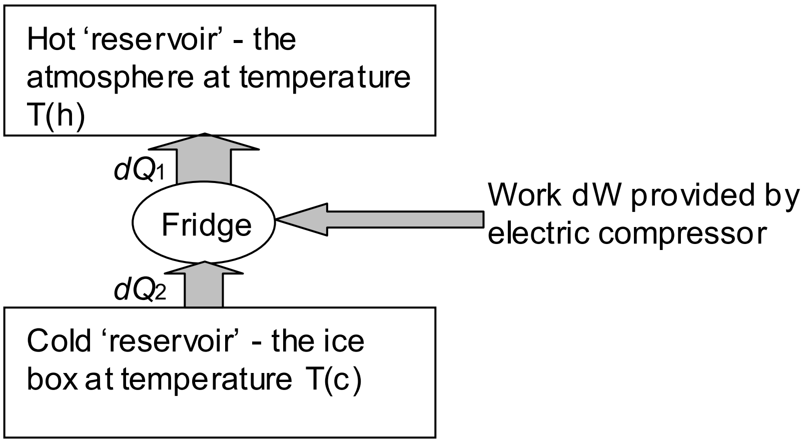
6.2 *The Second Law*

While the First Law is useful, there are certain things it can never tell us. For example – think about an ice cube sitting on a dish in an oven. We know what happens next – the ice cube melts as heat flows from oven to ice, warming it up until it reaches melting point. However the First Law doesn't tell us that. As far as it is concerned it is just as possible for heat to flow from the ice to the oven, cooling the ice and heating the oven.

We stumbled across our next law – called the second law of thermodynamics. This can be stated in several ways, but we shall start with this: **Heat will never flow from a cold object to a hotter object by itself**.

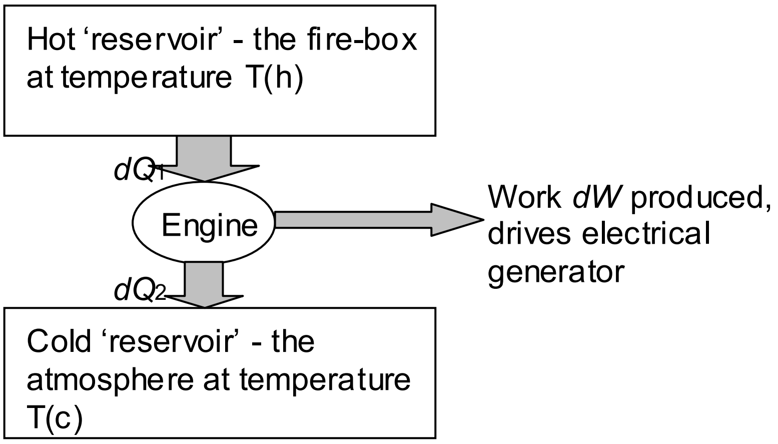
This helps us with the ice in the oven, but you may be wondering what the significance of the "by itself" is. Actually heat *can* be transferred from a cold object to a hotter one – that is what fridges and air conditioning units do. However they can only do it because they are plugged into the electricity supply. If you are prepared to do some work – then you can get heat out of a cold object and into a hotter one, but as soon as you turn the power off and leave it to its own devices, the heat will start flowing the other way again.

6.3 *Heat Engines and Fridges*


Hot 'reservoir' - the atmosphere at temperature T(h)
dQ1
Fridge
Work dW provided by electric compressor
dQ2
Cold 'reservoir' - the ice box at temperature T(c)

The fridge is shown diagrammatically above. It is a device which uses work dW (usually provided by an electric compressor) to extract heat dQ_2 from the ice-box (cooling it down), and pump it out into the surroundings (warming them up). However, by the conservation of energy, the amount of energy pumped out dQ_1 is bigger than the amount of energy removed from the ice-box. By convention $dQ_2>0$, and $dQ_1<0$, since heat flowing in is regarded as positive. The First Law therefore states that $dQ_1 + dQ_2 + dW = 0$.

The fridge is a device that uses work to move heat from cold objects to hot. The opposite of a fridge is a heat engine. This allows heat to flow its preferred way – namely from hot to cold – but arranges it to do some work on the way. Petrol engines, steam engines, turbo-generators and jet engines are all examples of heat engines.


Hot 'reservoir' - the fire-box at temperature T(h)
dQ1
Engine
Work *dW* produced, drives electrical generator
dQ2
Cold 'reservoir' - the atmosphere at temperature T(c)

It was Carnot who realised that the most efficient heat engine of all was a 'reversible' heat engine. In other words – one that got the same amount of work out of the heat transfer as would be needed to operate a perfect fridge to undo its operation.

In order to do this, it is necessary for all the heat transfers (between one object and another) to take place with as small a temperature difference as possible. If this is not done, heat will flow from hot objects to cold – a process which could have been used to do work, but wasn't. Therefore not enough work will be done to enable the fridge to return the heat to the hot object.

Carnot therefore proposed that the ratio of heat coming in from the hot object to the heat going out into the cold object has a maximum for this most-efficient engine. This is because the difference between heat in and heat out is the work done, and we want to do as much work as possible. Furthermore, he said that this ratio must be a function of the temperatures of the hot and cold objects only.

This can be stated as

$$\left|\frac{dQ_1}{dQ_2}\right| = f(T_1, T_2) \tag{3}$$

where T_1 is the temperature of the hot object, and T_2 is that of the cold object. More light can be shed on the problem if we stack two heat engines in series, with the second taking the heat dQ_2 from the first (at temperature *T*2), extracting further work from it before dumping it as heat (dQ3) into a yet colder object at temperature T3.

The two heat engines separately and together give us the equations:

$$\left|\frac{dQ_1}{dQ_2}\right| = f(T_1,T_2) \quad \left|\frac{dQ_2}{dQ_3}\right| = f(T_2,T_3) \quad \left|\frac{dQ_1}{dQ_3}\right| = f(T_1,T_3)$$
$$\Rightarrow \quad f(T_1,T_3) = f(T_1,T_2)\times f(T_2,T_3) \qquad . \qquad (4)$$
$$\Rightarrow \quad f(T_1,T_2) = \frac{g(T_1)}{g(T_2)}$$

6.3.1 Thermodynamic Temperature

However if g(T) is a function of the temperature alone, we might as well call g(T) the temperature itself. This is the thermodynamic definition of temperature.

To summarize: Thermodynamic temperature (T) is defined so that in a reversible heat engine, the ratio of heat extracted from the hot object (*Q*1) to the heat ejected into the cold object (*Q*2):

$$\left|\frac{Q_1}{Q_2}\right| = \frac{T_1}{T_2} \qquad (5)$$

The ‘kelvin’ temperature scale obtained using the gas laws satisfies this definition. For this reason, the kelvin is frequently referred to as the unit of ‘thermodynamic temperature’.

6.3.2 Efficiency of a Heat Engine

The efficiency of a reversible heat engine can then be calculated. We define the efficiency (η) to be the ratio of the work done (the useful output) to Q1 (the total energy input). Therefore

$$\eta = \left|\frac{dW}{dQ_1}\right| = \frac{dQ_1 - |dQ_2|}{dQ_1} = 1 - \frac{T_2}{T_1}. \qquad (6)$$

This, being the efficiency of a reversible engine, is the maximum efficiency that can be achieved. A real engine will fall short of this goal. Notice that for a coal-fired power station, in which T1 (the temperature of the boiler) is frequently 840K, and T2 (the temperature of the stream outside) is 300K; the maximum possible efficiency is

$$\eta = 1 - \frac{300}{840} = 64\%.$$

In practice the water leaves the turbo generator at 530K, and so the efficiency can't go any higher than

$$\eta = 1 - \frac{530}{840} = 37\% .$$

The design of modern large power stations is such that the actual efficiency is remarkably close to this value.

6.4 Entropy

Now we need to take a step backwards before we can go forwards. Look back at the definition of thermodynamic temperature in equation (5). It can be rearranged to state

$$\frac{dQ_1}{T_1} = \frac{|dQ_2|}{T_2} \quad \Rightarrow \quad \frac{dQ_1}{T_1} + \frac{dQ_2}{T_2} = 0 \quad \text{REVERSIBLE.} \qquad (7)$$

Remember that this is for the ideal situation of a reversible process – as in a perfect fridge or heat engine. Suppose, then, that we start with some gas at pressure p and volume V. Then we do something with it (squeeze it, heat it, let it expand, or anything reversible), and finally do some more things to it to bring it back to pressure p and volume V. The list of processes can be broken up into tiny stages, each of which saw some heat (dQ) entering or leaving the system, which was at a particular temperature T. The only difference between this situation, and that in (7) is that there were only two stages in the process for the simpler case. The physics of (7) should still apply, no matter how many processes are involved. Therefore providing all the actions are reversible we can write

$$\sum_{\text{complete cycle}} \frac{dQ}{T} = 0 \quad \Rightarrow \quad \oint \frac{dQ}{T} = 0 \quad \text{REVERSIBLE} \qquad (8)$$

where the circle on the integral implies that the final position (on a p,V graph) is the same as where the gas started.

Now suppose that there are two points on the (p,V) graph which are of interest to us, and we call them A and B. Let us go from A to B and then back again (using a different route), but only using reversible processes. We call the first route I, and the second route II. Equation (8) tells us

$$\oint \frac{dQ}{T} = \int_A^B \frac{dQ_I}{T} + \int_B^A \frac{dQ_{II}}{T} = 0$$

$$\int_A^B \frac{dQ_I}{T} - \int_A^B \frac{dQ_{II}}{T} = 0 \qquad \text{REVERSIBLE} \qquad (9)$$

$$\int_A^B \frac{dQ_I}{T} = \int_A^B \frac{dQ_{II}}{T}$$

In other words the integral of dQ/T between the two points A and B is the same, no matter which reversible route is chosen. This is a very special property of a function – we label dQ/T as a function of state, and call it the entropy (S).

This means that the current entropy of the gas, like pressure, volume and temperature, is only a function of the state that the gas is in now – and does not depend on the preparation method.

6.5 *Irreversible Processes and the Second Law*

We must stress that entropy is only given by $\int dQ/T$ when the integral is taken along reversible processes in which there is no wastage of heat. Heat is wasted when it is allowed to flow from a hot object to a cold one without doing any work on the journey. This would be irreversible, since you could only get the heat back into the hot object if you expended more energy on it.

Let us make an analogy. Reversible processes are like a world in which purchasing prices and selling prices are the same. If you started with £100, and spent it in various ways, you could sell the goods and end up with £100 cash at the end.

Irreversible processes are like the real world in that a trader will want to sell you an apple for more than she bought it for. Otherwise she won't be able to make a profit. If you started with £100, and spent it, you would never be able to get the £100 back again, since you would lose money in each transaction. You may end up with £100 'worth of goods', but you would have to be satisfied with a price lower than £100 if you wanted to sell it all for cash.

Let us now return to the physics, and the gas in the piston. What does irreversibility mean here? We haven't lost any energy – the First Law has ensured that. But we have lost usefulness.

Equation (8) tells us that if we come back to where we started, and only use reversible processes on the way, the total entropy change will be zero. There is another way of looking at this, from the point of view of a heat engine.

Let us suppose that the temperature of the boiler in a steam engine is T_A. In a perfect heat engine, the cylinder will receive the steam at this temperature. Suppose Q joules of heat are transferred from boiler to cylinder. The boiler loses entropy Q/T_A, the cylinder gains entropy Q/T_A, and the total entropy remains constant.

Now let us look at a real engine. The boiler must be hotter than the cylinder, or heat would not flow from boiler to cylinder! Suppose that the boiler is still at T_A, but the cylinder is at T_C. We have now let irreversibility loose in the system, since the heat Q now flows from hot to cooler without doing work on the way.

What about the entropy? The boiler now loses Q/T_A to the connecting pipe[16], but the cylinder gains Q/T_C from it. Since $T_C<T_A$, the cylinder gains more entropy than the boiler lost.

This is an alternative definition of the Second Law. Processes go in the direction to maximize the total amount of entropy.

6.6 Re-statement of First Law

For reversible processes, dW = -p dV, and dQ = T dS. Therefore the First Law (1), can be written as

$$T\,dS = dU + p\,dV\ . \tag{9}$$

We find that this equation is also true for irreversible processes. This is because T, S, U, p and V are all functions of state, and therefore if the equation is true for reversible processes, it is true for all processes. However care must be taken when using it for irreversible processes, since TdS is no longer equal to the heat flow, and pdV is no longer equal to the work done.

6.7 The Boltzmann Law

The Boltzmann Law is simple to state, but profound in its implications.

$$\text{Probability that a particle has energy E} \propto e^{-E/kT} \tag{10}$$

[16] What's the pipe got to do with it? Remember that we said that change in entropy dS is only given by dQ/T for reversible processes. The passing of Q joules of heat into the pipe is done reversibly (at temperature T_A), so we can calculate the entropy change. Similarly the passing of Q joules of heat from pipe to cylinder is done reversibly (at T_C), so the calculation is similarly valid at the other end. However something is going on in the pipe which is not reversible – namely Q joules of heat passing from higher to lower temperature. Therefore we mustn't apply any dS = dQ/T arguments inside the pipe.

where *k* is the Boltzmann constant, and is about 1.38×10^{-23} J/K. We also find that the probability that a system has energy E *or greater* is also proportional to $e^{-E/kT}$ (with a different constant of proportionality).

There is common sense here, because (10) is saying that greater energies are less likely; and also that the higher the temperature, the more likely you are to have higher energies.

Let's give some examples:

6.7.1 Atmospheric Pressure

The pressure in the atmosphere at height *h* is proportional to the probability that a molecule will be at that height, and is therefore proportional to $e^{-mgh/kT}$. Here, the energy *E*, is of course the gravitational potential energy of the molecule – which has mass *m*.

The proof of this statement is in several parts. Firstly we assume that all the air is at the same temperature. This is a dodgy assumption, but we shall make do with it. Next we divide the atmosphere into slabs (each of height dh and unit area), stacked one on top of the other. Each slab has to support all the ones above it. From the Gas Law (pV = NkT where N is the number of molecules under consideration), and the definition of density (Nm=ρV) we can show that ρ=pm/kT. Furthermore, if you go up by a small height dh, the pressure will reduce by the weight of one slab – namely ρgdh. Therefore

$$\begin{aligned} \frac{dp}{dh} &= -\rho g = -\frac{pmg}{kT} \\ \Rightarrow \quad p &= p_0 \exp\left(-\frac{mgh}{kT}\right) \end{aligned} \tag{11}$$

where p_0 is the pressure at ground level. We see that the Boltzmann Law is obeyed.

6.7.2 Velocity distribution of molecules in a gas

The probability that a molecule in the air will have x-component of its velocity equal to u_x is proportional to $\exp(-mu_x^2/2kT)$. Here the energy E is the kinetic energy associated with the x-component of motion.

From this statement, you can work out the mean value of u_x^2, and find it to be:

$$\begin{aligned} \overline{u_x^2} &= \frac{\int x^2 \exp\left(-mx^2/2kT\right)dx}{\int \exp\left(-mx^2/2kT\right)dx} = \frac{\frac{1}{2}(2kT/m)^{3/2}}{(2kT/m)^{1/2}} = \frac{kT}{m} \\ \overline{\tfrac{1}{2}mu_x^2} &= \tfrac{1}{2}kT \end{aligned} \tag{12}$$

The mean kinetic energy is given by

$$\overline{K} = \tfrac{1}{2} m\overline{u^2} = \tfrac{1}{2} m\overline{\left(u_x^2 + u_y^2 + u_z^2\right)} = \tfrac{3}{2} kT \qquad (13)$$

so the internal energy of a mole of gas (due to linear motion) is

$$U = N_A \overline{K} = \tfrac{3}{2} N_A kT = \tfrac{3}{2} RT \qquad (14)$$

where $R \equiv N_A k$ is the gas constant. From this it follows that the molar heat capacity of a perfect gas[17], $C_V = \tfrac{3}{2} R$.

6.7.3 Vapour Pressure

The probability that a water molecule in a mug of tea has enough (or more than enough) energy to leave the liquid is proportional to exp(-*E*L/*kT*) where *E*L is the energy required to escape the attractive pull of the other molecules (latent heat of vaporization per molecule).

