Physics

for the IB DIPLOMA

John Allum
Christopher Talbot

FREE ONLINE RESOURCES

HODDER EDUCATION

Dear Student

The Option chapters and extra chapters on:

- Teaching and learning IB Physics
- Theory of knowledge
- Extended essay
- Internal assessment
- Data analysis
- Preparing for the IB Physics exam

are available online at **www.hodderplus.co.uk/ibphysics**

You will need the following user name and password to access the online materials:

User name: ibphysics

Password: electromagnetic

Please note that the user name and password are both case sensitive.

Physics
for the IB DIPLOMA

John Allum
Christopher Talbot

HODDER EDUCATION
AN HACHETTE UK COMPANY

Orders: please contact Bookpoint Ltd, 130 Milton Park, Abingdon, Oxon OX14 4SB. Telephone: (44) 01235 827720. Fax: (44) 01235 400454. Lines are open 9.00 – 5.00, Monday to Saturday, with a 24-hour message answering service. Visit our website at www.hoddereducation.com

First published in 2012 by
Hodder Education,
An Hachette UK Company
338 Euston Road
London NW1 3BH

Impression number 3 2 1
Year 2014 2013 2012

Cover photo © Eliza Snow/istockphoto.com
Illustrations by Pantek Media, Maidstone, Kent
Typeset in 10pt Goudy Old Style by Pantek Media, Maidstone, Kent
Printed in Italy

A catalogue record for this title is available from the British Library

ISBN: 978 1444 146 523

Contents

Section 2: Additional higher level

Section 3: Options available on the website accompanying this book

Introduction

The International Baccalaureate Diploma programme, a pre-university course for 16- to 19-year-olds, is designed to develop not only a breadth of knowledge, skills and understanding, but well-rounded individuals and engaged world citizens. One of the diploma's key principles is the requirement to study in six different academic areas, at least one of which must be an experimental science (Group 4). Of these sciences, physics, whether taken at Standard or Higher Level, is the choice of many students. This book is designed to guide these students.

Within the IB Diploma programme, the theory content for physics is organized into compulsory core topics plus a number of options. The organization of this resource exactly follows that syllabus sequence:

- Section 1, Chapters 1–8, is the common core topics for Standard and Higher Level students.
- Section 2, Chapters 9–14, is the additional topics for Higher Level students.
- Section 3 consists of the most popular Options, A, B, C, E and G. (Options A, B and C are for Standard Level students; Options E and G are for both Standard Level and Higher Level students.)

Each of the core topics is the subject of a corresponding single chapter in the *Physics for the IB Diploma* printed book. The option chapters are can be found on the website; details of how to access the website can be found on the inside front cover of the book.

Special features of the chapters of *Physics for the IB Diploma* are described below.

- Each chapter begins with **Starting points** that summarize the knowledge assumed of the reader before beginning the chapter.
- The text is written in **straightforward language**, without phrases or idioms that might confuse students for whom English is a second language. The depth of treatment of topics has been carefully planned to accurately reflect the objectives of the IB syllabus and the requirements of the examinations.
- The IB Diploma physics syllabus **assessment statements** being covered in each section of text are shown clearly, so that the links between the text and the syllabus are easily understood.
- The bar at the foot of each page is colour coded to show the **syllabus level**: Standard Level (pink bar); Higher Level (dark red bar); or both (striped bar).
- **Worked examples** are provided in each chapter whenever new equations are introduced. Numerous **self-assessment questions** are placed throughout the chapters close to the relevant theory. **Answers** to most questions are provided at the end of the book.
- Some typical **IB examination style questions** are provided at the end of each chapter as well as a selection of past IB examination questions.
- The **Additional perspectives** sections are designed to take students beyond the limits of the IB syllabus in a variety of ways. They may, for example, provide a historical context, extend theory or offer an interesting application. They are often accompanied by more challenging, or research style, questions. They do *not* contain any knowledge which is essential for the IB examinations.
- Links to the interdisciplinary **Theory of Knowledge** element of the IB Diploma course are made in all chapters.
- A **Summary** at the end of each chapter lists the key points required by the IB syllabus.
- A comprehensive **Glossary** of words and terms is included both in the printed book and on the website.
- There are additional short chapters offering physics-specific advice on **Theory of Knowledge** and **Extended Essay** assignments, as well as on the **Internal Assessment** of practical work. Other chapters cover the skills necessary for the **drawing and interpreting graphs**, as well as advice about **answering examination questions** and understanding **command terms**. These chapters can be found on the website.

Using this book

The sequence of chapters in *Physics for the IB Diploma* deliberately follows the sequence of the syllabus contents. However, the IB Diploma physics syllabus is not designed as a teaching syllabus, so the order in which the syllabus content is presented is not necessarily the order in which it will be taught. Different schools and colleges should design a course based on their individual circumstances.

In addition to the study of the physics principles contained in this book, IB science students carry out experiments and investigations, as well as collaborating in a Group 4 Project. These are assessed within the school (Internal Assessment) based on well-established criteria. For the benefit of students and teachers who may be new to the IB physics course, a chapter on the website explains how all the various aspects of the IB physics course are integrated together. This chapter is based on material originally written for *Biology for the IB Diploma* by guest author Gary Seston, an experienced and enthusiastic teacher of the IB Diploma who also has examiner experience.

The contents of Chapter 1 (Physics and physical measurement) have applications that recur throughout the rest of the book and also during practical work. For this reason, it is intended more as a source of reference, than as material that should be fully understood before progressing to the rest of the course.

Author profiles

John Allum

John has been teaching pre-university physics courses as a Head of Department in a variety of international schools for more than 30 years. He has taught IB Physics in Malaysia and in Abu Dhabi, and he has been an examiner for IB Physics for many years.

Christopher Talbot

Chris taught IB Chemistry, IB Biology and IGCSE Physics at the Overseas Family School, Republic of Singapore. He has previously been Head of Science and a Theory of Knowledge Coordinator at the Overseas Family School, and has prepared students for the International Chemistry Olympiad.

Author's acknowledgements

We are indebted to the following teachers and lecturers who reviewed early drafts of the chapters: Professor Robert Smith, University of Sussex (Astrophysics); Dr Tim Brown, University of Surrey (Communications and Digital Technology); Dr David Cooper (Quantum Physics); Mr Bernard Taylor (Theory of Knowledge, Internal Assessment and Fields and Forces); Professor Christopher Hammond, University of Leeds (Electromagnetic Waves); Professor Phil Walker, University of Surrey (Nuclear Physics); Dr David Jenkins, University of York (Nuclear Physics) and Trevor Wilson, Enka Schools, Istanbul, Turkey.

We also like to thank David Talbot, who supplied some of the photographs for the book, and Terri Harwood and Jon Homewood, who drew a number of physicists. Chris Talbot would like to thank Cesar Reyes (Overseas Family School) for his invaluable assistance with the production of his chapters and Jenelyn Ramos for proofreading. John Allum would like to thank Li Hongmei for her invaluable help with proofreading.

Finally, we would like to express our gratitude for the tireless efforts of the Hodder Education team that produced the book you have in front of you, led by Nina Konrad and Eleanor Miles, together with freelancers Anne Trevillion and Anna Clark.

1 Physics and physical measurement

STARTING POINTS

- Experimental observations and measurements provide the evidence for almost all advances in scientific understanding of the world and the universe around us.
- A measurement is described as being accurate if it is the same as the 'true' value; however, in many cases the 'true' value may not be known.
- No measurements made in any experiment can ever be completely accurate.

1.1 The realm of physics

Physics is the fundamental science that tries to explain how and why everything in the universe behaves in the way that it does. Physicists study everything from the smallest parts of atoms to distant objects in our galaxy and beyond (Figure 1.1). Physicists use basic concepts like force and energy to explain the world in which we live.