6.7.4 Justification of Boltzmann Law

In this section, we introduce some statistical mechanics to give a taste of where the Boltzmann law comes from.

Suppose that you have *N* atoms, and *P* 'packets' of energy to distribute between them. How will they be shared? In statistical mechanics we assume that the energy will be shared in the most likely way.

In a simple example, we could try sharing 4 units of energy (*P*=4) among 7 atoms (*N*=7). Because the individual energy units are indistinguishable (as are the 7 atoms), the possible arrangements are:

- 1 atom with 4 energy units, 6 atoms with none
- 1 atom with 3 energy units, 1 with 1, 5 with none
- 2 atoms with 2 energy units, 5 with none
- 1 atom with 2 energy units, 2 with 1, 4 with none
- 4 atoms with 1 energy unit, 3 with none.

These are said to be the five macrostates of the system. We can work out how likely each one is to occur if the energy is distributed randomly by counting the ways in which each macrostate could have happened.

[17] This is the heat capacity due to linear motion. For a monatomic gas (like helium), this is the whole story. For other gases, the molecules can rotate or vibrate about their bonds as well, and therefore the heat capacity will be higher.

For example, in the first case (all of the energy is given to one of the atoms), there are seven ways of setting it up – because there are seven atoms to choose from.

In the second case, we have to choose one atom to take 3 units (7 to choose from), and then choose one from the remaining six to take the remaining unit. Therefore there are 7×6 = 42 ways of setting it up.

Similarly we can count the ways of rearranging for the other macrostates[18]:

• 1 atom with 4 energy units, 6 atoms with none	7 ways
• 1 atom with 3 energy units, 1 with 1, 5 with none	42 ways
• 2 atoms with 2 energy units, 5 with none	21 ways
• 1 atom with 2 energy units, 2 with 1, 4 with none	105 ways
• 4 atoms with 1 energy unit, 3 with none.	35 ways

We can see that there is one macrostate clearly in the lead – where 4 atoms have no energy, 2 atoms have one energy unit, and 1 atom has two energy units. This macrostate is interesting because there is a geometric progression in the number of atoms (4,2,1) having each amount of energy.

It can be shown that the most likely macrostate will always be the one with (or closest to) a geometric progression of populations.[19] In other

[18] The calculation is made more straightforward using the formula $W = N!/(n_0!\,n_1!\,n_2!\ldots)$ where W is the number of ways of setting up the macrostate, n_0 is the number of atoms with no energy, n_1 is the number of atoms with one unit of energy, and so on.

Here is one justification for this formula: *N*! gives the total number of ways of choosing the atoms in order. The division by n_0! prevents us overcounting when we choose the same atoms to have zero energy in a different order. A similar reason holds for the other terms on the denominator.

Alternatively, *W* = the number of ways of choosing n_0 atoms from *N* × the number of ways of choosing the n_1 from the remaining ($N–n_0$) × the number of ways of choosing the n_2 from the remaining ($N–n_0–n_1$) and so on. Thus

$$W = C_{n0}^{N}\, C_{n1}^{N-n0}\, C_{n2}^{N-n0-n1} \ldots$$
$$= \frac{N!}{n_0!\,(N-n_0)!} \times \frac{(N-n_0)!}{n_1!\,(N-n_0-n_1)!} \times \frac{(N-n_0-n_1)!}{n_2!\,(N-n_0-n_1-n_2)!} \ldots = \frac{N!}{n_0!\,n_1!\,n_2!\ldots}$$

[19] To show this, start with a geometric progression (in other words, say $n_i = Af^i$ where f is some number), and write out the formula for W. Now suppose that one of the atoms with *i* units of energy gives a unit to one of the atoms with *j* units. This means that n_i and n_j have gone down by 1, while n_{i-1} and n_{j+1} have each gone up by one. By comparing the old and

words, assuming that this macrostate is the true one (which is the best bet)[20], then the fraction of atoms with *n* energy units is proportional to some number to the power −*n*. The actual fraction is given by the formula $(1+P/N)^{-1} \times (1+N/P)^{-n}$. Assuming that the mean number of energy units per atom is large (so that we approximate the continuous range of values that the physical energy can take), this means that $1+N/P \approx e^{N/P}$, and so the probability that an atom will have *n* units of energy is proportional to $e^{-Nn/P}$.

Suppose that each packet contains ε joules of energy. Then the energy of one atom (with *n* packets) is $E = n\varepsilon$, and the probability that our atom will have energy E as $e^{-NE/P\varepsilon}$. The mean energy per atom is $P\varepsilon/N$. Now we have seen that the average energy of an atom in a system is about *kT*, where *T* is the temperature in kelvins, and so it should not seem odd that the Boltzmann probability is $e^{-E/kT}$ where we replace one expression for the mean energy per atom $P\varepsilon/N$, with another *kT*.

6.8 Perfect Gases

All substances have an *equation of state*. This tells you the relationship between volume, pressure and temperature for the substance. Most equations of state are nasty, however the one for an ideal, or perfect, gas is straightforward to use. It is called the Gas Law. This states that

$$p\,V = n\,R\,T \qquad (18)$$

$$p\,V = N\,k\,T \qquad (19)$$

where p is the pressure of the gas, V its volume, and T its absolute (or thermodynamic) temperature. This temperature is measured in kelvins *always*. There are two ways of stating the equation: as in (18), where n represents the number of moles of gas; or as in (19), where N represents the number of molecules of gas. Clearly $N=N_An$ where N_A is the Avogadro number, and therefore $R=N_Ak$.

You can adjust the equation to give you a value for the number density of molecules. This means the number of molecules per cubic metre, and is given by *N/V = p/kT*. The volume of one mole of molecules can also be worked out by setting *n*=1 in (18):

new values of *W*, you can show that the new *W* is smaller, and that therefore the old arrangement was the one with the biggest *W*.

[20] The bet gets better as the number of atoms increases. The combination (4,2,1) was the most popular in our example of N=7, P=4, however if you repeated the exercise with N=700 and P=400, you would find a result near (400,200,100) almost a certainty. In physics we deal with *huge* numbers of atoms in matter, so the gambling pays off.

$$V_m = \frac{RT}{p} \quad . \tag{20}$$

You can adjust this equation to give you an expression for the density. If the mass of one molecule is m, and the mass of a mole of molecules (the R.M.M.) is M, we have

$$\begin{aligned}\rho &= \frac{Mass}{Volume} = \frac{M}{RT/p} = \frac{Mp}{RT} \\ &= \frac{N_A mp}{N_A kT} = \frac{mp}{kT}\end{aligned} \quad . \tag{21}$$

Please note that this is the ideal gas law. Real gases will not always follow it. This is especially true at high pressures and low temperatures where the molecules themselves take up a good fraction of the space. However at room temperature and atmospheric pressure, the Gas Law is a very good model.

6.8.1 Heat Capacity of a Perfect Gas

We have already shown (in section 5.7.2) that for a perfect gas, the internal energy due to linear motion is $\frac{3}{2}RT$ per mole. If this were the only consideration, then the molar heat capacity would be $\frac{3}{2}R$. However there are two complications

6.8.1.1 The conditions of heating

In thermodynamics, you will see molar heat capacities written with subscripts – C_P and C_V. They both refer to the energy required to heat a mole of the substance (M kilograms) by one kelvin. However the energy needed is different depending on whether the volume or the pressure is kept constant as the heating progresses.

When you heat a gas at constant volume, all the heat going in goes into the internal energy of the gas ($dQ_V = dU$).

When you heat a gas at constant pressure, two things happen. The temperature goes up, but it also expands. In expanding, it does work on its surroundings. Therefore the heat put in is increasing both the internal energy and is also doing work ($dQ_P = dU + p\ dV$).

Given that we know the equation of state for the gas (18), we can work out the relationship between the constant-pressure and constant-volume heat capacities. In these equations we shall be considering one mole of gas.

$$\begin{aligned} C_V &= \frac{dQ_V}{dT} = \frac{dU}{dT} \\ C_P &= \frac{dQ_P}{dT} = \frac{dU}{dT} + p\frac{dV}{dT} \\ &= C_V + p\frac{d}{dT}\left(\frac{RT}{p}\right) = C_V + R \end{aligned} \tag{22}$$

6.8.1.2 The type of molecule

Gas molecules come in many shapes and sizes. Some only have one atom (like helium and argon), and these are called monatomic gases. Some gases are diatomic (like hydrogen, nitrogen, oxygen, and chlorine), and some have more than two atoms per molecule (like methane).

The monatomic molecule only has one use for energy – going places fast. Therefore its internal energy is given simply by $\frac{3}{2}kT$, and so the molar internal energy is $U = \frac{3}{2}RT$. Therefore, using equation (22), we can show that $C_V = \frac{3}{2}R$ and $C_P = C_V + R = \frac{5}{2}R$.

A diatomic molecule has other options open to it. The atoms can rotate about the molecular centre (and have a choice of two axes of rotation). They can also wiggle back and forth – stretching the molecular bond like a rubber band. At room temperature we find that the vibration does not have enough energy to kick in, so only the rotation and translation (the linear motion) affect the internal energy.

Each possible axis of rotation adds $\frac{1}{2}kT$ to the molecular energy, and so we find that for most diatomic molecules, $C_V = \frac{5}{2}R$ and $C_P = \frac{7}{2}R$.

6.8.1.3 Thermodynamic Gamma

It turns out that the ratio of C_P / C_V crops up frequently in equations, and is given the letter γ. This is not to be confused with the γ in relativity, which is completely different.

Using the results of our last section, we see that γ=5/3 for a monatomic gas, and γ=7/5 for one that is diatomic.

6.8.2 Pumping Heat

If a healthy examiner expects you to know about ideal gases and thermodynamics – you can bet that he or she will want you to be able to *do* thermodynamics with an ideal gas. In this section we show you how to turn a perfect gas (in a cylinder) into a reversible heat engine, and in doing so we will introduce the techniques you need to know.

6.8.2.1 *Isothermal Gas Processes*

As an introduction, we need to know how to perform two processes. Firstly we need to be able to get heat energy into or out of a gas without changing its temperature. Remember that we want a reversible heat engine, and therefore the gas must be at the same temperature as the hot object when the heat is passing into it. Any process, like this, which takes place at a constant temperature is said to be *isothermal*.

The Gas Law tells us (18) that pV=nRT, and hence that pV is a function of temperature alone (for a fixed amount of gas). Hence in an isothermal process

$$pV = \text{const} \ . \quad \text{ISOTHERMAL} \qquad (23)$$

Using this equation, we can work out how much we need to compress the gas to remove a certain quantity of heat from it. Alternatively, we can work out how much we need to let the gas expand in order for it to 'absorb' a certain quantity of heat. These processes are known as isothermal compression, and isothermal expansion, respectively.

Suppose that the volume is changed from V_1 to V_2, the temperature remaining T. Let us work out the amount of heat absorbed by the gas. First of all, remember that as the temperature is constant, the internal energy will be constant, and therefore the First Law may be stated dQ=pdV. In other words, the total heat entering the gas may be calculated by integrating pdV from V_1 to V_2:

$$Q = \int p dV = \int \frac{nRT}{V} dV = \left[nRT \ln V \right]_{V1}^{V2} = nRT \ln \frac{V_2}{V_1} \ . \qquad (24)$$

This equation describes an isothermal (constant temperature) process only. In order to keep the temperature constant, we maintain a good thermal contact between the cylinder of gas and the hot object (the boiler wall, for example) while the expansion is going on.

6.8.2.2 *Adiabatic Gas Processes*

The other type of process you need to know about is the adiabatic process. These are processes in which there is no heat flow (dQ=0), and they are used in our heat engine to change the temperature of the gas in between its contact with the hot object and the cold object. Sometimes this is referred to as an isentropic process, since if dQ=0 for a reversible process, TdS=0, and so dS=0 and the entropy remains unchanged.[21]

[21] While the terms 'isentropic' and 'adiabatic' are synonymous for a perfect gas, care must be taken when dealing with irreversible processes in more advanced systems. In this context dQ is not equal to TdS. If dQ=0, the process is said to be adiabatic: if dS=0, the process is

Before we can work out how much expansion causes a certain temperature change, we need to find a formula which describes how pressure and volume are related in an adiabatic process. Firstly, the First Law tells us that if dQ=0, then 0 = dU + p dV. We can therefore reason like this for *n* moles of gas:

$$\begin{aligned} 0 &= dU + pdV \\ &= nC_V dT + pdV \end{aligned}$$

Now for a perfect gas, nRT=pV, therefore $nRdT = pdV + Vdp$. So we may continue the derivation thus:

$$\begin{aligned} nC_V dT &= \frac{C_V}{R}\left(pdV + Vdp\right) \\ 0 &= \frac{C_V}{R}\left(pdV + Vdp\right) + pdV \\ &= C_V\,(pdV + Vdp) + RpdV \\ &= C_p\,pdV + C_V Vdp \\ &= \gamma\,pdV + Vdp \\ &= \gamma\frac{dV}{V} + \frac{dp}{p} \end{aligned} \quad . \qquad (25)$$

Integrating this differential equation gives

$$\begin{aligned} &\gamma \ln V + \ln p + C = 0 \\ &pV^{\gamma} = e^{-C} \\ &pV^{\gamma} = \text{const} \end{aligned} \quad . \text{ ADIABATIC} \qquad (26)$$

Equation (26) is our most important equation for adiabatic gas processes, in that it tells us how pressure and volume will be related during a change.

We now come back to our original question: what volume change is needed to obtain a certain temperature change? Let us suppose we have a fixed amount of gas (*n* moles), whose volume changes from V_1 to V_2. At the same time, the temperature changes from T_1 to T_2. We may combine equation (26) with the Gas Law to obtain:

isentropic. Clearly for a complex system, the two conditions will be different. This arises because in these systems, the internal energy is not just a function of temperature, but also of volume or pressure.

$$\begin{aligned} pV^{\gamma} &= \text{const} \\ pV\,V^{\gamma-1} &= \text{const} \\ nRTV^{\gamma-1} &= \text{const} \\ \frac{T_1}{T_2}\frac{V_1^{\gamma-1}}{V_2^{\gamma-1}} &= 1 \end{aligned} \quad . \qquad (27)$$

6.8.2.3 A Gas Heat Engine

We may now put our isothermal and adiabatic processes together to make a heat engine. The engine operates on a cycle:

1. The cylinder is attached to the hot object (temperature T_{hot}), and isothermal expansion is allowed (from V_1 to V_2) so that heat Q_1 is absorbed into the gas.

2. The cylinder is detached from the hot object, and an adiabatic expansion (from V_2 to V_3) is allowed to lower the temperature to that of the cold object (T_{cold}).

3. The cylinder is then attached to the cold object. Heat Q_2 is then expelled from the cylinder by an isothermal compression from V_3 to V_4.

4. Finally, the cylinder is detached from the cold object. An adiabatic compression brings the volume back to V_1, and the temperature back to T_{hot}.

Applying equation (24) to the isothermal processes gives us

$$\begin{aligned} Q_1 &= nRT_{hot}\ln\frac{V_2}{V_1} \\ Q_2 &= nRT_{cold}\ln\frac{V_4}{V_3} \end{aligned} \quad . \qquad (28)$$

Similarly, applying equation (27) to the adiabatic processes gives us

$$\begin{aligned} \frac{V_3}{V_2} &= \left(\frac{T_{hot}}{T_{cold}}\right)^{\gamma-1} \\ \frac{V_4}{V_1} &= \left(\frac{T_{hot}}{T_{cold}}\right)^{\gamma-1} \\ \Rightarrow\quad \frac{V_3}{V_2} &= \frac{V_4}{V_1} \quad \Rightarrow \frac{V_3}{V_4} = \frac{V_2}{V_1} \end{aligned} \quad . \qquad (29)$$

Combining equations (28) and (29), gives us

$$\frac{Q_1}{Q_2} = -\frac{T_{hot}}{T_{cold}}, \quad \left|\frac{Q_1}{Q_2}\right| = \frac{T_{hot}}{T_{cold}} \tag{30}$$

where the minus sign reminds us that $Q_2<0$, since this heat was leaving the gas.

To summarize this process, we have used a perfect gas to move heat from a hot object to a colder one. In doing this, we notice less heat was deposited in the cold object than absorbed from the hot one. Where has it gone? It materialized as useful work when the cylinder was allowed to expand. Had the piston been connected to a flywheel and generator, we would have seen this in a more concrete way.

We also notice that we have proved that the kelvin scale of temperature, as defined by the Gas Law, is a true thermodynamic temperature since equation (30) is identical to (5).

6.9 *Radiation of Heat*

And finally... there is an extra formula that you will need to be aware of. The amount of heat radiated from an object is given by:

$$P = \varepsilon\sigma A T^4 \tag{31}$$

P is the power radiated (in watts), *A* is the surface area of the object (in m^2), and *T* is the thermodynamic temperature (in K).

The constant σ is called the Stefan-Boltzmann constant, and takes the value of 5.671×10^{-8} W/(m^2K^4).

The amount radiated will also depend on the type of surface. For a perfect matt black (best absorber and radiator), the object would be called a *black body*, and the emissivity ε would take the value 1. For a perfect reflector, there is no absorption, and no radiation either, so ε=0.

6.10 Questions

1. Calculate the maximum efficiency possible in a coal fired power station, if the steam is heated to 700°C and the river outside is at 7°C.

2. Mechanical engineers have been keen to build jet engines which run at higher temperatures. This makes it very difficult and expensive to make the parts, given that the materials must be strong, even when they are almost at their melting point. Why are they making life hard for themselves?

3. Two insulated blocks of steel are identical except that one is at 0°C, while the other is at 100°C. They are brought into thermal contact. A long time later, they are both at the same temperature. Calculate the final temperature; the energy change and entropy change of each block if (a) heat flows by conduction from one block to the other, and if (b) heat flows from one to the other via a reversible heat engine. ++

4. There is a 'rule of thumb' in chemistry that when you raise the temperature by 10°C, the rate of reaction roughly doubles. Use Boltzmann's Law to show that this means the activation energy of chemical processes must be of order 10^{-19}J. +

5. The amount of energy taken to turn 1kg of liquid water at 100°C into 1kg of steam at the same temperature is 2.26 MJ. This is called the latent heat of vaporization of water. How much energy does each molecule need to 'free itself' from the liquid?

6. By definition, the boiling point of a liquid is the temperature at which the saturated vapour pressure is equal to atmospheric pressure (about 100kPa). Up a mountain, you find that you can't make good tea, because the water is boiling at 85°C. What is the pressure? You will need your answer to q4. +

7. Estimate the altitude of the mountaineer in q5. Assume that all of the air in the atmosphere is at 0°C. +

8. Use the Gas Law to work out the volume of one mole of gas at room temperature and pressure (25°C, 100kPa).

9. What fraction of the volume of the air in a room is taken up with the molecules themselves? Make an estimate, assuming that the molecules are about 10^{-10}m in radius.

10. Estimate a typical speed for a nitrogen molecule in nitrogen at room temperature and pressure. On average, how far do you expect it to travel before it hits another molecule? Again, assume that the radius of the molecule is about 10^{-10}m. ++

11. The fraction of molecules (mass *m* each) in a gas at temperature *T* which have a particular velocity (of speed *u*) is proportional to $e^{-mu^2/2kT}$, as predicted by the Boltzmann law. However the fraction of molecules which have speed *u* is proportional to $u^2 e^{-mu^2/2kT}$. Where does the u^2 come from? ++

12. One litre of gas is suddenly squeezed to one hundredth of its volume. Assuming that the squeezing was done adiabatically, calculate the work done on the gas, and the temperature rise of the gas. Why is the adiabatic assumption a good one for rapid processes such as this?

13. A water rocket is made using a 2 litre plastic drinks bottle. An amount of water is put into the bottle, and the stopper is put on. Air is pumped into the bottle through a hole in the stopper. When the pressure gets to a certain level, the stopper blows out, and the pressure of the air in the bottle expels the water. If the bottle was standing stopper-end downwards, it flies up into the air. If you neglect the mass of the bottle itself, what is the optimum amount of water to put in the bottle if you want your rocket to (a) deliver the maximum impulse, or (b) rise to the greatest height when fired vertically. ++

7 Sparks & Generation

7.1 *Electrostatics – when things are still*

The fundamental fact of electrostatics will be familiar to you – opposite charges attract: like charges repel. As a physicist it is not enough to know this, we also want to know how big the force is. It turns out that the equations describing the force, energy, potential and so on are very similar mathematically to the equations describing gravitational attraction.

Using the symbols F for force, U for potential energy, V for potential and E for field, we have the equations:

$$\begin{aligned} F_r &= \frac{1}{4\pi\varepsilon_0}\frac{Qq}{R^2} \\ E_r &= \frac{1}{4\pi\varepsilon_0}\frac{Q}{R^2} \\ U &= \frac{1}{4\pi\varepsilon_0}\frac{Qq}{R} \\ V &= \frac{1}{4\pi\varepsilon_0}\frac{Q}{R} \end{aligned} \qquad (1)$$

Notice that the symbols are slightly different for gravity – field is now given E rather than g, and in consequence we have to use another letter for energy – hence our choice of U. Here we have put a charge Q at the origin, and we measure the quantities associated with a small 'test' charge q at distance R. Notice that we do not have a minus sign in front – this allows like charges to repel rather than attract (whereas in gravity, positive mass attracts positive mass [and we have never found any lumps of negative mass – if we did this would upset a lot of our thinking]). Also, in place of the G of gravity, we have the constant $1/4\pi\varepsilon_0$ which is about 10^{20} times bigger. No wonder the theoreticians talk of gravity as a weaker force!

Now, you may wonder, why the factor of 4π? This comes about, because the equations above are not the nicest way of describing electrostatic forces. They are based on the 'Coulomb Force Law', which is the first equation in (1) – however there is another, equivalent, way of describing the same physics, and it is called Gauss' Law of Electrostatics. This Gauss Law is on the Olympiad syllabus, and you will find it useful because it will simplify your electrical calculations a lot.

7.1.1 Gauss Law of Electrostatics

Firstly, let us say what the law is. Then we will describe it in words, and then prove that it is equivalent to the Coulomb Law. Finally, we will show its usefulness in other calculations.

$$\oiint_S \mathbf{E} \bullet \mathbf{dS} = \frac{Q}{\varepsilon_0} \tag{2}$$

What does this mean? Firstly let us look at the individual symbols. S is a closed surface (that is what the circle on the integral sign means) – like the outer surface of an apple, a table, or a doughnut – but not the outer surface of a bowl (which ends at a rim). **dS** is a small part of the surface, with area dS, and is a vector pointing outward, perpendicular to the surface at that point. Q is the total electric charge contained inside the surface S. Finally, the vector **E** is the electric field (in volts per metre) – where the vector points in the direction a positive charge would be pulled.

The odd looking integral tells us to integrate the dot product of **E** with **dS** (a normal vector to the surface) around the complete surface. This may sound very foreign, strange, and difficult, but let us give some examples.

First of all, suppose S is a spherical surface of radius R, with one charge (+Q) at the centre of the surface. Assuming there are no other charges nearby, the field lines will be straight, and will stream out radially from the centre. Therefore **E** will be parallel to **dS**, and the dot product will simply be E dS – the product of the magnitudes. Now the size of E must be the same all round the surface, by symmetry. Therefore

$$\oiint_S \mathbf{E} \bullet \mathbf{dS} = \oiint_S E dS = E \oiint_S dS = E \times 4\pi R^2 \tag{3}$$

since the surface area of a sphere is given by 4πR^2. Now by Gauss' Law, this must equal Q/ε_0. Putting the two equations together gives us

$$4\pi R^2 E = \frac{Q}{\varepsilon_0} \quad \Rightarrow \quad E = \frac{Q}{4\pi\varepsilon_0 R^2} \tag{4}$$

in agreement with Coulomb's Law.

Similarly we may find the field at a distance R from a **wire** that carries a charge λ per metre – spread evenly along the wire – something which Coulomb's Law *could* do, but would need a horrendous integral to do it.

This time, our surface is a cylinder, one metre long, with the wire running down its axis. The cylinder has radius R. First, notice that the field lines will run radially out from the wire. This has the consequence that the two flat ends of the cylinder *do not count* in the integral. Think about this for a moment, because this is important. The vector **dS** for these ends will point normal to the surface – that means parallel to the wire. The vector **E**, on the other hand points outward. Therefore **E** is perpendicular to **dS**, and the dot product is zero.

Only the curved surface counts. Again, **E** will have the same magnitude at all points on it because of symmetry, and **E** will be parallel to **dS** on this surface. Once again, we have

$$\oiint_S \mathbf{E} \bullet \mathbf{dS} = \oiint_{curved\,S} E\,dS = E \oiint_{curved\,S} dS = E \times 2\pi RL \qquad (5)$$

where L is the length of the cylinder, and hence 2πRL is its curved surface area. By Gauss' Law, this must be equal to $Q/\varepsilon_0 = \lambda L/\varepsilon_0$, and so

$$E = \frac{\lambda}{2\pi\varepsilon_0 R}. \qquad (6)$$

We notice that for a line charge, the "inverse square" of the Coulomb Law has become an "inverse, not squared" law.

Before moving on, let us make two more points. Firstly, we have not *proved* the equivalence of Gauss' Law and Coulomb's Law – we have only shown that they agree in the case of calculating the field around a fixed, point charge. However the two can be proved equivalent – but the proof is a bit involved, and is best left to first (or second) year university courses.

Secondly, let us think about a surface S which is entirely inside the same piece of metal. **E** will be zero within a metal, because any non-zero E (i.e. voltage difference) would cause a current to flow until the E were zero. Therefore a surface entirely within metal can contain no total charge!

Impossible, I hear you cry! Let us take a hollow metal sphere, with a charge +Q at the centre of the cavity. How can the enclosed charge be zero – surely it's +Q! Oh, no it isn't. Actually the *total* enclosed charge is zero, and this enclosed charge is made up of +Q at the middle of the cavity, and –Q induced on the inside wall of the hollow sphere! If the sphere is electrically isolated, and began life uncharged, there must be a +Q charge somewhere on the metal, and it sits on the outer surface of the sphere.

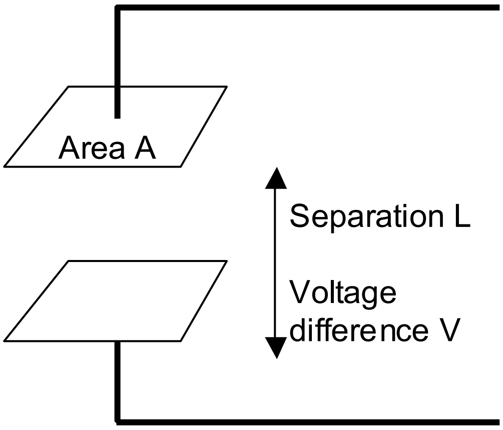
7.1.2 Capacitors

A capacitor is a device that stores charge. To be more precise, a capacitor consists of two conducting plates, with insulating space between them. When the positive plate carries charge +Q, an equal amount of negative charge is stored on the other. If certain insulating materials are used to separate the plates (instead of air or vacuum), the amount of charge that can be stored increases considerably. The charge stored is proportional to the potential difference across it, and we call the constant of proportionality the capacitance.

Gauss' Law gives us a wonderful way to calculate the capacitance of simple capacitors, and we will look at the calculation for a parallel plate capacitor.

7.1.2.1 Parallel Plate Capacitor

At its simplest, a capacitor is shown in the figure below. The two plates are square, and parallel. Each has area A and the distance between them is denoted L. Let's work out the capacitance. To do this, we first suppose that there were a charge Q stored. In other words, there is a charge –Q on the top plate, and +Q on the bottom plate. We can work out the electric field in the gap using Gauss' Law. We draw a rectangular box-shape surface, with one of its faces parallel to, but buried in the bottom plate, and the opposite face in the middle of the gap.



When working out the surface integral $\oiint_S \mathbf{E} \bullet \mathbf{dS}$, only this face in the middle of the gap counts. The face buried in the metal of the plate has **E=0**, while the other four surfaces' normals are perpendicular to the field. Therefore

$$\frac{+Q}{\varepsilon_0} = \oiint_S \mathbf{E} \bullet \mathbf{dS} = AE\,. \tag{7}$$

We next work out the voltage. This is not hard, as by analogy from gravitational work (chapter 1, equation 16)

Field = - d(Potential)/d(distance)

$$E = -dV/dx \qquad (8)$$

Here, E is constant and uniform throughout the inter-plate gap, and so V=Ex+c where c is a constant of integration. Thus the voltage difference between the plates can be calculated; and from this the capacitance can be worked out.

$$\Delta V = EL = \frac{QL}{\varepsilon_0 A} \quad \Rightarrow \quad C = \frac{Q}{\Delta V} = \frac{\varepsilon_0 A}{L} \qquad (9)$$

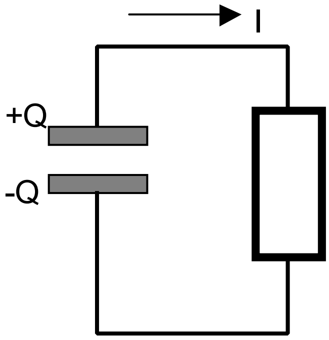
We ought to give a word of caution at this point. In a real parallel plate capacitor, the field near the edge of the plates will not point directly from one plate to the other, but will 'bow out' a bit. Therefore the equation given above is only true when these 'edge effects' are ignored. It turns out that the equation is pretty good providing that *L* is much smaller than both of the linear dimensions of the plates.

The equation also allows you to see the effects of wiring capacitors in series or parallel. When two identical capacitors, each of capacitance C, are connected together in parallel, the overall area A is doubled, so the capacitance of the whole arrangement is 2C. On the other hand if the capacitors are connected in series, the result is one capacitor with twice the gap thickness L. Therefore the overall capacitance is C/2.

7.1.2.2 Decay of Charge on Capacitor

We next come to the case where a capacitor is charged to voltage V (that is, a voltage V across the plates), and then connected in a simple circuit with a resistor R. How long will it take to discharge?

To work this out, we need to use our characteristic equations for capacitor and resistor. For the capacitor V=Q/C, for the resistor V=IR. To solve the circuit we need to clarify the relationship between Q and I.



Here we need to take care. Depending on how the circuit has been drawn, either I=dQ/dt or I=-dQ/dt. That is why it is essential that you include in your circuit diagrams arrows to show the direction of current flow for I>0, and which plate of the capacitor is the +ve one. Here a positive I will discharge the capacitor, so I=–dQ/dt.

The voltage across the capacitor is the same as that across the resistor, so

$$\begin{gathered} V = IR = -R\frac{dQ}{dt} = \frac{Q}{C} \\ \frac{dQ}{dt} = -\frac{Q}{RC} \end{gathered} . \qquad (10)$$

This differential equation can be solved with an exponential solution:

$$Q(t) = Q_0 e^{-t/RC} \qquad (11)$$

where Q_0 was the initial charge on the capacitor after it was charged up. Given that the voltage across the capacitor V(t)=Q(t)/C, the time dependence of the voltage obeys a similar equation. The constant RC is known as the time constant, and it is the time taken for the voltage (or charge) to fall by a factor e (approx 2.7).

7.1.2.3 *Energy considerations*

Next, we need to know how much energy has been stored in a capacitor, if its voltage is V and its capacitance C. The energy is actually 'stored' in the electric field between the plates – but more of this later.

To work the energy out, we charge a capacitor up from scratch (initial charge = 0), and continually measure the current flowing, and the voltage across it. The energy stored must be given by

$$\begin{aligned} U &= \int P\,dt = \int VI\,dt = \int V\frac{dQ}{dt}dt = \int V\frac{d(CV)}{dt}dt \\ &= \int VC\frac{dV}{dt}dt = \int CVdV = C\int VdV = C\left[\tfrac{1}{2}V^2\right] = \left[\tfrac{1}{2}CV^2\right] \end{aligned} \qquad (12)$$

Given that the energy stored must be zero when V=0, and there is no electric field in the gap, this fixes the constant of integration as zero, and we obtain the fact that energy stored = half the capacitance × the square of the voltage across the gap. You could equally well say that the energy is given by half the charge multiplied by the voltage.