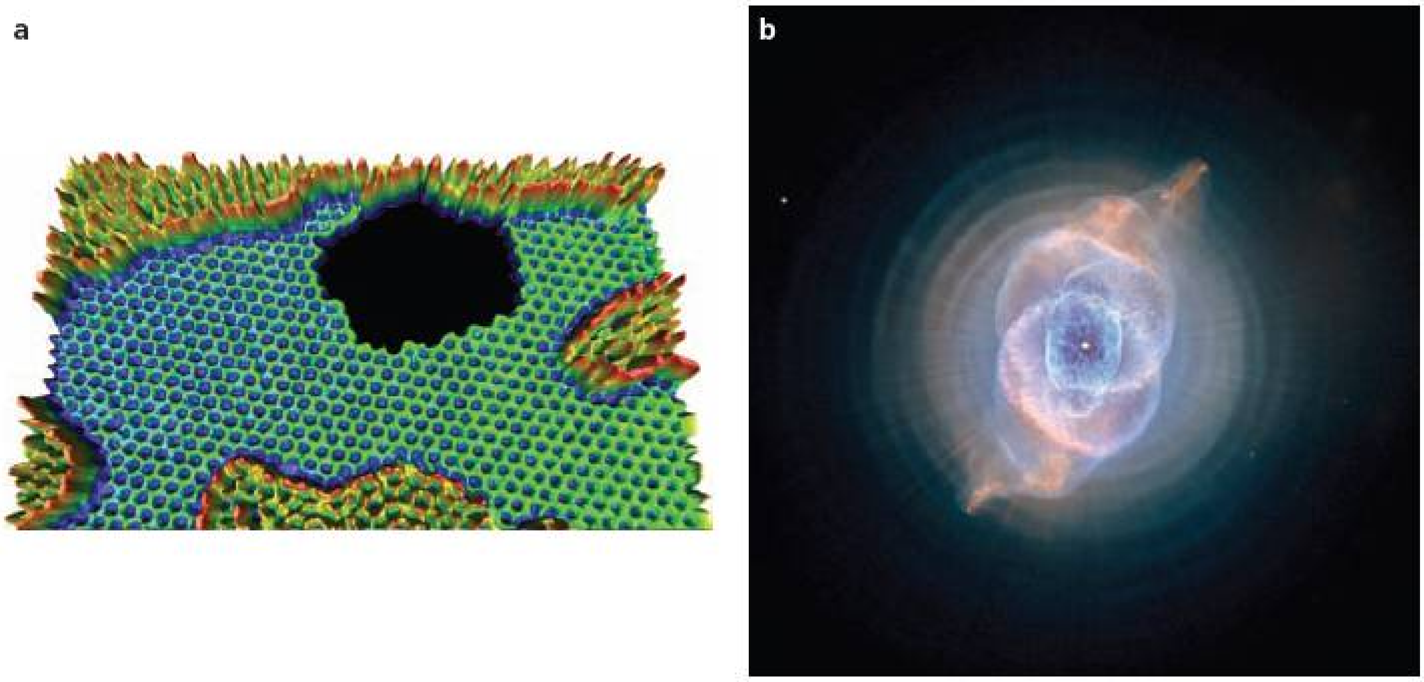


Figure 1.1 a The behaviour of individual atoms in graphene (a material made from a single layer of carbon atoms) can be seen using a special type of electron microscope **b** Complex gas and dust clouds in the Cat's Eye Nebula, 3000 light years away

Range and magnitudes of quantities in our universe

1.1.1 **State** and **compare** quantities to the nearest order of magnitude.
1.1.2 **State** the ranges of magnitude of distances, masses and times that occur in the universe, from smallest to greatest.
1.1.3 **State** ratios of quantities as differences of orders of magnitude.

Physics is a **quantitative** subject that makes much use of mathematics. Measurements and calculations commonly relate to the world that we can see around us (the **macroscopic** world), but our observations may require **microscopic** explanations, often including an understanding of molecules, atoms and subatomic particles. **Astronomy** is a branch of physics that deals with the other extreme – quantities that are very much greater than anything we experience in everyday life.

The study of physics therefore involves dealing with both very large and very small numbers. When numbers are so different from our everyday experiences, it can be difficult to appreciate their true size. For example, the age of the universe is believed to be about 10^{18} s, but just how big is that number? The only sensible way to answer that question is to compare the quantity with something else. For example, the age of the universe is about one hundred million human lifetimes (see below).

When comparing quantities of very different sizes (**magnitudes**), for simplicity we often make approximations to the nearest power of 10. When numbers are approximated and quoted to the nearest power of 10, it is called giving them an **order of magnitude**. For example, when comparing the lifetime of a human (worldwide average is about 70 years) to the age of the universe (1.4×10^{10} y), we can use the approximate ratio $10^{10}/10^2$. That is, the age of the universe is about 10^8 human lifetimes, or we could say that there are eight orders of magnitude between them.

Here are three further examples:

- The mass of a hydrogen atom is 1.67×10^{-27} kg. To an order of magnitude this is 10^{-27} kg.
- The distance to the nearest star (*Proxima Centauri*) is 4.01×10^{16} m. To an order of magnitude this is 10^{17} m. (Note: log of $4.01 \times 10^{16} = 16.60$, which is nearer to 17 than to 16)
- There are 86 400 seconds in a day. To an order of magnitude this is 10^5 s.

Tables 1.1 to 1.3 list the ranges of mass, distance and time that occur in the universe. Students are recommended to look at computer simulations representing these ranges.

Table 1.1 The range of masses in the universe

Object	Mass/kg
the observable universe	10^{53}
our galaxy (the Milky Way)	10^{42}
the Sun	10^{30}
the Earth	10^{24}
a large passenger plane	10^5
a large adult human	10^2

Object	Mass/kg
a large book	1
a raindrop	10^{-6}
a virus	10^{-20}
a hydrogen atom	10^{-27}
an electron	10^{-30}

Table 1.2 The range of distances in the universe

Distance	Size/m
distance to edge of visible universe	10^{27}
diameter of our galaxy (the Milky Way)	10^{21}
distance to nearest star	10^{16}
distance to the Sun	10^{11}
distance to the Moon	10^8
radius of the Earth	10^7
height of a cruising plane	10^4
height of a child	1
how much human hair grows in one day	10^{-4}
diameter of an atom	10^{-10}
diameter of a nucleus	10^{-15}

Table 1.3 The range of times in the universe

Time period	Time interval/s
age of the universe	10^{18}
time since dinosaurs became extinct	10^{15}
time since humans first appeared on Earth	10^{13}
time since the pyramids were built in Egypt	10^{11}
typical human lifetime	10^9
one day	10^5
time between human heartbeats	1
time period of high-frequency sound	10^{-4}
time for light to travel across a room	10^{-8}
time period of oscillation of a light wave	10^{-15}
time for light to travel across a nucleus	10^{-23}

Making estimates

1.1.4 Estimate approximate values of everyday quantities to one or two significant figures and/or to the nearest order of magnitude.

Sometimes we do not have the data needed for accurate calculations, or maybe the calculations need to be made quickly. Sometimes a question is so vague that an accurate answer is simply not possible. The ability to make sensible estimates is a very useful skill that needs plenty of practice. The worked example and questions 1–4 below are typical of calculations that do not have precise answers.

When making estimates, different people will produce different answers and it is sensible to use only one (maybe two) significant figures. Sometimes only an order of magnitude is needed. (There is more about precision and significant figures later in this chapter.)