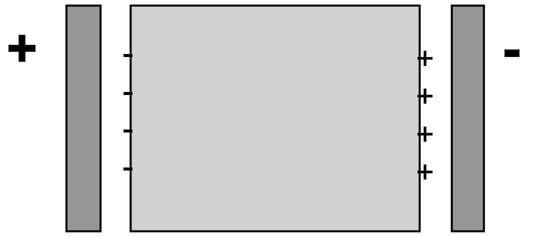
Before we leave this formula, let us do some conjuring tricks with this energy, assuming that the capacitor is a simple parallel plate device:

$$U = \tfrac{1}{2}CV^2 = \frac{\varepsilon_0 AV^2}{2L} = \frac{\varepsilon_0 A(EL)^2}{2L} = \tfrac{1}{2}\varepsilon_0 E^2 \times AL \qquad (13)$$

thus the energy stored per unit volume of gap is $\frac{1}{2}\varepsilon_0 E^2$. Although we have only shown this to be true for a perfect parallel plate capacitor, it is possible to make any electric field look like rows and columns of parallel plate capacitors arranged like a mosaic, and from this the proof can be generalized to all electric fields.

7.1.2.4 Polarization

When we introduced capacitors, we mentioned that the capacitance can be raised by inserting insulating stuff into the gap. How does this work? Look at the diagram below. The stuff in the middle contains atoms, which have positive and negative charges within them.



When the plates are charged, as shown, this pulls the nuclei of the stuff to the right, and the electrons of the atoms to the left. The left plate now has a blanket of negative charge, and the right plate has a blanket of positive charge. This reduces the overall total charge on the plates, and therefore reduces the voltage across the capacitor. Of course the circuit can't remove the "polarization" charges in the 'blankets' – as that would require the chemical breakdown of the substance. So we have stored the same 'circuit charge' on the capacitor for less voltage, and so the capacitance has gone up. The ratio by which the capacitance increases is called the relative permittivity of the substance (it used to be called the dielectric constant), and is given the symbol ε_r. In the presence of such a material, the ε_0 of all the equations derived so far in this chapter needs to be multiplied by ε_r.

7.2 *Magnetism – when things move*

So far, we have just considered electric charges at rest. Our next job is surely to look at electric charges that have gone roaming, and then to study magnetism – two things still to do? No. Actually we only have one job, because magnetism *is* all about moving charges.

We ought to give one warning, though. Just because magnetism is about moving charges, we can't derive its formulae simply from Coulomb's Law and Classical mechanics. We need Relativistic mechanics! That is actually one route into relativity – it is the kind of mechanics needed if electricity and magnetism are to be described together. Put another way, your nearest piece of evidence for special relativity is not in a particle accelerator or airborne atomic clock, but in your wrist-watch (if it has hands), credit card, vacuum cleaner, fridge, CD player, printer, hard disk drive, or wherever your nearest magnet is.

We shall demonstrate this at the end of the chapter. However, for the moment, let us stick to what we need for the Olympiad, and let us carry on calling it "magnetism" as opposed to "relativistic electricity".

7.2.1 Magnetic Flux Density

If there is a magnetic field, there must be a measurement of the field strength, and we call this the flux density, and give it the symbol **B**. The field has a direction (from North to South), and is therefore a vector. The fundamental fact of magnetism can be stated in two ways:

1. If a wire of length L is carrying current **I**, and the wire is in a magnetic field **B**, it will experience a force **F**, where $\mathbf{F} = L\,\mathbf{I} \times \mathbf{B}$. Written without the vector cross product, this is $F = B\ I\ L \sin\theta$, where θ is the angle between the direction of the current, and the direction of the magnetic field.

2. If a charged particle, of charge Q is moving in a magnetic field **B**, and it has velocity **u**, it will experience a force **F**, where $\mathbf{F} = Q\ \mathbf{u} \times \mathbf{B}$. Written without the vector cross product, this is $F = Q\ u\ B \sin\theta$ where θ is the angle between the direction of motion, and the direction of the magnetic field.

You can show that these descriptions are equivalent, by imagining the wire containing N charges (each Q coulombs) per metre. If the wire has length L, the total charge is Q_{TOT}=NQL, and this moves when the current is flowing. If the current is I, this means that the charge passing a point in one second is I, and hence I=NQu, where u is the speed. Therefore

$$\begin{aligned} \mathbf{F} &= L\,\mathbf{I} \times \mathbf{B} \\ &= L\,(NQ\mathbf{u}) \times \mathbf{B} \\ &= (NQL)\ \mathbf{u} \times \mathbf{B} \\ &= Q_{TOT}\ \mathbf{u} \times \mathbf{B}. \end{aligned} \tag{14}$$

7.2.2 Doing the Corkscrew

Now that we have an expression for the force on a charge moving in a magnetic field, we can work out the motion if a charge is thrown into the vicinity of a magnet.

The most important fact is that the force is always at right angles to the velocity. Therefore it never does any work at all, and it never changes the kinetic energy (hence speed) of the object.

The next important fact is that if the velocity is parallel to the magnetic field, there is no force – and the particle will just carry on going: as if the magnet weren't there at all.

This second fact is useful, because any velocity can be broken down (or resolved) into two components – one parallel to the magnetic field (u cosθ), and one perpendicular to it (u sinθ). The component parallel to the field will be unchanged by the motion – it will stay the same, just as Newton's First Law requires.

We next need to calculate what the effect of the other component will be. This will cause a force perpendicular to both the velocity and magnetic field, and we already know from classical mechanics that when a force consistently remains at right angles to the motion, a circular path is obtained. We can calculate the radius from the equations of circular motion:

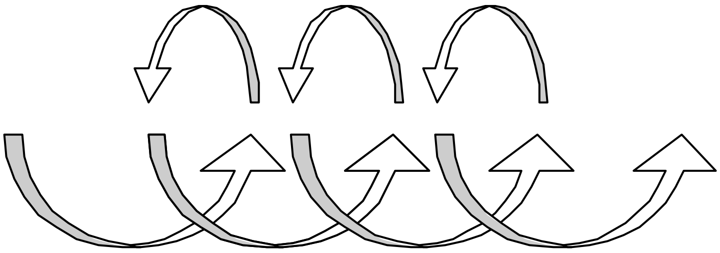
Magnetic Force = Mass × Centripetal Acceleration

$$BQu\sin\theta = \frac{m(u\sin\theta)^2}{R} \quad \Rightarrow \quad R = \frac{mu\sin\theta}{BQ} \qquad (15)$$

Another useful measurement is the time taken for the particle to 'go round the loop' once. Since its rotational speed is u sinθ, the time taken to go round a 2πR circumference is $T = 2\pi R/u\sin\theta$. We can also work out the angular velocity:

$$\omega = \frac{u\sin\theta}{R} = \frac{BQ}{m} \qquad (16)$$

The overall motion is therefore a helix, or corkscrew shape, with the axis of the corkscrew parallel to the magnetic field. The radius of the helix is given by R in equation (15), and the 'pitch' (that is, the distance between successive revolutions), is equal to $D = T\,u$ cosθ.



Helix, or corkscrew motion of an electron in a magnetic field.

Please notice that all these formulae remain valid when the particle starts going very quickly. The only correction that special relativity requires is that we use the enhanced mass $m = \gamma m_{rest}$. No further correction is needed, because the speed remains constant, and therefore γ does not change.

7.2.3 Calculation of magnetic field strengths

So far, we have just thought about the effects of a magnetic field. However before you can study what an electron will do in such a field, you have to make the field! How do you do that?

At its simplest, magnetic fields are caused by electric currents. The bigger the current: the bigger the field. These may be 'real' currents of electrons in wires, or they can be the effective currents of electrons 'orbiting' their nuclei in atoms. The latter is responsible for the permanent magnetic property of iron (and some other metals) – however the process is quite involved and needs no further consideration for the Olympiad.

We do need to worry about the magnetism caused by wires, however, although only a brief description is necessary. The Olympiad syllabus precludes questions that involve large amounts of calculus (hence integration), and most field calculations require an integral. Therefore if you need to know how big a magnetic field is, you will probably be given the equation you need.

Nevertheless it will help to see how the calculations are done. There is a method akin to Coulomb's Law, and an alternative called Ampere's Law. We shall introduce both.

7.2.3.1 Magnetic Coulomb – The Biot Savart Law

To use this method, the wire (carrying current **I**) is broken down into small lengths (dL), linked head-to-tail. To work out the size of the magnetic field at point **r**, we sum the effects of all the current elements. Let us take one current element at point **s**. Its contribution to the B field at **r** is:

$$d\,\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{\mathbf{I}\times(\mathbf{r}-\mathbf{s})}{|\mathbf{r}-\mathbf{s}|^3}\,dL \qquad (17)$$

where (**r-s**) is the vector that points from the bit of wire to the point at which we are calculating **B**. To work out the total field **B**(**r**), we integrate the expression along the wire. In most cases, this can be put more simply as

$$dB = \frac{\mu_0}{4\pi}\frac{I\sin\theta}{d^2}\,dL \qquad (18)$$

where d is the distance from wire element to point of measurement, and θ is the angle between the current flow and the vector from wire to measurement point.

It is possible to use this expression to calculate the magnetic field B on the axis of a coil of wire with N turns, radius R, carrying current I, if we measure the field at a distance D from the central point:

$$B = \frac{\mu_0 INR^2}{2\left(R^2 + D^2\right)^{3/2}} \tag{19}$$

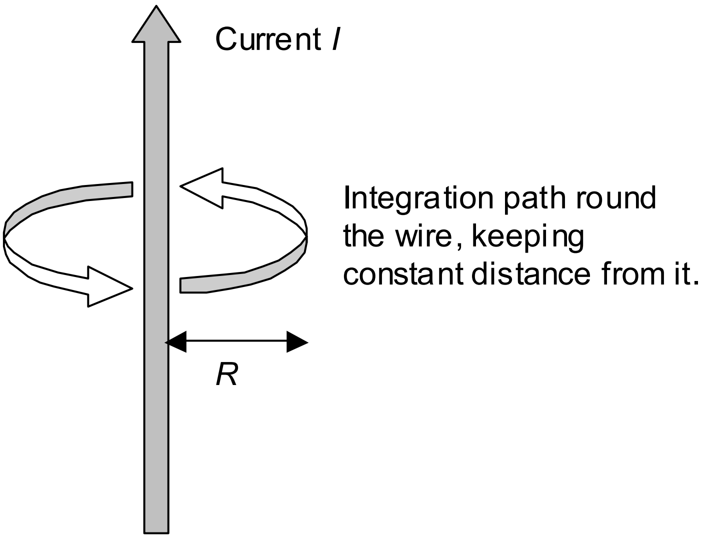
7.2.3.2 The Ampere Law

The alternative method of calculating fields is called the Ampere Law. You remember that Gauss' electrostatic law involved integrating a dot product over a surface? Well, the Ampere Law involves integrating a dot product along a line:

$$\oint_L \mathbf{B} \bullet \mathbf{dL} = \mu_0 I \tag{20}$$

In other words, if we choose a loop path, and integrate the magnetic flux density around it, we will find out the current enclosed by that loop.

Let us give an example. This formula is very useful for calculating the magnetic field in the vicinity of a long straight wire carrying current *I*. Suppose we take path *L* to be a circle of radius *R*, with the wire at its centre, and with the wire perpendicular to the plane of the circle, as shown in the diagram.

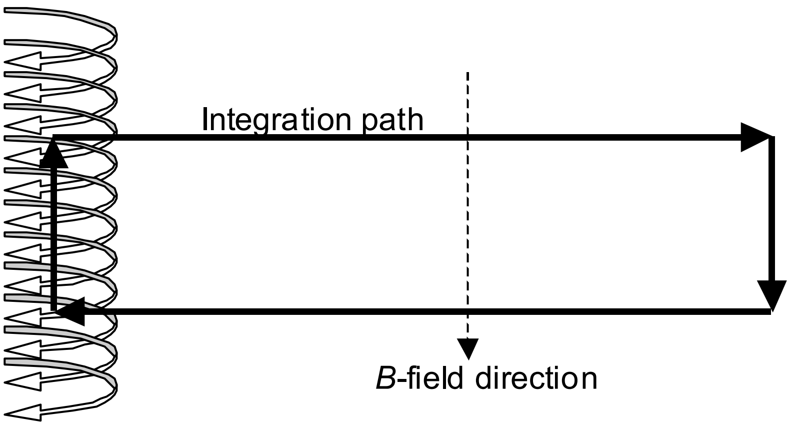

Current *I*
Integration path round the wire, keeping constant distance from it.
R

We know that **B** points round the wire, and therefore that **B** is parallel with **dL** (**dL** being the vector length of a small part of the path).

Furthermore, B=|**B**| must be the same all round the path by symmetry. Therefore:

$$\oint_L \mathbf{B} \bullet \mathbf{dL} = \oint_L B\,dL = B\oint_L dL = 2\pi RB = \mu_0 I$$
$$B = \frac{\mu_0 I}{2\pi R} \qquad (21)$$

Another splendid use of the Ampere Law is in calculating the magnetic field within the middle of a long solenoid with *n* turns per metre. This time we use a rectangular path, as shown in the diagram.



Only one of the sides of the rectangle "counts" in the integration – the one completely inside. Of the other three, one is so far away from the coil that there is no magnetic field, and the other two are perpendicular to the **B**-field lines, so the dot product is zero. The rectangle encloses *nL* turns, and hence a current of *nLI*. Therefore, Ampere's Law tells us:

$$BL = \mu_0 nLI$$
$$B = \mu_0 nI \qquad (22)$$

Before leaving Ampere's Law, we ought to give a word of caution. The Ampere Law is actually a simplified form of an equation called the Ampere-Maxwell Law[22]; and the simplification is only valid if there are no

[22] For the curious, the full Ampere-Maxwell Law states $\oint_L \mu_0^{-1}\mathbf{B} \bullet \mathbf{dL} = I + \iint_S \varepsilon_0 \varepsilon_{rel} \dot{\mathbf{E}} \bullet \mathbf{dS}$ where the surface S is any surface that has as its edge the loop L. This reduces to equation (20) when **E** is constant.

changing *electric* fields in the vicinity. Therefore you would be on dodgy ground using Ampere's Law near a capacitor that is charging up!

7.2.4 Flux, Inductance & Inductors

To sum up the last section: if there is a current, there will be a magnetic field. Furthermore, the strength of the magnetic field is proportional to the size of the current. It turns out, however, to be more useful to speak of the *magnetic flux* Φ. This is the product of the field strength B and the cross sectional area of the region enclosed by the magnetic field lines. We visualize this as the total 'number of field lines' made by the magnet.

We then write

$$\Phi = LI . \qquad (23)$$

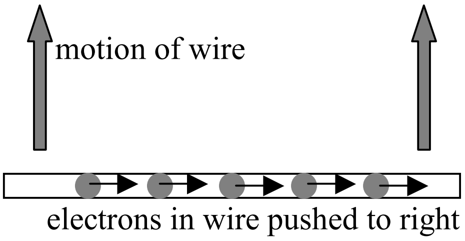
The total amount of magnetic field (the flux) is proportional to the current, and we call the constant of proportionality *L* – the self-inductance (or inductance for short). Any coil (or wire for that matter) will have an inductance, and this gives you an idea of how much magnetic field it will make when a current passes. You might think of an analogy with capacitors – the capacitance gives a measure of how much electric field a certain charge will cause (since C^{-1} = V/Q and V is proportional to E).

Now, this magnetic field is important, because a changing current will cause a changing magnetic field, and this will generate (or induce) a voltage, and therefore upset the circuit it is in. This is something we need to understand better – but before we do so, let us remind ourselves of the laws of electromagnetic induction:

7.2.5 Generators & Induction

If a wire 'thinks' it is moving magnetically, a voltage is induced in it. It doesn't matter whether the wire is still, and the magnetic field is moving or changing; or whether the wire is moving and the magnetic field is still. However for the two situations, different equations are used. The two equations are equivalent – however it is easier to remember them both than to prove the equivalence.

7.2.5.1 Stationary field: Moving wire



Magnetic field **B** down into paper.

Here the voltage is easy to calculate. The electrons in the wire must move with the wire, and if the wire is in a magnetic field, they will experience a force pushing them along the wire. This causes them to bunch up at one end of the wire – which in turn sets up an electric field which discourages further electrons to join the party. Assuming that the wire has length *L*, and is being moved at speed *u* through a uniform magnetic field **B** (perpendicular to *u* and *L* – if **B** is not so inclined, take only the component of **B** which is), once equilibrium is established

Electric force balances magnetic force

$$\begin{aligned} qE &= quB \\ \frac{V}{L} &= uB \\ V &= LuB \end{aligned} \tag{24}$$

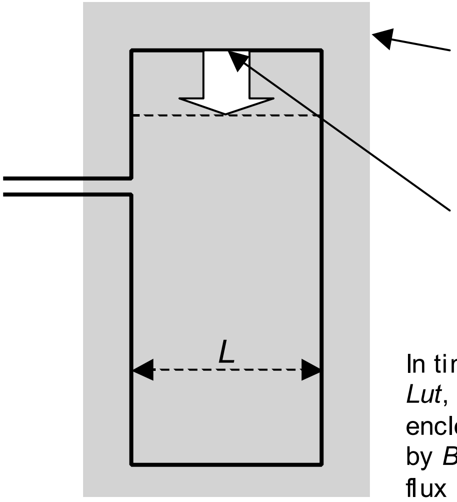
7.2.5.2 Stationary wire: Changing field

Here the voltage induced across the ends of a circular loop of wire (complete circuit, apart from the small gap) is given by

$$V = \frac{d}{dt}\iint_S B \bullet dS = \frac{d\Phi}{dt} \tag{25}$$

7.2.5.3 Equivalence?

The two expressions are very similar, as we shall see when we look at the first from a different perspective. Look at the diagram – we have completed a loop by using a very long wire.



Region of magnetic field *B*, pointing down into paper

End wire moved in direction of white arrow. Distance moved equals *ut*, where *u* is speed.

In time t, the area is reduced by *Lut*, so the magnetic flux enclosed by the wire is reduced by *BLut*. The rate of change of flux = induced voltage = *BLu*.

The total flux enclosed in the loop is equal to B × Area. In one second, the Area changes (decreases) by *Lu*, and hence the flux changes by *BLu*, and so we see that the first equation is a special case of the second. The second is better as it also allows us to calculate what will happen if the wire is stationary.

7.2.5.4 Direction of Induced Voltage

When equation (25) is written down, it is customary to put a minus sign in front of the derivative. The significance of this negation was Lenz's discovery. When the voltage is induced in a complete circuit, it will try to (and succeed in) driving a current. This current will produce a magnetic field. Lenz postulated that this 'produced' magnetic field always opposed the change being made.

Let us have an example. Imagine a large coil of wire (say, in a motor), with a decent sized current flowing in it. Now let us try and lower the current by reducing the voltage of the supply. This causes a reduction in the magnetic field, which in turn induces a voltage in the wire, which pushes a current in a desperate attempt to keep the original current going. On the other hand, if I were to try an increase the current in the motor (by increasing the supply voltage), the opposite would happen: the greater current causes the magnetic field to grow, which induces a voltage, which pushes a current to oppose this increase.

This is the origin of the phrase "back emf" to refer to the voltage induced across an inductor.

Now for a word about the minus sign. Yes, the voltage does go in opposition to the change in current, so I suppose one ought to write equation (25) with a minus sign. However if you do, please also write Ohm's law as $V = -IR$, since the voltage opposes the current in a resistor. I would prefer it, however, if you used common sense in applying your notation and were not stuck in ruts of "always" or "never" using the minus sign. We all remember which way V and I go in a resistor without being nagged about conventions, so I hope that there is no need for me to nag you when inductors come on the scene.

7.2.6 Inductors in circuits

Just as a capacitor requires an energy flow to change the voltage across it, an inductor requires an energy flow to change the current through it. It doesn't give in without a fight.

Let me illustrate this with a demonstration – or at least the story of one. A nasty physics teacher (yes, they do exist...) asked a pupil to come and hold one wire in his left hand, and one in his right – completing the circuit with his body. He grasped the first wire, and then the second – steeling himself for the shock which never came. The teacher then stalled him with questions, and kept him there, while he surreptitiously, slowly increased the current in the circuit, which also included an inductor.

Finally, the teacher said, “OK, now you can go back to your seat.” The unsuspecting pupil let go of the wires suddenly, and was very surprised when the inductor – indignant to have its current shut off so quickly – made its displeasure known with an arc from the wire to the pupil. It is foolish in the extreme to starve an inductor of its current. Its revenge will be short, but not sweet.

To see this, let us combine equations (23) and (25)

$$V = \frac{d\Phi}{dt} = \frac{d}{dt} LI = L\frac{dI}{dt} \tag{26}$$

Equation (26) is the definitive equation for inductors, just like Q=CV was for capacitors, and V=IR is for a resistor.

Again, we need to bear in mind the comments above, that the voltage is in the direction needed to oppose the change in current. You will often see (26) with a minus sign in it for that reason.

Let us now calculate how much energy is stored in the device (actually in its magnetic field)

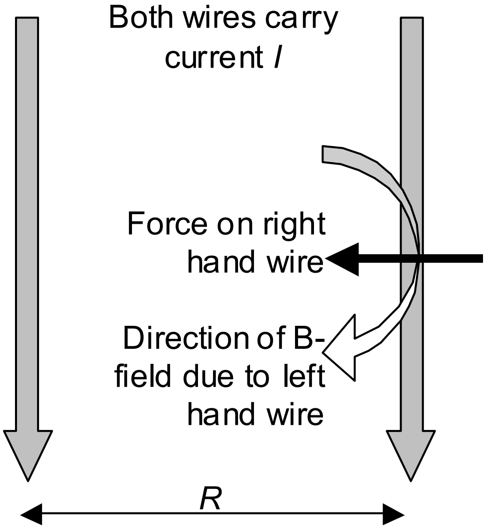
$$U = \int P dt = \int VI dt = \int L\frac{dI}{dt} I dt = L\int I dI = \left[\tfrac{1}{2} LI^2\right] \tag{27}$$

Since it seems sensible for the device to hold no energy when there is no current and no field, we take the constant of integration to be zero.

7.2.7 Relativity and Magnetism

At the beginning of the chapter we stated boldly that magnetism could be derived entirely from electrostatics and special relativity. Now is the time to justify this. We shall do so by deriving the same result two ways – once using magnetism, and once using relativity.

The phenomenon we choose is the mutual attraction of two parallel wires carrying equal current in the same direction, and we shall calculate the attractive force per metre of wire.


Both wires carry current *I*
Force on right hand wire
Direction of B-field due to left hand wire
R

7.2.7.1 A Classical Magnetic calculation

Equation (21) tells us that the magnetic field at a distance R from a straight infinite wire is

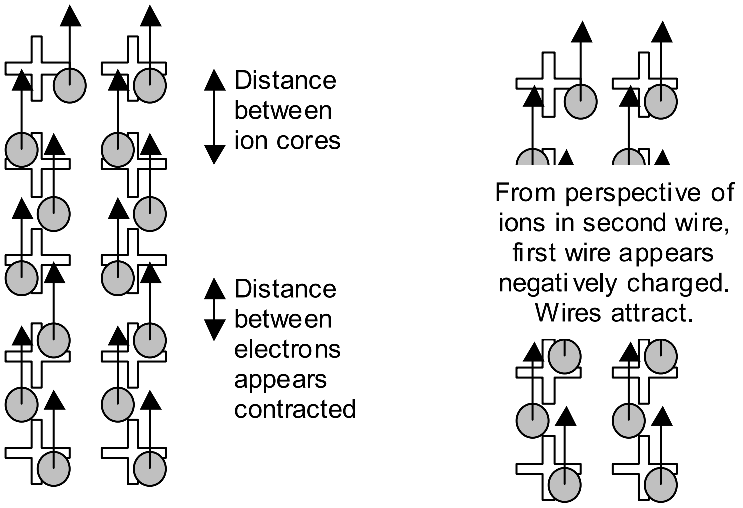
$$B = \frac{\mu_0 I}{2\pi R}$$

Therefore, the attractive force experienced by one metre of parallel conductor also carrying current I is

$$\frac{F}{L} = IB = \frac{\mu_0 I^2}{2\pi R} \qquad (28)$$

7.2.7.2 A Relativistic calculation

Each wire contains positive ion cores (say Cu^+ for a copper wire), and free electrons. Let us imagine the situation in the diagram below, with conventional current flowing downwards in both wires. The ion cores are stationary, while the electrons move upwards. If the free (electron) charge per metre of cable is called λ_0, then the current is related to the electron speed u by the equation $I=\lambda_0 u$.



Let's imagine the situation as perceived by an ion core in the second (right hand) wire. It sees the ion cores in the other wire stationary, with charge density λ0 and finds them repulsive. However it also sees the electrons on the other wire, and is attracted by them. The electrons are travelling, and therefore our observant ion core sees the length between adjacent electrons contracted. Therefore as far as it is concerned, the electron charge density is higher than the ion core charge density by factor $\gamma=\left(1-u^2/c^2\right)^{-1/2}$. Therefore its overall impression is attraction – with a total effective electric field (as derived in equation 6)

Total field = Field due to ion cores + Field due to loose electrons

$$\begin{aligned} E &= \frac{\lambda_0}{2\pi\varepsilon_0 R}-\frac{\gamma\lambda_0}{2\pi\varepsilon_0 R} \\ &= -(\gamma-1)\frac{\lambda_0}{2\pi\varepsilon_0 R} \\ &\approx -\left(\frac{u^2}{2c^2}\right)\frac{\lambda_0}{2\pi\varepsilon_0 R} \end{aligned} \tag{29}$$

where the final stage has made use of a Binomial expansion of (γ-1) to first order in u/c. Now the total charge of ion cores experiencing this field per metre is of course λ0. Therefore the total attraction of the ion cores in the second wire to the first wire (electrons & ion cores) is

$$\begin{aligned} F_{ions} &= QE \approx -\left(\frac{u^2}{2c^2}\right)\frac{\lambda_0^2}{2\pi\varepsilon_0 R} \\ &= -\frac{I^2}{4c^2\pi\varepsilon_0 R} \end{aligned} \tag{30}$$

This is the force experienced by the ion cores in the right hand wire. By an exactly equivalent argument, the electrons in the right hand wire see their counterparts in the left wire as stationary, and see the left hand ion cores bunched up, and therefore more attractive. Therefore the electrons in the right hand wire experience an equal attraction, and the total attractive force between the wires is twice the figure in equation (30).

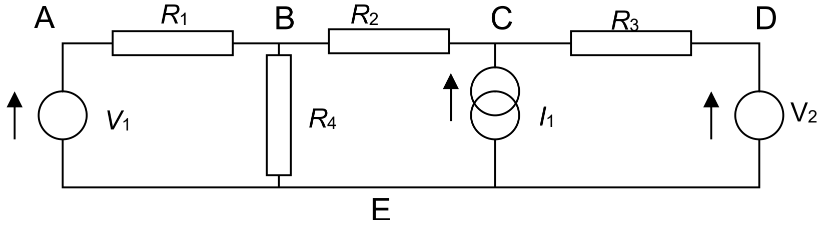
Finally, if you get a book of physical constants and a calculator, you will discover that $\mu_0 \varepsilon_0 = c^{-2}$. Therefore the total force agrees exactly with our magnetic calculation in equation (28).

7.3 Circuits – putting it together

In this section, we look at combining resistors, capacitors and inductors in electrical circuits. There are two reasons for doing this. Firstly, once you have left school, you will be faced with complicated electronic networks, and you need to be able to analyse these just as well as the simple series and parallel arrangements you dealt with in the classroom. Secondly, engineers frequently use electric circuits as models or analogies for other systems (say, an oscillating bridge or the control of the nervous system over the muscles in a leg) – the better you understand electric circuits, the better you will understand any linked system.

7.3.1 Circuit Analysis

Our aim here is to be able to solve a circuit like the one below. The circles represent constant-voltage sources (a bit like cells or batteries) and the linked circles represent constant-current sources. Our aim is to find voltage difference across each component, and also to work out the current in each resistor.



In order to solve the circuit, we use two rules – the Kirchoff Laws. Kirchoff's 1st says that the total current going into a junction is equal to the total current leaving it. Therefore, at B in the circuit below, we would say that $I_{BE} = I_{AB} + I_{CB}$, where I_{BE} means the current flowing from B to E (through R_4).

Kirchoff's 2nd Law is that voltages always add up correctly. In other words, no matter which route we took from E to B, say, we would agree on the voltage difference between E and B. In symbols, if V_{BE} means the difference in potential (as measured by a voltmeter) between B and E, then we have $V_{BE} = V_{AE} + V_{BA}$. This is basically the same thing as the law of conservation of energy. The voltage (or, more strictly, the *potential*) at B, V_B, is the energy content of one coulomb of charge at B. In travelling to E, it will lose V_B-V_E joules, irrespective of the route taken.[23] In fact, we assume the truth of Kirchoff's 2nd Law whenever we say, "let's call the voltage at A 'V_A'," for we are assuming that the voltage of A does not depend on the route used to measure it.

Using these two rules, and the equation for the current through a resistor (for example, $V_{BA} = I_{BA} R_1$), we may write down a set of equations for the circuit. Notice that because currents are said to go from + to –, this means that if V_{AB} (the voltage of A, measured relative to B) is positive, then V_A is bigger than V_B, and hence I_{AB} will be positive too. To make the notation easier we will take the potential at E to be zero. In symbolic form, this means that we shall call V_{BE} (that is, $V_B–V_E$) V_B for short.

Kirchoff's First Law:

$$I_{EA} = I_{AB}; \quad I_{BE} = I_{AB}+I_{CB}; \quad I_1 = I_{CB}+I_{CD}; \quad I_{CD} = I_{DE}$$

Kirchoff's Second Law:

$$\begin{aligned} V_B &= I_{BE} R_4 \\ &= V_A + V_{BA} = V_1 - I_{AB} R_1 \\ &= V_D + V_{CD} + V_{BC} = V_2 + I_{CD} R_3 - I_{CB} R_2 \end{aligned}$$

After elimination, the equations reduce to two:

$$V_1 - I_{AB} R_1 = V_2 + I_1 R_3 - (R_3 + R_2) I_{CB} = (I_{AB} + I_{CB}) R_4,$$

and from these the currents I_{AB} and I_{CB} can be found (after a bit of messy algebra). After this, the remaining currents and voltages are straightforward to determine.

These principles can be used to solve any circuit. However, as networks get bigger, it is useful to find more prescriptive methods of solution, which could be used by a computer. We shall cover two methods here – for certain problems, they may be more efficient than the direct application of Kirchoff's Laws.

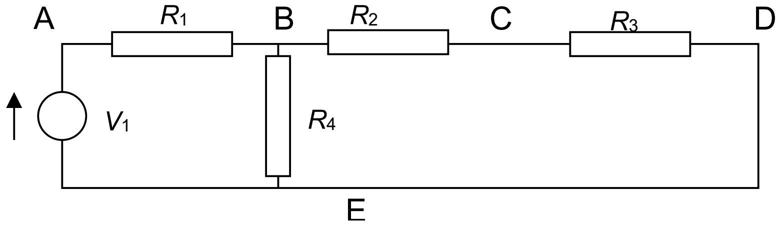
[23] To see why the Law of Conservation of Energy is involved, let us suppose that our coulomb of charge would lose 5J going from B to E via A, whereas it would lose 3J in going directly. All it would have to do is go direct from B to E, then back to B via A and it would be back where it started, having gained 2J of energy! This is not allowed.

7.3.1.1 Method of Superposition

The method of superposition relies on the fact that for a simple resistor, the current is proportional to the voltage. It follows that if current I_1 causes a voltage difference of 3V, and current I_2 causes a voltage difference of 5V, then current I_1+I_2 will cause an 8V p.d. across the component. Here is the procedure:

- Choose one of the supply components.
- Remove the other supply components from the circuit. Replace voltage sources with direct connections (short circuits), and leave breaks in the circuit where the current sources were (open circuits).
- Calculate the current in each wire, and the voltage across each component.
- Repeat the procedure for each supply component in turn.
- The current in each wire for the original (whole) circuit is equal to the sum of the currents in that wire due to each supply unit.
- The voltage across each component in the original (whole) circuit is equal to the sum of the voltages across that component due to each supply unit.

Let's use this method to analyse the circuit above. We start by considering only source V_1. Removing the other supply components gives us a circuit like this.