Worked example

1 Estimate the mass of air in a classroom. (Density of air is $1.3\,\text{kg}\,\text{m}^{-3}$.)

A typical classroom may have dimensions of $7\,\text{m} \times 8\,\text{m} \times 3\,\text{m}$, so that its volume is about $170\,\text{m}^3$.

$$\text{mass} = \text{density} \times \text{volume} = 170 \times 1.3 = 220\,\text{kg}$$

Since this is an estimate, an answer of $200\,\text{m}^3$ may be more appropriate. To an order of magnitude it would be $10^2\,\text{kg}$.

1 Estimate the mass of:
 a a page of a book
 b air in a bottle
 c a dog
 d water in the oceans of the world.

2 Give an estimate for each of the following:
 a the height of a house with three floors
 b how many times a wheel on a car rotates during the lifetime of the car
 c how many grains of sand would fill a cup
 d the thickness of a page in a book.

3 Estimate the following periods of time:
 a how many seconds there are in an average human lifetime
 b how long it would take a person to walk around the Earth (ignore the time not spent walking)
 c how long it takes for light to travel across a room.

4 Research the relevant data so that you can compare the following measurements. (Give your answer as an order of magnitude.)
 a the distance to the Moon to the circumference of the Earth
 b the mass of the Earth to the mass of an apple
 c the time it takes light to travel 1 m to the time between your heartbeats.

1.2 Measurement and uncertainties

The SI system of fundamental and derived units

To talk to each other we need to share a common language, and to share numerical information we need to use common **units of measurement**. An internationally agreed system of units is used by scientists around the world. It is called the **SI system** (from the French 'Système International').

The fundamental units of measurement

1.2.1 State the fundamental units in the SI system.

There are seven **fundamental (basic)** units in the SI system: kilogram, metre, second, ampere, mole, kelvin (and candela, which is not part of this course). The quantities, names and symbols for these fundamental SI units are given in Table 1.4.

They are called 'fundamental' because their definitions are not combinations of other units (unlike metres per second, for example). You do not need to learn the definitions of these units, but they are included for interest in the Additional Perspectives box on page 4.

Table 1.4 Fundamental units

Quantity	Name	Symbol
length	metre	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
temperature	kelvin	K
amount of substance	mole	mol

Additional Perspectives

Definitions of fundamental units

- A **kilogram, kg**, is the mass of a cylinder of platinum–iridium alloy kept at the International Bureau of Weights and Measures in France.

Figure 1.2 Replicas of the International Prototype Kilogram (IPK) are used across the world to acts as mass standards. They are stored in air, in nested bell jars, and need periodic cleaning to remove atmospheric contamination

- A **metre, m**, is the distance travelled by light in a vacuum in 1/299 792 458 seconds. Previously it was considered to be one ten millionth the distance between the equator and the North Pole.
- A **second, s**, is the duration of 9 192 631 770 oscillations of the electromagnetic radiation emitted in the transmission between two specific energy levels in caesium-133 atoms. Previously a second was defined as 1/84 600 of a day.
- An **ampere, A**, is that current which, when flowing in two parallel conductors one metre apart in a vacuum, produces a force of 2×10^{-7} N on each metre of the conductors.
- A **kelvin, K**, is 1/273.16 of the thermodynamic temperature of the triple point of water.
- A **mole, mol**, is an amount of substance that contains as many particles as there are atoms in 12 g of carbon-12.

Questions

1 Suggest possible reasons why the modern definitions of the metre and the second are different from their original definitions.

2 Can you suggest a better way of defining mass, rather than by comparing it to a piece of metal in France?

Derived units of measurement

1.2.2 **Distinguish** between fundamental and derived units and give examples of derived units.
1.2.3 **Convert** between different units of quantities.
1.2.4 **State** units in the accepted SI format.

Table 1.5 Some named derived units

Derived unit	Quantity	Combined fundamental units
newton (N)	force	$\mathrm{kg\,m\,s^{-2}}$
pascal (Pa)	pressure	$\mathrm{kg\,m^{-1}\,s^{-2}}$
hertz (Hz)	frequency	$\mathrm{s^{-1}}$
joule (J)	energy	$\mathrm{kg\,m^2\,s^{-2}}$
watt (W)	power	$\mathrm{kg\,m^2\,s^{-3}}$
coulomb (C)	charge	As
volt (V)	potential difference	$\mathrm{kg\,m^2\,s^{-3}\,A^{-1}}$
ohm (Ω)	resistance	$\mathrm{kg\,m^2\,s^{-3}\,A^{-2}}$
weber (Wb)	magnetic flux	$\mathrm{kg\,m^2\,s^{-2}\,A^{-1}}$
tesla (T)	magnetic field strength	$\mathrm{kg\,s^{-2}\,A^{-1}}$
becquerel (Bq)	radioactivity	$\mathrm{s^{-1}}$

All other units in science are combinations of the fundamental units. For example, the unit for volume is $\mathrm{m^3}$ and the unit for speed is $\mathrm{m\,s^{-1}}$. Combinations of fundamental units are known as **derived** units.

Sometimes derived units are also given their own name (Table 1.5). For example, the unit of force is $\mathrm{kg\,m\,s^{-2}}$, but it is usually called the newton, N. All derived units will be introduced and defined when they are needed during the course.

Note that you will be expected to write and recognize units using superscripts in the format $\mathrm{m\,s^{-1}}$, rather than as m/s. For example, acceleration has the unit $\mathrm{m\,s^{-2}}$.

Occasionally physicists use units that are not part of the SI system. For example, the electronvolt, eV, is a conveniently small unit of energy that is often used in atomic physics. Units such as this will be introduced when necessary during the course. You will be expected to be able to convert from one unit to another. A more common conversion would be changing time in years to time in seconds.

TOK Link: Fundamentals

As well as some units of measurement, many of the ideas and principles used in physics can be described as being *fundamental*. Indeed, physics itself is often described as the fundamental science. But, what exactly do we mean when we describe something as fundamental? We may replace the word with 'elementary' or 'basic', but that does not really help us to understand the true meaning of the word.

One of the central themes of physics is the search for *fundamental particles* – particles that are the basic building blocks of the universe and are not, themselves, made up of smaller and simpler particles. It is the same with *fundamental laws and principles*: a physics principle cannot be described as fundamental if it can be explained by 'simpler' ideas. Most scientists also believe that a principle cannot be really fundamental unless it is relatively simple to express (probably with mathematics). If it is complicated, maybe the underlying simplicity has not yet been discovered.

Fundamental principles must also be 'true' everywhere and for all time. The fundamental principles of physics that we use today have been tested, re-tested and re-tested again to check if they are truly fundamental. Of course, there is always a possibility that a principle believed to be fundamental may, in the future, be discovered to be explained by simpler ideas.

Consider two well known *laws* in physics. *Hooke's law* describes how some materials stretch when forces act on them. It is a simple law, but it is not a fundamental law because it is certainly not always true. The *law of conservation of energy* is also simple, but it is described as fundamental because there are no known exceptions.

Question

1 We use many constants in physics and mathematics. Which constants can be described as fundamental?

Scientific notation

1.2.5 **State** values in scientific notation and in multiples of units with appropriate prefixes.

When writing and comparing very large or very small numbers it is convenient to use **scientific notation** (sometimes called standard form).

In scientific notation every number is expressed in the following form: $a \times 10^b$, where a is a decimal number larger than one and less than ten, and b is a whole number (**integer**), called the **exponent**.

Table 1.6 Standard prefixes

Power	Prefix	Symbol
10^{-12}	pico-	p
10^{-9}	nano-	n
10^{-6}	micro-	μ
10^{-3}	milli-	m
10^{-2}	centi-	c
10^{-1}	deci-	d
10^{3}	kilo-	k
10^{6}	mega-	M
10^{9}	giga-	G
10^{12}	tera-	T

For example, in scientific notation the number 434 is written as 4.34×10^2; similarly, 0.000 316 is written as 3.16×10^{-4}.