This circuit is easier to analyse as it only has one supply. Supply V_1 feeds a circuit with resistance

$$R_1 + \{R_4 \,//\, (R_2+R_3)\}$$

$$= R_1 + \frac{R_4\left(R_2+R_3\right)}{R_4+R_2+R_3}$$

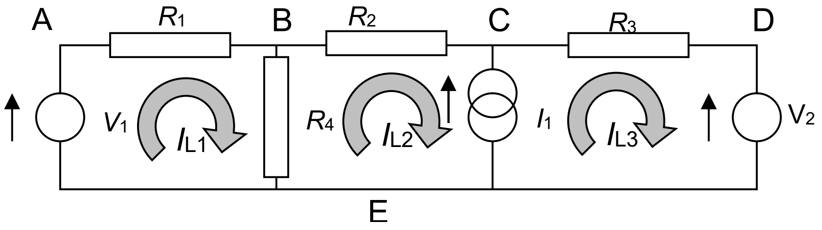
where // means 'in parallel with.' Accordingly, the current supplied by V_1 (and the current through R_1 which is in series with it) is equal to V_1 divided by this resistance. The voltages of points B, C and D can be calculated, as can the current in each wire. We make a note of the values, and add to them the results of analyses of circuits only containing I_1 and only containing V_2.

You may find this method good in the sense that you only have to deal with one supply component at a time – and therefore all you need to know is how to combine resistors (something you've done before). Having said that, we end up analysing three circuits rather than one, so it is more time consuming.

Before leaving the method, you may be curious why voltage sources were replaced with short circuits, and current sources with open circuits. Here's the reason. A voltage source does not change the voltage across its terminals, no matter what the current is (d Voltage / d Current = $R_{\text{equivalent}}$ = 0). The only type of resistor which behaves likewise is a perfect conductor (0Ω). Similarly, a current source does not change its current, no matter what the voltage (d Current / d Voltage = 1 / $R_{\text{equivalent}}$ = 0). The equivalent resistor in this case is a perfect insulator (∞Ω) which lets no current through ever.

7.3.1.2 Method of Loop Currents

Here we break the circuit down into the smallest loops it contains. Here there are three loops:



- E to A to B and back to E (loop 1),
- E to B to C to E (loop 2), and
- E to C to D to E (loop 3).

We call the current in loop 1 "loop current" number one (I_{L1}), with I_{L2} and I_{L3} representing the currents in the other two loops. We then express all other currents in terms of the loop currents. Clearly, $I_{AB} = I_{L1}$, since R_1 is in the first loop alone. Similarly, $I_{BC} = I_{L2}$, and $I_{CD} = I_{L3}$.

The current through R_4 is more complex, since this resistor is part of two of the loops. We write $I_{BE} = I_{L1} – I_{L2}$. Here I_{L1} is positive, since I_{BE} is in the same direction as I_{L1}, whereas I_{L2} (which goes from E to B then on to C) is in the opposite direction. These designations automatically take care of Kirchoff's First Law. Notice that by this method, $I_1 = I_{L3} – I_{L2}$.

Each loop now contributes one equation – Kirchoff's 2nd law around that loop. Clearly, if you go all the way round the loop, you must return to the voltage you started with. Taking the first loop as an example, we have:

$$0 \quad = V_{AE} + V_{BA} + V_{EB}$$

$= V_1 - I_{AB}\,R_1 - I_{BE}\,R_4$
$= V_1 - I_{L1}\,R_1 + (I_{L1} - I_{L2})\,R_4.$

In a similar way, we write equations for each of the other two loops[24]. We then have three equations in three unknowns (the three loop currents), which can be solved. The end result is the same as for a direct 'sledgehammer' approach with Kirchoff's Laws – but the method is more organized.

7.3.2 Alternating Current

Having looked at circuits with resistors in them, we next turn our attention to circuits with inductors and capacitors as well. For a direct current, the situation is easy. After a brief period of settling down, there is no voltage drop across an inductor (because the current isn't changing), and a capacitor doesn't conduct at all.

For alternating currents the situation is more complicated. Let us suppose that the supply voltage is given by $V=V_0 \cos \omega t$. It turns out that the circuit will settle down to a steady behaviour (called the steady state). Once this has happened, the voltage across each component (and the current through each component) will also be a cosine wave with frequency ω, however it may not be in phase with the original V.

7.3.2.1 Resistor, capacitor and inductor

We start with the three simplest circuits – the lone resistor, the lone capacitor and the lone inductor, each supplied with a voltage $V=V_0 \cos \omega t$.

For the resistor, I=V/R, so the result is straightforward.

For the capacitor, Q=VC, and if we take I as positive in the direction which charges the capacitor, then

$$\begin{aligned} I &= \frac{dQ}{dt} = \frac{d}{dt} CV_0 \cos \omega t \\ &= -\omega CV_0 \sin \omega t \\ &= \omega CV_0 \cos\left(\omega t + \tfrac{1}{2}\pi\right) \\ &= -\omega CV_0 \cos\left(\omega t - \tfrac{1}{2}\pi\right) \end{aligned} \tag{31}$$

For the inductor, $V = L\, dI/dt$, so

[24] It may help when writing the equations to notice the pattern: voltage sources count positively if you go through them from – to +, but negatively if you go from + to –. The voltages across resistors (e.g. $V_{BA} = I\,R_1$) count negatively if you go through them in the same direction as the current, and positively if you go through them the opposite way to the current.

$$\begin{aligned}\frac{dI}{dt} &= \frac{V_0}{L}\cos\omega t \\ I &= \frac{V_0}{L\omega}\sin\omega t \\ &= \frac{V_0}{L\omega}\cos\left(\omega t - \tfrac{1}{2}\pi\right)\end{aligned} \tag{32}$$

where we have taken the constant of integration to be zero. Failure to do so would lead to a non-zero mean current, which is clearly impossible as the mean supply voltage is zero.

7.3.2.2 Reactance and Impedance

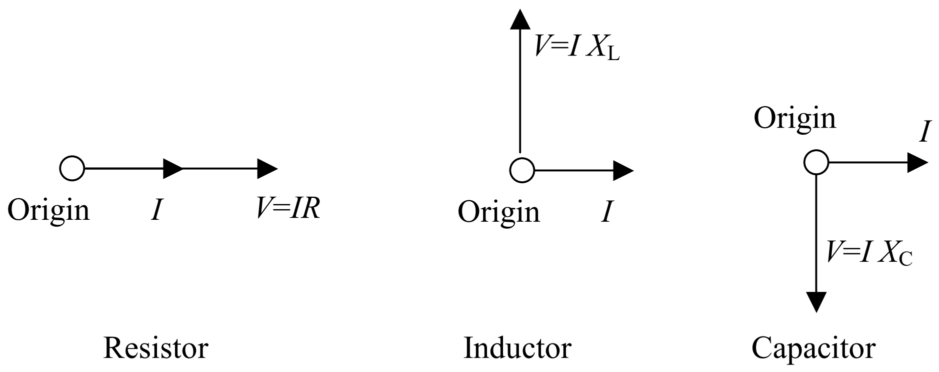
For resistors, the current and voltage are proportional, and consequently are in phase – one peaks at the same time as the other. For the other two components, this is not the case. The voltage is $\pi/2$ radians (or 90°) out of phase with respect to the current. Inductor currents peak 90° later than the voltage (the current *lags* the voltage), whereas capacitor currents peak 90° before the voltage (the current *leads* the voltage). Nevertheless, the amplitude of the voltage is still proportional to the amplitude of the current, and we call the ratio of the amplitudes the reactance (X).

$$\begin{aligned}X_L &= \omega L \\ X_C &= -\frac{1}{\omega C}\end{aligned} \tag{33}$$

By convention, we take reactance to be positive if the current lags the voltage by 90°, and negative if it leads by 90°. For capacitors and inductors in series, the total reactance is equal to the sum of the individual components' reactances – just as resistances add in series. Similarly, the formula for combining reactances in parallel is the same for that used for the resistance of resistors wired in parallel.

When a circuit is constructed with resistors, capacitors and inductors, then we need a way of analysing a circuit with both resistances and reactances. We visualise the situation using a 2D (phasor) diagram.

For any component or circuit, both voltage and current are represented by vectors. The length of the lines gives the amplitude, and the angle between the vectors gives the phase difference. By convention, we imagine the vectors to rotate about the origin in an anticlockwise direction (once per time period of the alternating current). The vectors for a resistor, capacitor and inductor are shown below.



As the arrows rotate anticlockwise, for the inductor, V comes before I. With the capacitor, I comes before V. This accurately represents the phase relationships between voltage and current for these components.

For a set of components in series, the current I will be the same for all of them. We usually draw the current pointing to the right. Voltages across inductors will then point up, those across resistors point right, and those across capacitors point down. By adding these voltages vectorially, we arrive at the voltage across the set of components – and can calculate its amplitude and phase relationship with respect to the current.

Similarly, for components in parallel, the voltage will be the same for each. We thus put voltage pointing to the right. Currents in capacitors now point up, currents in resistors point right, and currents in inductors point down. The total current is given by the vector sum of the individual currents.

In all cases, we call the ratio of the voltage amplitude to the current amplitude the impedance (Z) irrespective of the phase difference between the current and voltage.[25] In general the impedance of a component is related to resistance and reactance by $Z^2 = R^2 + X^2$.

7.3.2.3 Complex Numbers and Impedance

If you are familiar with complex numbers, there is an easier way of describing all of this, using the Argand diagram in place of 2-dimensional vectors. The impedance Z is a now a complex number $Z = R + iX$, with R as its real part and X as its imaginary part.

The complex impedances of a resistor, capacitor and inductor are accordingly written as R, $-i/\omega C$ and $i\omega L$ respectively. The impedance of a set of components in series is given by the sum of the individual impedances. For a parallel network, Z^{-1} of the network is given by the

[25] In other words, a resistance is a special kind of impedance with zero phase difference, and a reactance is a special kind of impedance when the phase difference is 90°.

sum of Z^{-1} for each component, where 'inverse' (or 'reciprocal') is calculated in the usual way for complex numbers.

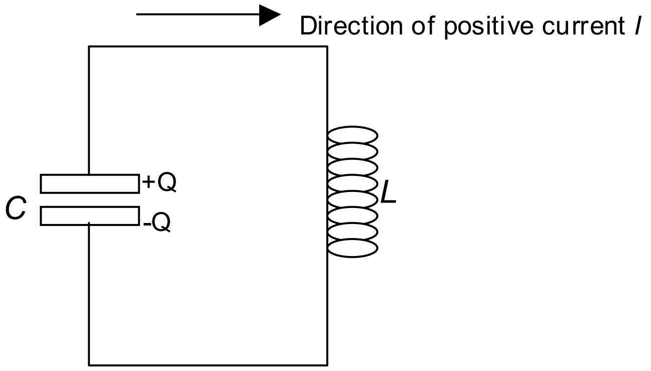
7.3.2.4 *Root Mean Square values*

You will also need to remember the definition of RMS voltage and current in an a.c. circuit. For a resistor, remember that the RMS supply voltage is the d.c. voltage which would supply the same mean **power** to the device.

$$\begin{aligned} \overline{P} &\equiv \frac{V_{rms}^2}{R} = \frac{\overline{V_0^2 \cos^2 \omega t}}{R} = \frac{V_0^2}{2R} \\ \Rightarrow \quad V_{rms} &= \frac{1}{\sqrt{2}} V_0 \end{aligned} \tag{34}$$

7.3.3 Resonance

One further circuit needs a mention, and that is the simple circuit of an inductor and a capacitor connected together, as shown in the diagram below. Both the voltage and current for the two components must be the same, and so with the sign conventions chosen in the diagram:



$$\begin{aligned} \frac{Q}{C} &= L\frac{dI}{dt} = -L\frac{d^2Q}{dt^2} \\ \frac{d^2Q}{dt^2} &= -\frac{1}{LC}Q \end{aligned} \tag{35}$$

This is an equation of 'simple harmonic motion' with angular frequency ω, where $\omega^2 = 1/LC$. This circuit can therefore oscillate at this frequency, and this makes it useful in radio receivers for selecting the frequency (and hence radio station) which the listener wants to detect.

7.4 Questions

1. Calculate the size of the repulsion force between two electrons 0.1nm apart.

2. In this question, you will make an estimate for the size of a hydrogen atom. Suppose an electron moves in a circular path around the proton, with radius *r*. Calculate, in terms of r, the potential energy of the atom (it will be negative, of course), the speed of the electron, and its kinetic energy. Now write down an expression for the total energy of the electron. Find the value of *r* which minimizes this total energy, and compare it to the measured radius of a hydrogen atom, which is about 5×10^{-11} m.

3. What fraction of the electrons in the solar system would have to be removed in order for the gravitational attractions to be completely cancelled out by the electrostatic repulsion?

4. A cloud of electrons is accelerated through a 20kV potential difference (so that their kinetic energy of each coulomb of electrons is 20kJ). Calculate their speed.

5. A beam of 20kV electrons is travelling horizontally. An experimenter wishes to bend their path to make them travel vertically (at the same speed) using a region with a uniform electric field. This region is square with side length 5cm. Calculate the size and direction of electric field needed to do this. What would happen to a beam of 21kV electrons passing this region?

6. A different experimenter wishes to bend the beam of 20kV electrons using a magnetic field. She chooses to bend the beam round a circular path of radius 3cm. What magnitude and direction of magnetic field is needed? What would happen to a beam of 21kV electrons passing this region?

7. I wish to make a 1T magnetic flux density inside a long coil (or solenoid) with radius 5mm. I use wire which can carry a current of 4A. How many turns per metre of coil are needed?

8. ‘Clamp’ ammeters used by electricians can measure the current in a wire without needing to break the wire. A metal loop encloses the wire, and the magnetic field around the wire is measured. If the loop is circular, with radius 3cm, and is centred on the wire, calculate the magnetic flux density measured when the current in the wire is 100A.

9. Calculate the impedance of a 20Ω resistor wired in series with a 3mH inductor when fed with alternating current of 50Hz. A capacitor wired in parallel with this combination causes the overall reactance of the circuit to become zero at 50Hz (in other words, the voltage is in phase with the total current). Calculate the capacitance of the capacitor.

8 Small Physics

The rules, or laws, of classical mechanics break down typically in three cases. We have seen that when things start going quickly, we need to take special relativity into account. Another form of relativity – the general theory – is needed when things get very heavy, and the gravitational fields are strong. The third exception is very mysterious – and occurs often when we deal with very small objects like atoms and electrons. This is the realm of quantum physics, and many of its discoverers expressed horror or puzzlement at its conclusions and philosophy.

Having said that, there is no need to be frightened. While there is much we do not understand, a set of principles have been set up which allow us to perform accurate calculations. Furthermore, those calculations agree with experiment to a high degree of accuracy. The development of the transistor, hospital scanner, and many other useful devices testify to this. The situation is analogous to a lion-tamer who can get the lion to jump through a hoop, though she doesn't know what is going on inside the lion.

8.1 *Waves and Particles*

Quantum objects, like electrons and photons (packets of light) are difficult to describe. As physicists, we have two models, or descriptions, which we are comfortable using – the wave and the particle.

Waves can interfere, they have a wavelength, frequency and intensity, and they carry energy by means of fluctuations in a medium. The intensity is continuous – it can take any value.

Particles on the other hand, are lumps. They possess individual masses, energies and momenta. They most certainly do not interfere – if you add 1 apple to 1 apple, you always get 2 apples. Finally they only come in integer numbers. You can have one, or two, or 45 678 543; but you can‘t have half.

The electron fits neither description. Light fits neither description. The descriptions are too simplistic. However there are instances when the particle description fits well – but it doesn't always fit. There are also instances when the wave description fits well – but it doesn't always fit.

Given that a particular electron beam may behave like particles one minute, and waves the next, we need some kind of 'phrase book' to convert equivalent measurements from one description to the other. Quantum theory maintains that such a 'phrase book' exists.

The total number of particles (in the particle picture) is related to the intensity (in the wave picture). The exact conversion rate can be determined using the principle of conservation of energy.

The energy per particle (in the particle picture) is related to the frequency (in the wave picture) by the relationship

$$\text{Energy of one particle (in J)} = h \times \text{Frequency of wave (Hz)}, \quad (1)$$

where h is the Planck constant, and has a value of 6.63×10^{-34} Js. You may also come across the constant 'h-bar' $\hbar \equiv h/2\pi$, which can be used in place of *h* if you wish to express your frequency as an angular frequency in radians per second.