Standard notation is useful in making the number of significant figures clear (see page 9). Scientific notation is also used for entering and displaying large and small numbers in calculators. The letter E is sometimes used on calculators to represent 'times ten to the power of …'. For example, 4.62E3 represents 4.62×10^3, or 4620.

Standard prefixes

In everyday language we use the words thousand and million to help represent large numbers. The scientific equivalents are the prefixes kilo- and mega-. For example, a kilowatt is one thousand watts, and a megajoule is one million joules. Similarly, a thousandth and a millionth are represented scientifically by the prefixes milli- and micro-. A list of standard prefixes is shown in Table 1.6. A full list is provided in the IB *Physics data booklet*.

Uncertainty and error in measurement

Errors in measurement

1.2.6 **Describe** and **give examples** of random and systematic errors.

An **error** occurs in a measurement when it is not exactly the same as the correct ('true') value. *All* measurements involve errors, whether they are large or small, for which there are many possible reasons, but they should not be confused with mistakes. Errors can be described as either *random* or *systematic*, although all measurements involve both kinds of error to some extent.

Random errors

Random errors cannot be avoided because exact measurements are not possible. Readings may be bigger or smaller than the correct value and they are scattered randomly around that value. There are many possible reasons for random errors, including for example:

- limitations of the scale or display being used
- reading scales from wrong positions
- irregular human reaction times when using a stopwatch
- difficulty in making an observation which changes quickly with time.

The reading obtained from a measuring instrument is limited by the smallest **division** of its scale. This is sometimes called a **readability (or reading) error**. For example, a liquid-in-glass thermometer with a scale marked in degrees (23 °C, 24 °C, 25 °C, etc.) cannot reliably be used to measure to every 0.1 °C. It is usually assumed that the error for **analogue** (continuous) scales, like a liquid-in-glass thermometer, is half of the smallest division, in this example ±0.5 °C. For **digital** instruments the error is assumed to be the smallest division that the meter can display. Figure 1.3 shows analogue and digital ammeters which can be used for measuring electric current.

Figure 1.3 Analogue and digital ammeters

A common reason for random errors is reading an analogue scale from an incorrect position. This is called a **parallax error** and an example is shown in Figure 1.4.

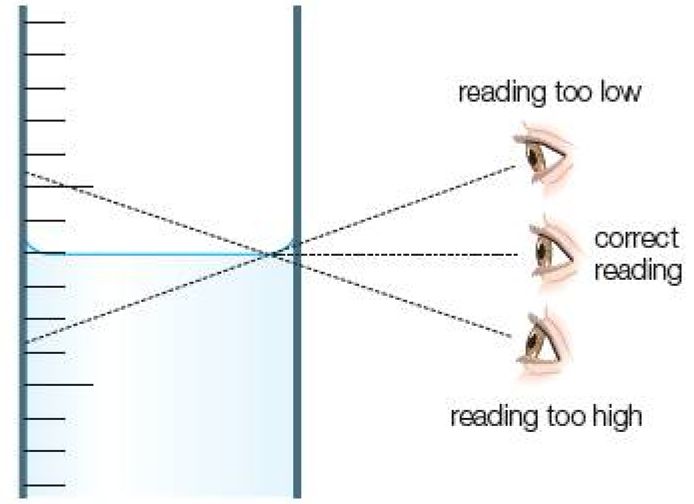


Figure 1.4 Parallax error when reading the level of liquid in a measuring cylinder

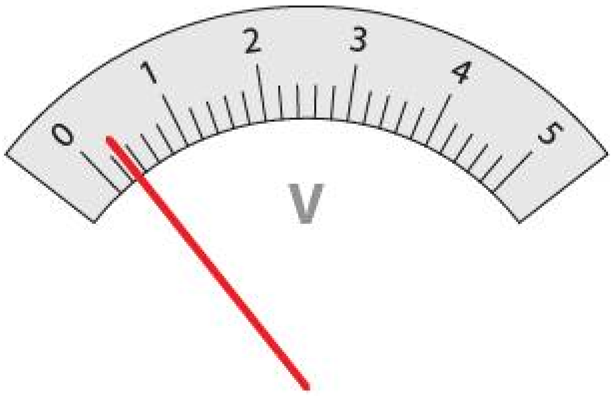


Figure 1.5 This voltmeter has a zero offset error of 0.3V, so the reading will always be too large by this amount

Systematic errors

A **systematic error** occurs because there is something *consistently* wrong with the measuring instrument or the method used. A reading with a systematic error is always either bigger or smaller than the correct value by the same amount. Common causes are instruments that have an incorrect scale (wrongly **calibrated**), or instruments that have an incorrect value to begin with, such as a meter that displays a reading when it should read zero. This is called a **zero offset error** and an example is shown in Figure 1.5. A thermometer that incorrectly records room temperature will produce systematic errors when used to measure other temperatures.

Precision and accuracy

1.2.7 **Distinguish** between precision and accuracy.

In everyday conversation we would consider measurements to be accurate if they were close to the correct value, but in physics the word **accurate** means that the measurements have a small systematic error. This means that accurate measurements are approximately evenly distributed around the correct value (whether they are close to it or not).

In many experiments the 'correct' result may not be known, which means that the accuracy of measurements cannot be known with any certainty. In such cases, the quality of the results can best be judged by whether values can be repeated in further measurements.

Measurements can be described as **precise** if we are confident that similar values will be obtained if they are repeated.

The difference between precise and accurate can be illustrated by considering arrows fired at a target, as in Figure 1.6. The aim is precise if the arrows are grouped close together and accurate if the arrows are approximately evenly distributed around the centre of the target. The last diagram shows *both* accuracy and precision, although in everyday conversation we would probably just describe it as accurate.

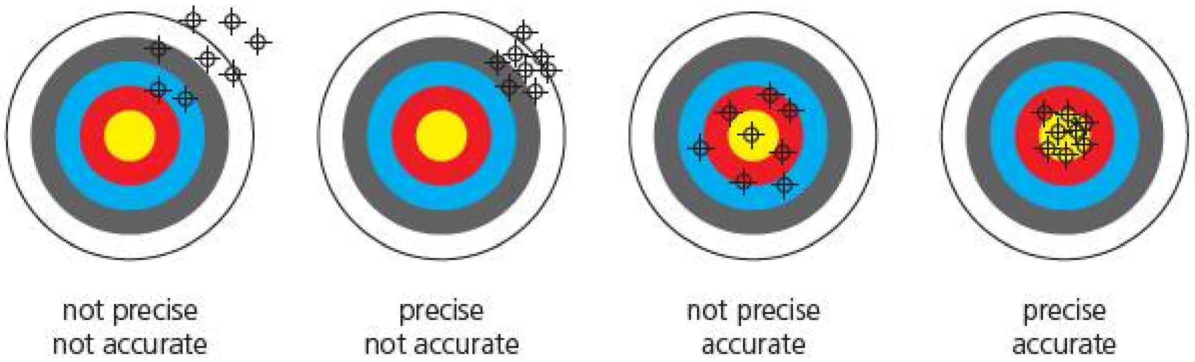


Figure 1.6 Difference between precision and accuracy

A watch that is always five minutes fast can be described as precise but not accurate. This is an example of a systematic zero offset error. Using stopwatches to time a 100 m race might give accurate results (if there are no systematic errors), but they are unlikely to be precise because human reaction times will produce significant random errors.

Uncertainties

Uncertainty in a measurement is the range, above and below the stated value, over which we would expect any repeated measurements to fall. If measurements are precise they have a low uncertainty. There is more about uncertainties in the next section.

The words *error* and *uncertainty* are sometimes used to mean the same thing, although this is only strictly true when referring to experiments with a known correct result and no systematic errors in the measurements.

TOK Link: Limits to the accuracy of data

No matter how hard we try, even with the best of measuring instruments, it is simply not possible to measure anything *exactly*. For one reason, the things that we can measure do not exist as perfectly exact quantities; there is no reason why they should.

This means that *every* measurement is an approximation. A measurement could be the most accurate ever made, for example the width of a ruler might be 2.283 891 034 cm, but that is still not perfect, and even if it was, we would not know because we would always need a more accurate instrument to check it. In this example we also have the added complication of the fact that when measurements of length become very small we have to deal with the atomic nature of the things we are measuring. (Where is the edge of an atom?)

Question

1 a What were the most accurate measurements that you ever made in your science experiments?

b Explain why you believe that they were accurate.

Reducing the effects of errors

1.2.8 **Explain** how the effects of random errors may be reduced.

The most common way of reducing the effects of *random errors* is by repeating measurements and calculating a mean value, which should be closer to the correct value than most, or all, of the individual measurements. Any unusual (**anomalous**) values should be checked and probably excluded from the calculation of the mean. Many experiments will involve taking a range of measurements, each under different experimental conditions, so that a graph can be drawn to show the pattern of the results. (For example, changing the voltage in an electrical circuit to see how it affects the current.) Increasing the number of pairs of measurements made will also reduce the effects of random errors because the line of best-fit can be placed with more confidence.

Experiments should be designed, wherever possible, to produce large readings. For example, a metre ruler may only be readable to the nearest half a millimetre and this will be the same for all measurements that are made with it. When measuring a length of 90 cm this error will probably be considered as acceptable (it is a percentage error of 0.56%), but the same sized error when measuring only 2 mm is 25%, which is probably an unacceptable result. The larger a measurement (that is made with a particular measuring instrument), the smaller the percentage error will be. If this is not possible, then the measuring instrument may need to be changed to one with smaller divisions.

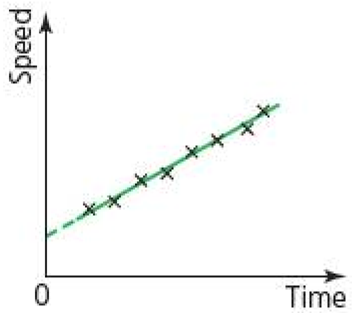


Figure 1.7 The best-fit line for this speed–time graph for a trolley rolling from rest down a slope does not pass through the origin. There was probably a systematic error

It is possible to carry out an experiment carefully with good quality instruments, but still have large random errors. There could be many different reasons for this and the experiments may have to be redesigned to get over the problems. Using a stopwatch to time the fall of an object dropped from a hand to the floor, or measuring the height of a bouncing ball, are two examples of simple experiments which may have significant random errors.

The effects of *systematic errors* cannot be reduced by repeating measurements. Instruments should be checked for errors before they are used, but a systematic error may not even be noticed until a graph has been drawn of the results and a line of best-fit does not pass through an expected intercept, as shown in Figure 1.7. In such a case it may then be sensible to adjust all measurements up or down by the same amount if the cause of the systematic error can be determined.

Using appropriate numbers of significant figures

1.2.9 Calculate quantities and results of calculations to the appropriate number of significant figures.

The more precise that a measurement is, the greater the number of significant figures (digits) that can be used to represent it. For example, an electric current stated to be 4.20 A (as distinct from 4.19 A or 4.21 A) suggests a much greater precision than a current stated to be 4.2 A.

Significant figures (digits) are all the digits used in data to carry meaning, whether they are before or after a decimal point, and *this includes zeroes*. But sometimes zeroes are used without thought or meaning, and this can lead to confusion. For example, if you are told that it is 100 km to the nearest airport, you may be unsure whether it is approximately 100 km, or 'exactly' 100 km. This is a good example of why scientific notation is useful. Using 1.00×10^3 km makes it clear that there are three significant figures. 1×10^3 represents much less precision.

When making *calculations*, the result cannot be more precise than the measurements used to produce it. As a general (and simplified) rule, when answering questions, or processing experimental data, the result should have the same number of significant figures as the data used. If the number of significant figures is not the same for each piece of data, then the number of significant figures in the answer should be the same as the least precise of the data (which has the fewest significant figures). This is illustrated in Worked example 2.

Worked example

2 Use the equation $P = mgh/t$ to determine the power, P, of an electric motor which raises a mass, m, of 1.5 kg, a distance, h, of 1.128 m in a time, t, of 4.79 s. ($g = 9.81\ \text{m s}^{-2}$)

$$P = \frac{mgh}{t} = \frac{1.5 \times 9.81 \times 1.128}{4.79}$$

A calculator will display an answer of 3.4652……, but this answer suggests a very high precision which is not justified by the data. The data used which has the lowest precision is 1.5 kg, and the answer should also have the same number of significant figures:

$P = 3.5$ W

'Rounding off'

'Rounding off' should be done at the *end* of a multi-step calculation, when the answer has to be given. If further calculations using this answer are then needed, *all* the digits shown previously on the calculator should be used. The answer to this calculation should then be rounded off to the correct number of significant figures. This process can sometimes result in small apparent inconsistencies between answers.

Uncertainties in calculated results

1.2.10 State uncertainties as absolute, fractional and percentage uncertainties.

Uncertainties in experimental data or calculations may be expressed in one of three ways:

1. The **absolute uncertainty** of a measurement is the range, above and below the stated value, within which we would expect any repeated measurements to fall. For example, the mass of a pen may be stated to be 53.2 g ± 0.1 g, where the uncertainty is ± 0.1 g.
2. The **fractional uncertainty** is the ratio of the absolute uncertainty to the measured value.
3. The **percentage uncertainty** is the fractional uncertainty expressed as a percentage. Uncertainties expressed in percentages are often the most informative. Experiments which produce results with uncertainties of less than 5% may be desirable, but are not always possible.

Worked example

3 The mass of a piece of metal is quoted to be 346 g ± 2.00%.
  a What is the absolute uncertainty?
  b What is the range of values that the mass could be expected to have?
  c What is the fractional uncertainty?

a 2.00% of 346 g is ± 7 g

b 339 g to 353 g (to 3 significant figures)

c $\frac{7}{346} = \frac{1}{50}$

Ideally, uncertainties should be quoted for all experimental measurements, but this can be repetitive and tedious in a learning environment, so they are often omitted unless being taught specifically.

It is usually easy to decide on the size of an uncertainty associated with taking a single measurement with a particular instrument. It is often assumed to be the readability error, as described above. However, the overall uncertainty in a measurement, allowing for all experimental difficulties, is sometimes more difficult to decide. For example, the readability error on a hand-operated stopwatch may be 0.01 s, but the uncertainty in its measurements will be much greater because of human reaction times. The amount of scattering of the readings around a mean value is a useful guide to random uncertainty, but not systematic uncertainty. After the mean value of the readings has been calculated, the random uncertainty can be assumed to be the largest difference between any single reading and the mean value. This is shown in the following example.

Worked example

4 The following measurements (in cm) were recorded in an experiment to measure the height to which a ball bounced: 32, 29, 33, 32, 37 and 28. Estimate values for the absolute and percentage random uncertainties in the experiment.

The mean of these six readings is 31.83 cm, but it would be sensible to quote this to two significant figures (32 cm), as in the original data. The measurement which has the greatest difference from this value is 37 cm, so an estimate of the uncertainty is 5 cm, which means a percentage uncertainty of $(5 \times 100)/37 = 14\%$.

Note that if the same data had been obtained in the order 28, 29, 32, 32, 33, 37, it would be difficult to believe that the uncertainties were random, and another explanation for the variation in results would need to be found.

Calculations

1.2.11 Determine the uncertainties in results.

When making further calculations based on experimental data, the uncertainty in individual measurements should be known. It is then important to know how to use these uncertainties to determine the uncertainty in any results that are calculated from that data.