The momentum per particle (in the particle picture) is related to the wavelength of the wave (in the wave picture) by the relationship

$$\text{Momentum of one particle (kg m/s)} = \frac{h}{\text{Wavelength (m)}} \ . \quad (2)$$

8.2 *Uncertainty*

The bridge between wave and particle causes interesting conclusions. We have seen in the chapter on Waves that a wave can have a well-defined frequency or duration (in time), but not both. This was expressed in the bandwidth theorem:

$$\Delta f \, \Delta t > 1 .$$

When combined with our wave-particle translation, we obtain a relationship between *energy* and time:

$$\Delta E \, \Delta t > h . \quad (3)$$

In other words, only something that lasts a long time can have a very well known energy.

Let us have an example. Suppose a nucleus is unstable (radioactive), with a half-life *T*. Seeing as the emission of the radiation is a process that typically 'takes' a time *T*, the energy of the alpha particle (or whatever) has an inherent uncertainty of $\Delta E \approx h/T$. If we were watching a spectrometer, monitoring the radiation emitted, we would expect to see a spread of energies showing this level of uncertainty.

The bandwidth theorem also has something to say about wavelength:

$$\Delta\left(\frac{1}{\lambda}\right)\Delta x > 1 .$$

This has the quantum consequence:

$$\Delta p \, \Delta x > h \ . \tag{4}$$

This is frequently stated as, "You can't know both the momentum and the position of a particle accurately." It might be better stated as, "Since it is a bit like a wave, it can not *have* both a well defined position and momentum."

We can use this to make an estimate for the speed of an electron in an atom. Atoms have a size of about 10^{-10} m. Therefore, for an electron in an atom, $\Delta x \approx 10^{-10}\,\mathrm{m}$. So, using equation (4), $\Delta p \approx 10^{-23}\,\mathrm{kg\,m/s}$. Given the electron mass of about 10^{-30} kg, this gives us a speed of about 10^{7} m/s – about a tenth the speed of light!

Caution: Please note that we haven't defined precisely what we mean by uncertainty (Δ). That is why we have only been able to work with approximate quantities. In more advanced work, the definition can be tightened up (to mean, say, standard deviation). However it is better for us to leave things as they are. In any case, it is never wise to state uncertainties to more than one significant figure!

8.3 ***Atoms***

Putting things classically for a moment, the electron orbits the nucleus. While a quantum mechanic thinks this description very crude, we shall use it as a starting point.

Now, let's imagine the electron as a wave. For the sake of visualization, think of it as a transverse wave on a string that goes round the nucleus, at a distance *R* from it. If the electron wave is to make sense, the string must join up to form a complete circle. Therefore the circumference must contain a whole number of wavelengths.

$$\begin{aligned} 2\pi R &= n\lambda \\ 2\pi R &= \frac{nh}{p} \\ pR &= \frac{nh}{2\pi} = n\hbar \\ L &= n\hbar \end{aligned} \tag{5}$$

The conclusion of this argument is that the angular momentum of the electron, as it goes round the nucleus, must be in the $\hbar$-times table.

The argument is simplistic, in that the quantum picture does not involve a literal orbit. However, the quantum theory agrees with the reasoning above in its prediction of the angular momentum.

Given that the angular momentum can only take certain values (we say that it is quantized – it comes in lumps), we conclude that the electron can only take certain energies. These are called the energy levels, and we can work out the energies as follows:[26]

$$\begin{aligned}\text{Kinetic Energy} &= \frac{L^2}{2I} = \frac{n^2\hbar^2}{2mR^2} \\ \text{Potential Energy} &= -\frac{Ze^2}{4\pi\varepsilon_0 R}\end{aligned}, \qquad (6)$$

where *Z* is the number of protons in the nucleus. We are, of course, ignoring the other electrons in the atom – hence this model is only directly applicable to hydrogen.[27] Next, we use the relationships derived in section 1.2.2, where we showed that for a Coulomb attraction,

Potential energy = -2 × Kinetic Energy

$$\begin{aligned}\text{Potential energy} &= 2E \\ \text{Kinetic energy} &= -E\end{aligned} \qquad (7)$$

where *E* is the total energy of the electron. We may use this information to eliminate the radius in equations (6), obtaining:

$$E = -\frac{Z^2 e^4 m}{2(4\pi\varepsilon_0\hbar)^2} \times \frac{1}{n^2}. \qquad (8)$$

When dealing with atoms, the S.I. units can be frustrating. A more convenient unit for atomic energies is the electron-volt. This is the energy required to move an electron through a potential difference of one volt, and as such it is equal to about 1.60×10^{-19} J. In these units, equation (8) can be re-written:

$$E = -\frac{Z^2}{n^2} \times 13.6\,\text{eV}. \qquad (9)$$

This form should be remembered. It will help you to gain a ‘feel’ for the energies an electron can have in an atom, and as a result, it will help you spot errors more quickly.

[26] The kinetic energy is calculated using the relationships derived in chapter 3. If you do not wish to go in there, a simpler derivation can be employed. L=mvR, where v is the speed. Therefore the kinetic energy $mv^2/2 = L^2/(2mR^2)$.

[27] Hydrogen, that is, and hydrogen-like ions: which are atoms that have had all the electrons removed apart from one.

When an electron moves from one level (*n* value) to another, energy is either required or given out. This is usually in the form of a photon of light that is absorbed or emitted. The energy of the photon is, as usual, given by the Planck constant, multiplied by the frequency (in Hz) of the light.

If an electron moves from orbit n_1 to n_2, where $n_2<n_1$, the frequency of photon emitted is therefore given by:

$$f = Z^2\left(\frac{1}{n_2^2} - \frac{1}{n_1^2}\right)3.29\times 10^{15}\ \text{Hz}\,. \qquad (10)$$

Similarly, the formula gives the frequency of photon required to promote an electron from n_2 to n_1. The frequency of photon required to remove the electron completely from the atom (if it starts in level n_2) is also given by equation (10), if n_1 is taken as infinite.

8.4 Little Nuts

As far as Romans were concerned, the stones in the middle of olives were 'little nuts' or **nuclei**. We shall thus turn our attention to 'nutty physics'.

The nuclear topics required for the International Olympiad are common to the A-level course. In this book we shall merely state what knowledge is needed. You will be able to find out more from your school textbook.

8.4.1 Types of radiation

Alpha decay: in which a helium nucleus (two protons and two neutrons) is ejected from the unstable nucleus.

Beta decay: in which some weird nucleonic processes go on. In all beta decays, the total number of nucleons (sometimes called the *mass number*) remains constant.

In β- decay (the most common), a neutron turns into a proton and an electron. The electron is ejected at speed from the nucleus.

There are two other forms of beta radiation. In β+ decay, a proton turns into a neutron and an anti-electron (or positron). The positron flies out of the nucleus, and annihilates the nearest electron it sees. The annihilation process produces two gamma rays.

The other permutation is electron capture (ε) in which an electron is captured from an inner (low *n*) orbit, and 'reacts' with a proton to make a neutron. This phenomenon is detected when another electron descends to fill the gap left by the captive – and gives out an X-ray photon as it does so.

Gamma decay: in which the nucleus re-organizes itself more efficiently, leading to a drop in its internal potential energy. This energy is released as a burst of electromagnetic radiation – a gamma ray photon. By convention high energy photons are called X-rays if they come from the electrons in an atom, and gamma rays if they come from a nucleus.

8.4.2 Radioactive decay

It is beyond the wit of a scientist to predict when a particular nucleus will decay. However we have so many radionuclides in a sample that the average behaviour can be modelled well.

The rate of decay (number of decays per second) is proportional to the number of nuclei remaining undecayed. This 'rate of decay' is called the activity, and is measured in Becquerels (Bq). We define a parameter λ to be the constant of proportionality:

$$\begin{aligned} I &= -\frac{dN}{dt} = \lambda N \\ N &= N_0 e^{-\lambda t} \\ I &= \lambda N_0 e^{-\lambda t} = I_0 e^{-\lambda t} \end{aligned} \tag{11}$$

where N_0 is the initial number of radionuclides, and I_0 is the initial activity.

The half-life (T) is the time taken for the activity (or the number of undecayed nuclei) to halve. This is inversely proportional to λ, as can be seen:

$$\begin{aligned} \tfrac{1}{2} &= \exp(-\lambda T) \\ \ln \tfrac{1}{2} &= -\lambda T \\ \ln 2 &= \lambda T \quad . \\ T &= \frac{\ln 2}{\lambda} \end{aligned} \tag{12}$$

If a half-life is too long to measure directly, the value of λ can be determined if *I* and *N* are known separately. *I* would be measured simply by counting the decays in one year (say), while *N* would be measured by putting a fraction of the sample through a mass spectrometer.

8.4.3 Nuclear Reactions

Now for the final technique: You will need to be able to calculate the energy released in a nuclear reaction. For this, add up the mass you started with, and add up the mass at the end. Some mass will have gone missing. Remembering that mass and energy are basically the same thing – the 'lost mass' is the energy released from the nuclei.

Finally, to convert from kilograms to joules, multiply by c^2 (the speed of light squared).

A particular case of a nuclear reaction is the 'annihilation' reaction, in which a particle and its antiparticle (say an electron and a positron) react together. The matter vanishes, and the energy appears in the form of two gamma rays.[28]

8.5 Questions

1. Calculate the wavelength and frequency of the quantum associated with a 60g ball travelling at 40m/s. Why don't we observe interference effects with balls such as this?

2. Blue light has a wavelength of approximately 400nm, while red light has a wavelength of approximately 650nm. Calculate the energies of photons of blue and red light (a) in joules (b) in electron-volts (eV). One electron-volt is equal to 1.602×10^{-19} J.

3. Work out the wavelengths of light emitted when electrons from the n=5, 4, and 3 levels 'descend' to the n=2 level. Why do you think that these transitions were more important in the historical development of atomic theory than the 'more fundamental' transitions going down to the n=1 level?

4. Calculate the energies of the n=1, 2, 3, 4 and 10 levels for an ionized helium atom (a helium nucleus with a single electron).

5. Calculate the size of a muonic hydrogen atom in comparison with a normal hydrogen atom. A muonic hydrogen atom has a muon rather than an electron moving near a proton. The muon has a charge equal to that of an electron, but its mass is 207 times greater.

6. In this question, you will make an estimate for the size of a hydrogen atom. Suppose that the atom's radius is *r*. Then the uncertainty in the electron's position is 2*r*. Use the uncertainty principle to work out the uncertainty in its momentum, and from this work out its typical kinetic energy, in terms of *r*. The electron's typical electrostatic energy is $-e^2/4\pi\varepsilon_0 r^2$. Find the value of *r* which minimizes the total energy of the electron.

7. Calculate the energy liberated in the fusion reaction ${}^{2}_{1}\mathrm{H}+{}^{3}_{1}\mathrm{H}\rightarrow{}^{4}_{2}\mathrm{He}+{}^{1}_{0}\mathrm{n}$. The masses of the particles are given in the table in unified mass units (u). 1u = 1.66043×10^{-27} kg.

[28] When analysing the collision, you find that you can not satisfy momentum and energy conservation at the same time if only one photon is produced.

^{2}H	2.014102
^{3}H	3.016049
^{4}He	4.002604
n	1.008665

9 Practical Physics

9.1 Errors, and how to make them[29]

Every dog has its day, every silver lining has its cloud, and every measurement has its error.

If you doubt this, take (sorry – borrow with permission) a school metre stick, and try and measure the length of a corridor in your school. Try and measure it to the nearest centimetre. Then measure it again. Unless you cheated by choosing a short corridor, you should find that the measurements are different. What's gone wrong?

Nothing has gone wrong. No measurement is exact, and if you take a series of readings, you will find that they cluster around the 'true value'. This spread of readings is called random error – and will be determined by the instrument you use and the observation technique. To be more precise and polite, this kind of 'error' is usually called uncertainty, as this word doesn't imply any mistake or incompetence on the part of the scientist.

So, whenever you write down a measurement, you should also write down its uncertainty. This can be expressed in two ways – absolute and relative.

The absolute uncertainty gives the size of the spread of readings. You might conclude that your corridor was (12.3±0.2)m long. In other words, your measurements are usually within 20cm of 12.3m. In this case the absolute uncertainty is 20cm.

The absolute uncertainty only gives part of the story. A 10cm error in the length of a curtain track implies sloppy work. A 10cm error in the total length of the M1 motorway is an impressive measurement. To make this clearer, we often state errors (or uncertainties) in percentage form – and this is called relative uncertainty. The relative uncertainty in the length of the corridor is

[29] A mathematician would probably be appalled at some of the statements I make. The study of errors and uncertainties is embedded in statistics, which is a well-established discipline. There are many refinements to the results I quote which are needed to satisfy the rigour of a professional statistician. However, the thing about uncertainties in measurements is that quoting them to more than one significant figure is missing the point, and therefore our methods only need to be accurate to this degree. If you are doing statistics and you want to take things more seriously, then you will understand (2) from the addition of variances; and you will realise that in 8.2.1 we really ought to be adding variances not errors. You will also appreciate that (2) ought to have an (n-1) in the denominator to take into account the difference between population and sample statistics, and that our section 8.2.2 is a form of the Binomial theorem to first order.

$$\text{Relative Error} = \frac{\text{Abs. Uncertainty}}{\text{Measurement}} = \frac{0.2\text{ m}}{12.3\text{ m}} = 1.6\% \approx 2\%. \qquad (1)$$

Notice the rounding off at the end. It is usually pointless to give uncertainties to more than one significant figure.

Every measurement has its uncertainty, and the only way of determining this is to take more than one measurement, and work out the standard deviation – to measure the spread. In practice the spread can be 'eyeballed' rather than calculated. If the measurements were 54.5cm, 54.7cm and 54.3cm, then there is no need to use a calculator and the technical definition of deviation. The observation that the spread is about ±0.2cm is perfectly good enough.

Notice that the more readings you take, the better idea you get of the spread of the measurements – and hence the better estimate you can make for the middle, which is indicative of 'true' value. Therefore we find, from statistics, that if you take *n* measurements, and the absolute uncertainty is *x*, then the uncertainty of the mean of those measurements is approximately:[30]

$$\text{Uncertainty of mean} = \frac{x}{\sqrt{n}}. \qquad (2)$$

Therefore, the more measurements you take, the more accurate the work. Notice that if you wish to halve the uncertainty, you need to take *four* times as many readings. This is subject to one proviso:

Measurements also have a resolution. This is the smallest distinguishable difference that the measuring device (including the technique) can detect. For a simple length measurement with a metre ruler, the resolution is probably 1mm. However if, by years of practice with a magnifying lens, you could divide millimetres into tenths by eye, you would have a resolution of 0.1mm using the same metre stick. That is why we say that the resolution depends on the technique as well as on the apparatus.

The uncertainty of a measurement can never be less than the resolution. This is the proviso we mentioned below equation (2). Why should this be the case? Let us have a parable.

Many years ago, the great nation of China had an emperor. The masses of the population were not permitted to see him. One day, a citizen had

[30] This result will be proved in any statistics textbook. To give a brief justification – the more readings you take, the more likely you are to have some high readings cancelling out some low readings when you take the average.

the sudden desire to know the length of the emperor's nose. He could not do this directly, since he was not permitted to visit the emperor. So, using the apparatus of the imperial administration, he asked all the regional mandarins to ask the entire population to make a guess. Each person would make some guess at the imperial nasal length – and the error of each guess would probably be no more than ±2cm – since nose lengths tend not to vary by more than about 4cm.

However, the mean would be a different matter. Averaged over the 1000 million measurements, the error in the mean would be 0.7μm. So the emperor's nose had been measured incredibly accurately – without a single observation having been made!

The moral of the story: uncertainties *are* reduced by repeated measurement, but the error can never be reduced below the resolution of the technique – here 2cm – since ignorance can not be circumvented by pooling it with more ignorance.

9.2 Errors, and how to make them worse

Errors are one thing. The trouble is that usually we want to put our measurements into a formula to calculate something else. For example, we might want to measure the strength of a magnetic field by measuring the force on a current-carrying wire $B = F/IL$.

If there is a 7% uncertainty in the current, 2% in the force and 1% in the length – what is the uncertainty in the magnetic field?

There are two rules you need:

9.2.1 Rule 1 – Adding or subtracting measurements

If two measurements are added or subtracted, the **absolute** uncertainty in the result equals the sum (never the difference) of the absolute uncertainties of the individual measurements.