Consider a simple example: a trolley moving with constant speed was measured to travel a distance of 76 cm ± 2 cm (± 2.6%) in a time of 4.3 s ± 0.2 s (± 4.7%).

The speed can be calculated from distance/time = 76/4.3 = 1.767…, which is 1.8 m s^{-1} when rounded to two significant figures, consistent with the experimental data.

To determine the uncertainty in this answer we consider the uncertainties in distance and time. Using the largest distance and smallest time, the largest possible answer for speed is 78/4.1 = 1.902… . Using the smallest distance and the greatest time, the smallest possible answer for speed is 74/4.5 = 1.644…. . (The numbers will be rounded at the end of the calculation.)

The speed is therefore between 1.644 m s^{-1} and 1.902 m s^{-1}. The value 1.902 has the greater difference (0.135) from 1.767. So our final result can be expressed as 1.767 ± 0.135 m s^{-1}, which is a maximum uncertainty of 7.6%. Rounding to two significant figures, the result becomes 1.8 ± 0.1 m s^{-1}.

Uncertainty calculations like these can be very time consuming and, for this course, *approximate methods are acceptable*. For example, in the calculation for speed shown above, the uncertainty in the data was ± 2.6% for distance and ± 4.7% for time. The percentage uncertainty in the final result is approximated by adding the percentage uncertainties in the data: 2.6 + 4.7 = 7.3%. This gives approximately the same value as calculated using the largest and smallest possible values for speed. Rules for finding uncertainties in calculated results are given below.

Rules for uncertainties calculations

- For quantities that are added or subtracted: *add* the *absolute* uncertainties.
  In the IB *Physics data booklet* this is given as:

$$\text{If } y = a \pm b \text{ then } \Delta y = \Delta a + \Delta b$$

- For quantities that are multiplied or divided, or raised to powers: *add* the individual *fractional* or *percentage* uncertainties.
  In the IB *Physics data booklet* this is given as:

$$\text{If} \quad y = \frac{ab}{c} \quad \text{then} \quad \frac{\Delta y}{y} = \frac{\Delta a}{a} + \frac{\Delta b}{b} + \frac{\Delta c}{c}$$

- For other functions (such as trigonometric functions, logarithms or square roots): calculate the highest and lowest *absolute* values possible and compare with the mean value.

Worked example

5 An angle, θ, was measured to be $34° \pm 1°$. What is the uncertainty in the tangent of this angle?

$\tan 34° = 0.675$; $\tan 33° = 0.649$; $\tan 35° = 0.700$
$0.700 - 0.675 = 0.026$ $\quad$ ($0.675 - 0.649 = 0.025$, which is smaller than 0.026)
$\tan \theta = 0.67 \pm 0.03$
(to be consistent with the experimental data)

5 A mass of 346 ± 2 g was added to a mass of 129 ± 1 g.
a What was the overall absolute uncertainty?
b What was the overall percentage uncertainty?

6 The equation $s = \frac{1}{2}at^2$ was used to calculate a value for s when a was $4.3 \pm 0.2\,\text{m s}^{-2}$ and t was $1.4 \pm 0.1\,\text{s}$.
a Calculate a value for s.
b Calculate the percentage uncertainty in the data provided.
c Calculate the percentage uncertainty in the answer.
d Calculate the absolute uncertainty in the answer.

7 A certain quantity was measured to have a magnitude of (1.46 ± 0.08). What is the maximum uncertainty in the square root of this quantity?

Using computer spreadsheets to calculate uncertainties

Computer spreadsheets can be very helpful when it is necessary to make multiple calculations of uncertainties in experimental results. For example, the resistivity, ρ, of a metal wire can be calculated using the equation $\rho = R\pi r^2/l$, where r and l are the radius and length of the wire, and R is its resistance. Figure 1.8 shows the **raw data** (shaded green) of an experiment that measured the resistance of various wires of the same metal. The rest of the spreadsheet shows the calculations involved with processing the data to determine resistivity and the uncertainty in the result. A computer program can then be use to draw a suitable graph of the results, and this may include error bars (see next section).

RESISTANCE		RADIUS			LENGTH		RESISTIVITY		
Resistance, R / Ω ±0.2 Ω	Percentage uncertainty in R	Radius, r / mm ±0.01 mm	Percentage uncertainty in r	Percentage uncertainty in r^2	Length, l / cm ± 1 cm	Percentage uncertainty in l	Resistivity, $\rho = R\pi r^2/l$ / Ωm	Percentage uncertainty in ρ	Absolute uncertainty in ρ / Ωm
9.4	2.1	0.15	6.7	13.3	44	2.3	0.0000015	18	0.0000003
6.2	3.2	0.22	4.5	9.1	67	1.5	0.0000014	14	0.0000002
6.2	3.2	0.25	4.0	8.0	80	1.3	0.0000015	12	0.0000002
5.2	3.8	0.30	3.3	6.7	99	1.0	0.0000015	12	0.0000002
5.0	4.0	0.35	2.9	5.7	128	0.8	0.0000015	10	0.0000002
3.8	5.3	0.43	2.3	4.7	149	0.7	0.0000015	11	0.0000002
3.4	5.9	0.51	2.0	3.9	175	0.6	0.0000016	10	0.0000002
2.4	8.3	0.62	1.6	3.2	198	0.5	0.0000015	12	0.0000002

Figure 1.8 Using a spreadsheet to calculate uncertainties in a resistance experiment

8 **a** Use a computer spreadsheet to enter the same raw data as shown in Figure 1.8.
b Use the spreadsheet to confirm the results of the calculations shown.
c What difference would it make to the results if the radius of the wire could only be measured to the nearest half a millimetre?

Uncertainties in graphs

1.2.12 **Identify** uncertainties as error bars in graphs.
1.2.13 **State** random uncertainty as an uncertainty range (±) and represent it graphically as an 'error bar'.

The range of random uncertainty in a measurement or a calculated result can be represented on a graph by using crossed lines to mark the point (instead of a dot).

Figure 1.9 shows an example: a graph of distance against time for the motion of a train. Vertical and horizontal lines are drawn through each data point to represent the uncertainties in the two measurements. In this example, the uncertainty in time is ± 0.5 s and the uncertainty in distance is ± 1 m. These lines, which usually have small lines to indicate clearly their end, are called **error bars**. In Figure 1.9 the space outlined by each error bar has been shaded for emphasis – it is expected that a line of best fit should pass somewhere through each shaded area.

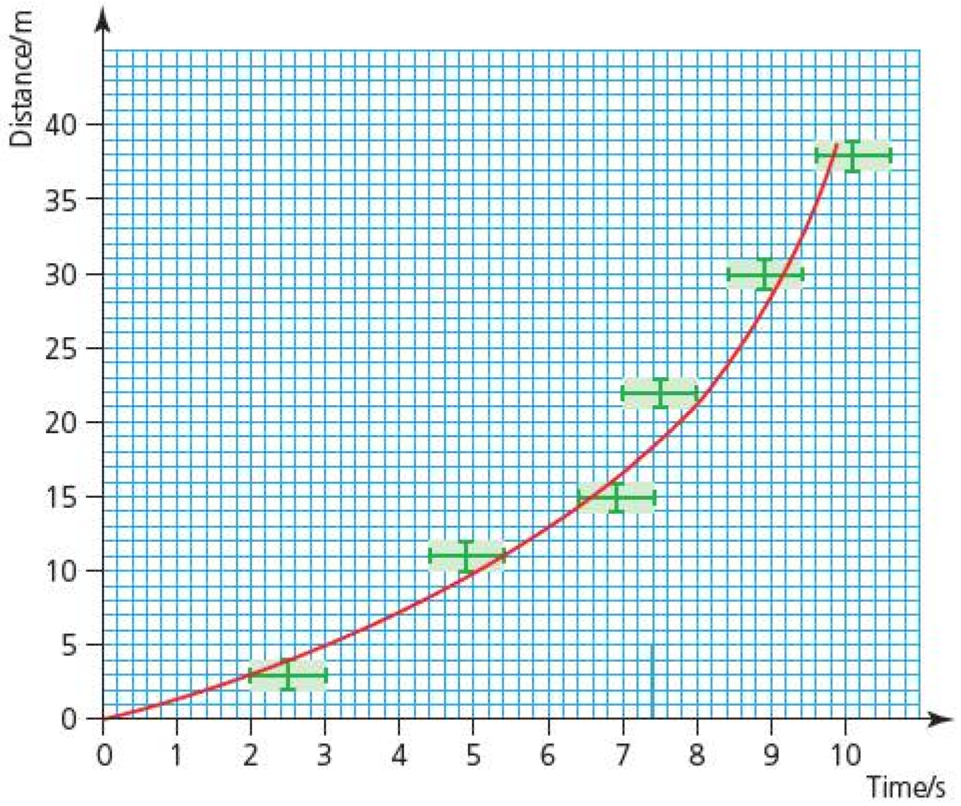


Figure 1.9 Showing uncertainty using error bars

In some experiments the error bars are so small and insignificant that they are not included on the graph. For example, a mass could be measured as 347.46 ± 0.01 g. The uncertainty in this reading would be too small to show as an error bar on a graph. (Note that error bars are not expected for trigonometric or logarithmic functions.)

Uncertainties in gradients and intercepts

1.2.14 **Determine** the uncertainties in the gradient and intercepts of a straight-line graph.

If the results of an experiment suggest a straight line graph, it is often important to determine values for the gradient and/or the intercept(s) with the axes. However, it will often be possible to draw a range of different straight lines, all of which pass through the error bars representing the experimental data.

We usually assume that the best-fit line is midway between the lines of maximum possible gradient and minimum possible gradient. Figure 1.10 shows an example. Note that, for simplicity when determining gradients and intercepts, in this course it is only necessary to draw the error bars for the first and last data points.

Figure 1.10 shows how the length of a metal spring changed when the force applied was increased. The measurements were not very precise. This graph shows the error bars for the first and last readings only. The line of best-fit has been drawn midway between the other two. This is a **linear graph** (a straight line) and it is known that the gradient of the graph represents the force constant (stiffness) of the spring and the *x*-intercept represents the original length of the spring. Taking measurements from the best-fit line, we can make the following calculations:

$$\text{force constant} = \text{gradient} = \frac{(90 - 0)}{(6.6 - 1.9)} = 19\,\text{N cm}^{-1}$$

original length = *x*-intercept = 1.9 cm

To determine the uncertainty in the calculations of gradient and intercept, we need only to consider the range of straight lines that could be drawn through the first and last error bars. The uncertainty will be the maximum difference between these extreme values and the value calculated from the best-fit line. In this example it can be shown that:

force constant is between 14 N cm^{-1} and 28 N cm^{-1}
original length is between 1.1 cm and 2.6 cm

The final result can be quoted as: force constant = 19 ± 9 N cm^{-1} and original length is 1.9 ± 0.8 cm. Clearly, the large uncertainties in these results confirm that the experiment lacked precision.

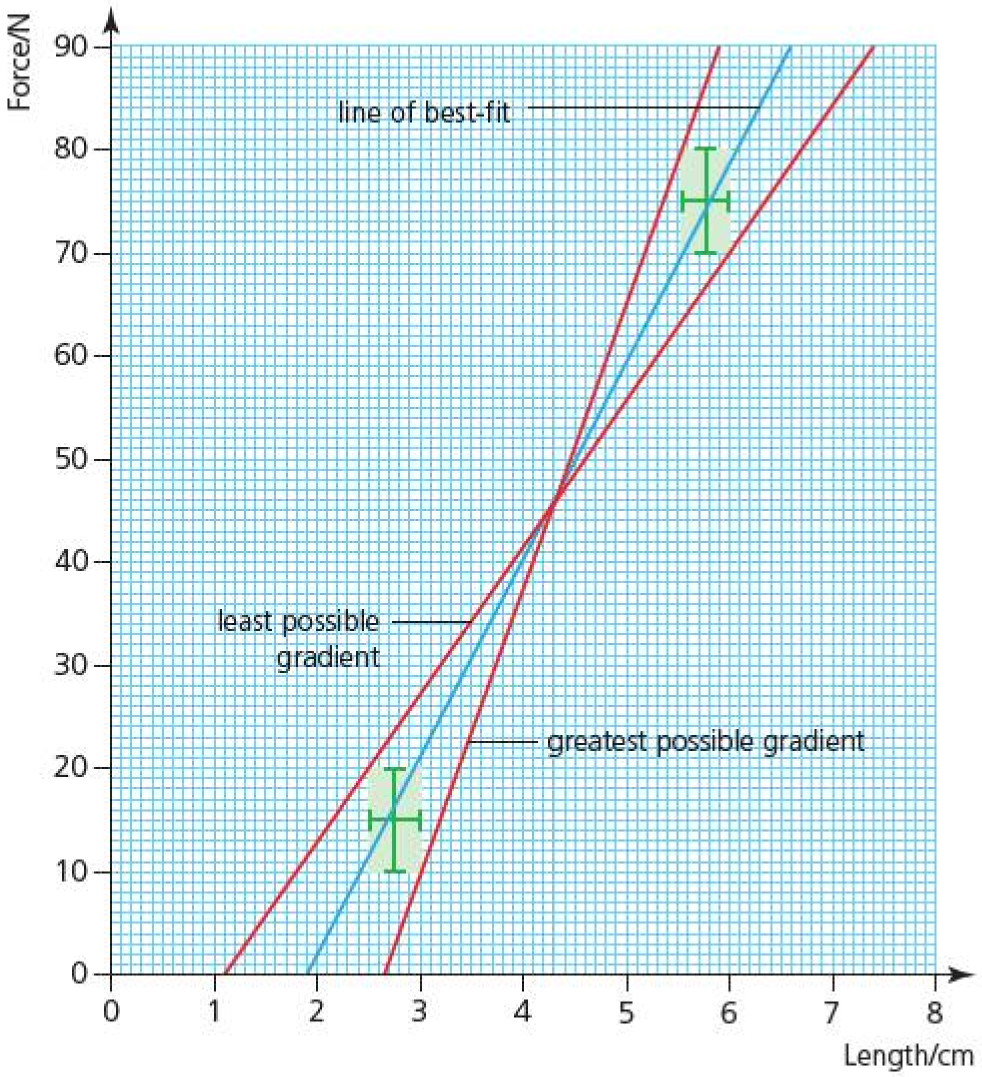


Figure 1.10 Finding maximum and minimum gradients for a spring stretching experiment

1.3 Vectors and scalars

1.3.1 **Distinguish** between vector and scalar quantities and give examples of each.

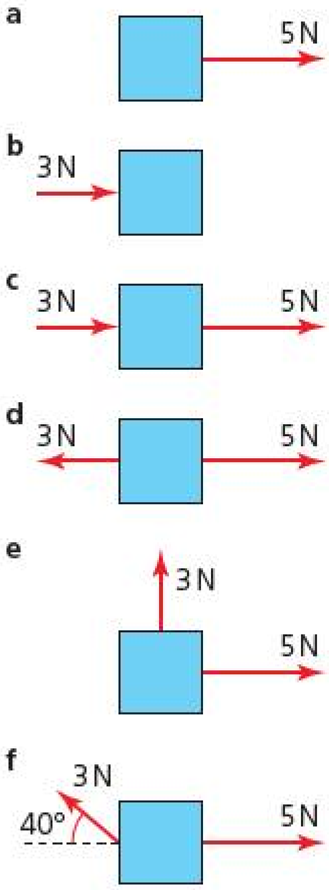


Figure 1.11 Forces are vector quantities

The diagrams in Figure 1.11 show the force(s) acting on a mass. In Figure 1.11a the mass is being pulled to the right with a force of 5 N. The length of the arrow represents the size of the force and the orientation of the arrow shows the direction in which the force acts. The length of the arrow is proportional to the force. In Figure 1.11b there is a smaller force (3 N) pushing the mass to the right. In both of these examples the mass will move (accelerate) to the right.