Therefore if a car is (3.2±0.1)m long, and a caravan is (5.2±0.2)m long, the total length is (8.4±0.3)m long. Similarly if the height of a two-storey house is (8.3±0.2)m and the height of the ground floor is (3.1±0.1)m, the height of the upper floor is (5.2±0.3)m.

Even in the second case, we do not subtract the uncertainties, since there is nothing stopping one measurements being high, while the other is low.[31]

[31] Of course, there is a good chance that the errors will partly cancel out, and so our method of estimating the overall error is pessimistic. Nevertheless, this kind of error analysis is good enough for most experiments – after all it is better to *over*estimate your errors. If you want to do more careful analysis, then you work on the principle that if the absolute uncertainties in a

9.2.2 Rule 2 – Multiplying or dividing measurements

If two measurements are multiplied or divided, the **relative** uncertainty in the result equals the sum (never the difference) of the relative uncertainties of the individual measurements.

Therefore if the speed of a car is 30mph ± 10%, and the time for a journey is 6 hours ± 2%, the uncertainty in the distance travelled is 12%.

Notice that one consequence of this is that if a measurement, with relative uncertainty *p*% is squared (multiplied by itself), the relative error in the square is 2*p*% - i.e. doubled. Similarly if the error in measurement *L* is *p*%, the error in L^n is *p*×*n*%. Notice that while a square root will halve the relative error, an inverse square (*n*=–2) doubles it. All the minus sign does is to turn overestimates into underestimates. It does not reduce the magnitude of the relative error.[32]

Now we can answer our question about the magnetic field measurement at the beginning of the section. All three relative errors (in length, force and current) must be added to give the relative error in the magnetic field, which is therefore 10%.

9.3 *Systematic Errors*

All the 'errors' mentioned so far are called 'random', since we assume that the measurements will be clustered around the true value. However often an oversight in our technique will cause a measurement to be overestimated more often than underestimated or vice-versa. This kind of error is called 'systematic error', and can't be reduced by averaging readings. The only way of spotting this kind of error (which is a true error in that there is something wrong with the measurement) is to repeat the measurement using a completely different technique, and compare the results. Just thinking hard about the method can help you spot some

set of measurements are A, B, C…, then the absolute uncertainty in the sum (or in any of the differences) is given by $\sqrt{A^2+B^2+C^2\cdots}$. This result comes from statistics, where we find that the variance (the square of the standard deviation) of a sum is equal to the sum of the variances of the two measurements.

[32] The conclusions of this paragraph can be justified using calculus. If measurement *x* has absolute uncertainty δ*x*, and *y* (a function of *x*) is given by $y = Ax^n$, then we find that the relative error in *y* is given by:

$$\frac{\delta y}{y} \approx \frac{dy}{dx}\frac{\delta x}{y} = \left(An\,x^{n-1}\right)\frac{\delta x}{y} = n\frac{y}{x}\frac{\delta x}{y} = n\frac{\delta x}{x},$$

that is *n* multiplied by the relative error in *x*.

systematic errors, but it is still a good idea to perform the experiment a different way if time allows.

9.4 Which Graph?

You will often have to use graphs to check the functional form of relationships. You may also have to make measurements using the graph. In order to do either of these, you usually need to manipulate the data until you can plot a straight line. A straight line is conclusive proof that you have got the form of the formula right!

The gradient and y-intercept can then be read, and these enable other measurements to be made. For example, your aim may be to measure the acceleration due to gravity. You may plot velocity of falling against time, in which case you will need to find the gradient of the line.

At its most general, you will have a suspected functional form *y=f(x)*, and you will need to work out what is going on in the function *f*. Notice that our experiment will give us pairs of (*x,y*) values – what is not known are the parameters in the function *f*. We find them by manipulating the equation:

$$\begin{gathered} y = f(x) \\ \vdots \\ g(x,y) = A\,h(x,y) + B \end{gathered}.$$

We can then plot *g(x,y)* against *h(x,y)*, and obtain the parameters *A* and *B* from the gradient and intercept of the line. Furthermore, the presence of the straight line on the graph assures us that our function *f* was a good guess. We shall now look at the most common examples.

9.4.1 Exponential growth or decay

Here we have the functional form $y = Ae^{Bx}$, where *A* and *B* need to be determined. We manipulate the equation:

$$\begin{gathered} y = Ae^{Bx} \\ \ln y = \ln A + Bx \end{gathered}.$$

So we plot (ln *y*) on the vertical axis, and (*x*) on the horizontal. The y-intercept gives ln *A*, and the gradient gives *B*.

9.4.2 Logarithmic growth or decay

Here we have the functional form $y = A + B\ln x$, and again, we need to work out the values of *A* and *B*. This equation is already in linear form – we plot *y* on the vertical, and (ln *x*) on the horizontal. The y-intercept gives *A*, and the gradient gives *B*.

9.4.3 Power laws

This covers all equations with unknown powers: $y = Ax^B$. The manipulation involves logarithms:

$$
\begin{aligned}
y &= Ax^B \\
\ln y &= \ln A + \ln x^B \\
\ln y &= \ln A + B \ln x
\end{aligned} .
$$

Here we plot (ln *y*) against (ln *x*), and find the power (*B*) as the gradient of the line. The *A* value can be inferred from the y-intercept, which is equal to ln *A*.

9.4.4 Other forms

Even hideous looking equations can be reduced to straight lines if you crack the whip hard enough. How about $y = A\sqrt{x} + Bx^3$? Is it tasty enough for your breakfast? Actually it's fine if digested slowly:

$$
\begin{aligned}
y &= A\sqrt{x} + Bx^3 \\
\frac{y}{\sqrt{x}} &= A + Bx^{5/2}
\end{aligned} .
$$

This looks even worse, doesn't it? But remember that it is *x* and *y* that are *known*. If we plot ($y/\sqrt{x}$) on the vertical, and ($x^{5/2}$) on the horizontal, a straight line appears, and we can read *A* and *B* from the y-intercept and gradient respectively.

9.5 *Questions*

1. Work out the relative uncertainty when a 5V battery is measured to the nearest 0.2V.

2. If I don't want to have to correct my watch more than once a week, and I never want my watch to be more than 1s from the correct time, calculate the necessary maximum relative uncertainty of the electronic oscillator which I can tolerate.

3. My two-storey house is 7.05±0.02m tall. The ground floor is 3.2±0.01m tall. How tall is the first floor?

4. I want to measure the resistance of a resistor. My voltmeter can read up to 5V, with an absolute uncertainty of 0.1V. My ammeter can read up to 1A with an absolute uncertainty of 0.02A. Assuming that my resistor is approximately 10Ω, calculate the absolute uncertainty of the resistance I measure using the formula R=V/I. Assume that I choose the current to make the relative uncertainty as small as possible.

10 Appendix

10.1 *Multiplying Vectors*

Physics is riddled with quantities which have both magnitude and direction – velocity, acceleration, displacement, force, momentum, angular velocity, torque, and electric field to name but eight. When describing these, it is very useful to use vector notation. At best this saves us writing out separate equations for each of the components. While the addition and subtraction of vectors is reasonably straightforward (you add, or subtract, the components to get the components of the result), multiplication is more tricky.

You can think of a vector as a little arrow. You can add them by stacking them nose-to-tail, or subtract them by stacking them nose-to-nose. But how do you go about multiplying them? It is not obvious!

To cut a long story short, you can't do it unambiguously. However there are two vector operators which involve multiplication and are useful in physics. Ordinary multiplication is commonly written with either the cross (×) or dot (●), so when it comes to vectors we call our two different 'multiplication' processes the dot product and cross product to distinguish them. These are the closest we get to performing multiplication with vectors.

10.1.1 The Dot product (or scalar product)

A ton of bricks is lifted a metre, then moved horizontally by 2m. How much work is done? Work is given by the product of force and distance, however only the vertical lifting (not the horizontal shuffling) involves work. In this case the work is equal to the weight (about 9.8kN) multiplied by the vertical distance (1m).

This gives us one useful way of 'multiplying' vectors – namely to multiply the magnitude of the first, by the component of the second which is parallel to the first.

If the two vectors are **A** and **B**, with magnitudes *A* and *B*, and with an angle θ between them, then the component of **B** parallel to **A** is $B\cos\theta$. Therefore the dot product is given by $AB\cos\theta$.

$$\mathbf{A} \bullet \mathbf{B} \equiv AB\cos\theta \qquad (1)$$

Notice that the dot product of two vectors is itself a *scalar*. Note that when we talk about the square of a vector, we mean its scalar product with itself. Since in this case, θ=0, this is the same as the square of the vector's magnitude.

The dot product is also commutative, in other words, the order of the two vectors **A** and **B** does not matter, since **A•B**=**B•A**.

The dot product of two vectors written using Cartesian co-ordinates is particularly easy to calculate. If we use **i**, **j**, and **k** to represent the unit vectors pointing along the *+x*, *+y* and *+z* axes, then

$$\begin{aligned}(a\mathbf{i}+b\mathbf{j}+c\mathbf{k})\bullet(u\mathbf{i}+v\mathbf{j}+w\mathbf{k}) &= au\mathbf{i}\bullet\mathbf{i}+av\mathbf{i}\bullet\mathbf{j}+aw\mathbf{i}\bullet\mathbf{k}\\ &\quad +bu\mathbf{j}\bullet\mathbf{i}+bv\mathbf{j}\bullet\mathbf{j}+bw\mathbf{j}\bullet\mathbf{k}\\ &\quad +cu\mathbf{k}\bullet\mathbf{i}+cv\mathbf{k}\bullet\mathbf{j}+cw\mathbf{k}\bullet\mathbf{k}\\ &= au+bv+cw\end{aligned} \tag{2}$$

10.1.2 The Cross product (or vector product)

If the dot product produced a scalar, what are we to do if a *vector* is needed as the result of our multiplication? Answer: a cross product.

Our first dilemma is to choose the direction of the result. Given that the vectors will, in general, not be parallel or antiparallel, we can't choose the direction of one of them – that would not be fair! The two vectors will usually define a plane, so perhaps we could use a vector in this plane as the result? No, that wouldn't do either – there is still an infinite number of directions to choose from! A solution is presented if we choose the vector perpendicular to this plane. This narrows the choice down to two directions – and we use a convention to choose which.

Notice that the result of the cross product must be zero if the two vectors are parallel, since in this case we can't define a plane using the vectors. It follows that the cross product of a vector with itself is zero. This means that we aren't going to be interested in the component of the second vector which is parallel to the first when calculating the product. On the contrary, it is the perpendicular component which matters.

The cross product of two vectors is defined as the magnitude of the first, multiplied by the component of the second which is perpendicular to the first. The product is directed perpendicular to both vectors. To be more precise, imagine a screw attached to the first vector. The cross product goes in the direction the screw advances when the first vector is twisted to line up with the second. The cross product of a vector lying along the *+x* axis with one lying along the *+y* axis is one lying along the *+z* axis. The cross product of 'up' with 'forwards' is 'left'.

$$|\mathbf{A}\times\mathbf{B}| \equiv AB\sin\theta \tag{3}$$

With a definition as obtuse as this, you could be forgiven for wondering whether it had any practical use at all! However they turn out to be very useful in physics – especially when dealing with magnetic fields and rotational motion.

Notice that, where the dot product was commutative, the cross product is anticommutative. In other words, **A×B=–B×A**, so make sure you don't swap the vectors over inadvertently.

The vector product of vectors written in Cartesian form can also be calculated:

$$
\begin{aligned}
(a\mathbf{i}+b\mathbf{j}+c\mathbf{k})\times(u\mathbf{i}+v\mathbf{j}+w\mathbf{k}) &= au\mathbf{i}\times\mathbf{i}+av\mathbf{i}\times\mathbf{j}+aw\mathbf{i}\times\mathbf{k} \\
&\quad +bu\mathbf{j}\times\mathbf{i}+bv\mathbf{j}\times\mathbf{j}+bw\mathbf{j}\times\mathbf{k} \\
&\quad +cu\mathbf{k}\times\mathbf{i}+cv\mathbf{k}\times\mathbf{j}+cw\mathbf{k}\times\mathbf{k} \\
&= 0+av\mathbf{k}-aw\mathbf{j} \\
&\quad -bu\mathbf{k}+0+bw\mathbf{i} \\
&\quad +cu\mathbf{j}-cv\mathbf{i}+0 \\
&= (bw-cv)\mathbf{i}+(cu-aw)\mathbf{j}+(av-bu)\mathbf{k} \\
&= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a & b & c \\ u & v & w \end{vmatrix}
\end{aligned}
\tag{4}
$$

where the most convenient way of remembering the result is as the determinant of the 3×3 matrix shown.

10.2 *Dimensional Analysis*

If you look back at the ‘flow equation’ (24 in section 1.3.3), you will see something interesting about the units.

Current (A) = Charge density (C/m^3) × Area (m^2) × Speed (m/s)

If we ‘do algebra’ with the units on the right hand side, we get

$$\frac{C}{m^3} \times m^2 \times \frac{m}{s} = \frac{C}{s} = A \,,$$

and this agrees with the units of the left hand side. Now this may all seem pretty obvious, but it gives us a useful procedure for checking whether our working is along the right lines. If, during your calculations, you find yourself adding a charge of 3C to a distance of 6m to get a result of 9N; or you multiply a speed of 13m/s by a time of 40s and get a current of 520A; then in either case you must have made a mistake!

We can also use the principle that units must balance to guess the form of an equation we do not know how to derive. For example, you may guess that the time period of a simple pendulum might depend on the length of the pendulum *L*, the strength of the local gravity *g* and the mass of the pendulum bob *m*. Now

- *L* is measured in m
- *g* is measured in N/kg or m/s^2
- *m* is measured in kg,
- and we want a time period, which will be measured in s.

The only way these measurements can be combined to make something in seconds is to take *L*, divide it by *g* (this gives something in s^2) and then take the square root. Therefore, without knowing any physics of the simple harmonic oscillator, we have shown that the time period of a pendulum is related to $\sqrt{L/g}$ and will be independent of the mass *m*.

Similarly, notice what happens if you multiply ohms by farads:

$$\Omega F = \frac{V}{A} \times \frac{C}{V} = \frac{C}{A} = s \,.$$

Yes, you get seconds. Therefore, it should come as little surprise to you that if you double the resistance of a capacitor-resistor network, it will take twice as long to charge or discharge. Furthermore, you have worked this out without recourse to calculus or the tedious electrical details of section 6.1.2.2.

Of course, one drawback of the method presented here is that some quantities have two different units. For example, gravitational field strength could be N/kg or m/s^2. Electric field strength could be N/C or V/m. How do you know which to choose? The answer is that if you restrict yourself to using the minimum number of units in your working, and express all others in terms of them, you will not have any difficulties. Usually people choose m, s, kg and A but any other combination of independent units[33] will do equally well.[34]

In books you may see folk use L, T, M, I and Θ to represent the 'dimensions' of length, time, mass, electric current and temperature. This is just a more formal way of doing what we have done here using the S.I. units. In these books, the dimension of speed would be written as

$$[\text{speed}] = \text{L T}^{-1},$$

and the dimensions of force would be written

$$[\text{force}] = \text{M L T}^{-2},$$

where the square brackets mean 'dimensions of'. Technically, this is more correct than using the S.I. units, because some quantities are dimensionally the same, but have very different meanings (and hence units). For example, torque and energy have the same dimensions, but you wouldn't want to risk confusing them by using the same unit for both. Similarly an angle in radians has no dimensions at all (being a an arc length in metres divided by a radius in metres), but we wouldn't want to confuse it with an ordinary number like 3.[35]

[33] By independent we mean that no one unit can be derived entirely from a combination of the others. For example, m, s, kg and J would be no good as a set of four since J can already be expressed in terms of the others J = kg × (m/s)^2, and hence we have ambiguity arising as to how we express quantities.

[34] OK, if you want to do work where there are electric currents and temperatures as well as mechanical quantities, you might need to go up to five (an extra one for temperature).

[35] Indeed, quantities with 'no units' are usually said to have the dimensions of the number one. Thus [angle] = 1. It follows that angular velocity has dimensions of [angle]÷[time] = 1÷T = T^{-1}. This comes from the property 1 has in being the 'unity' operator for multiplication.