In Figure 1.11c there are two forces acting. We can add them together to show that the effect is the same as if a single force of 8 N (= 3 + 5) was acting on the mass. We say that the **resultant** (net) force is 8 N.

In Figure 1.11d there are two forces acting on the mass, but they act in different directions. The overall effect is still found by 'adding' the two forces, but also taking their direction into account. This can be written as +5 + (–3) = +2 N, where forces to the right are given a positive sign and forces to the left are given a negative sign. The resultant will be the same as if there was only one force (2 N) acting to the right. In Figures 1.11e and 1.11f there are also two forces acting, but they are not acting along the same line. For these forces, the resultant can be determined using a scale drawing or trigonometry (see page 15).

Clearly, force is a quantity for which we need to know its direction as well as its **magnitude** (size). Quantities that have both magnitude and direction are called **vectors**.

Everything that we measure has a number and a unit. For example, we may measure the mass of a book to be 640 g. Here 640 g is the magnitude of the measurement, but mass has no direction.

Quantities that have only magnitude, and no direction, are called **scalars**. Most quantities are scalars. Some common examples of scalars used in physics are mass, length, time, energy, temperature and speed.

However, when using the following quantities, we need to know both the magnitude and the direction in which they are acting, so they are vectors:

- displacement (distance in a given direction)
- velocity (speed in a given direction)
- force (including weight)
- acceleration
- momentum
- field strength (gravitational, electric or magnetic).

The symbols for vector quantities are often shown in bold italic (for example, ***F***). Scalar quantities are shown with a normal italic font (for example, *m*).

In diagrams, all vectors are shown with straight arrows, pointing in the right direction, and which have a length proportional to the magnitude of the vector (as shown by the forces in Figure 1.11).

Determining the sum or difference of two vectors

1.3.2 **Determine** the sum or difference of two vectors by a graphical method.

Adding vectors to determine a resultant

When two or more scalar quantities are added together (for example masses of 25 g and 50 g), there is only one possible answer (resultant), 75 g; but when vector quantities are added, there is a range of different resultants possible, depending on the directions involved.

To determine the resultant of the two forces shown in Figures 1.11e or 1.11f there are two possible methods: by drawing (graphical method) or by trigonometry.

- **Graphical method** – The two vectors shown in Figure 1.11f are drawn carefully to scale (for example, by using 1 cm to represent 1 N), with the correct angle (140°) between them. A parallelogram is then completed. The resultant is the diagonal of the parallelogram (see Figure 1.12). Remember that the magnitude and the direction should *both* be determined from the diagram. In this example the resultant force is represented by the line drawn in red. Its length is 3.4 cm, which represents 3.4 N, at an angle of 36° to the 5.0 N force.

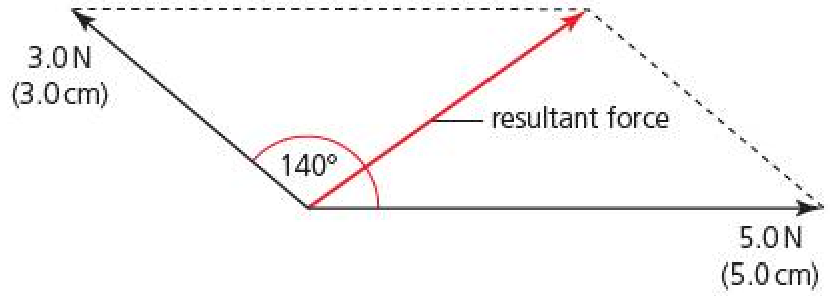


Figure 1.12 Using a parallelogram to determine a resultant force

- **Trigonometry** – The forces in Figure 1.11e are at right angles to each other. This means that a parallelogram drawn to represent these forces will be a rectangle (Figure 1.13) and the magnitude of the resultant of the forces, F, can be found using Pythagoras' theorem:

$$F^2 = 3.0^2 + 5.0^2 = 34$$
$$F = 5.8\,\text{N}$$

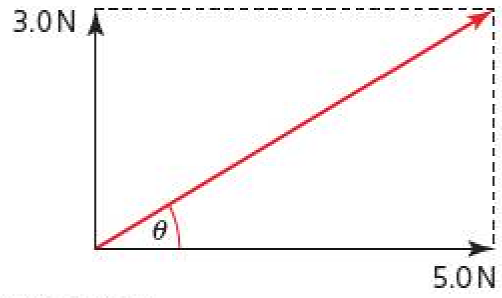


Figure 1.13

The direction of this force can be determined by using trigonometry:

$$\tan\theta = \frac{3.0}{5.0}$$ (θ is the angle that the resultant makes with the direction of the 5.0 N force)

$$\theta = 31°$$

Subtracting vectors to find their difference

We need to know the *difference* between two vectors when we are considering by how much a vector quantity has changed. It is determined by *subtracting* one vector from the other. A negative vector has the same magnitude, but opposite direction, to a positive vector, so when finding the difference between vectors **P** and **Q** we can write:

$$\mathbf{P} - \mathbf{Q} = \mathbf{P} + (-\mathbf{Q})$$

Figure 1.14 shows how vectors are subtracted graphically.

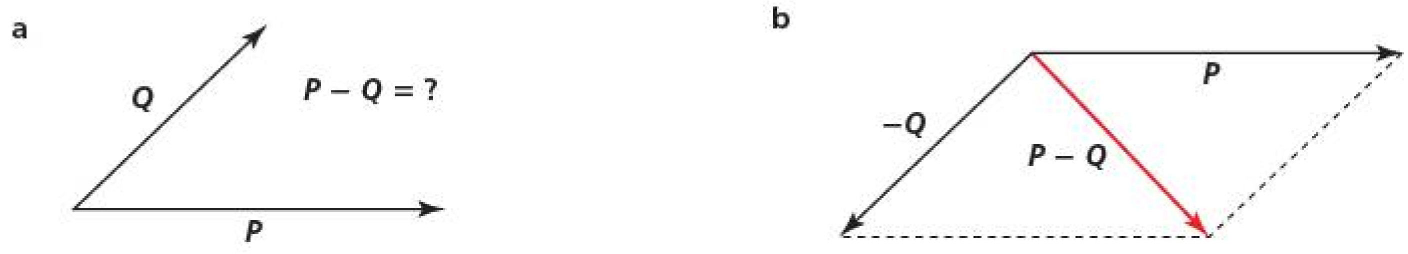


Figure 1.14 If we want to know the difference between ***P*** and ***Q*** (diagram **a**) we add ***P*** to –***Q*** (diagram **b**)

Multiplying and dividing vectors by scalars

If a vector **P** is multiplied or divided by a scalar number k, the resultant vectors are simply $k\mathbf{P}$ and $\mathbf{P}/k$. If k is negative, then the resultant vector becomes negative, meaning that the direction is reversed.

Resolving a single vector into two components

1.3.3 **Resolve** vectors into perpendicular components along chosen axes.

We have seen that two individual vectors can be combined mathematically to find a single resultant that has the same effect as the two separate vectors. This process can be reversed: a single vector can be considered as having the same effect as two separate vectors. This is called **resolving** a vector into two **components**. Resolving can be very useful because, if the two components are chosen to be perpendicular to each other (often horizontal and vertical), they will then both be independent of each other, so that they can both be considered totally separately.

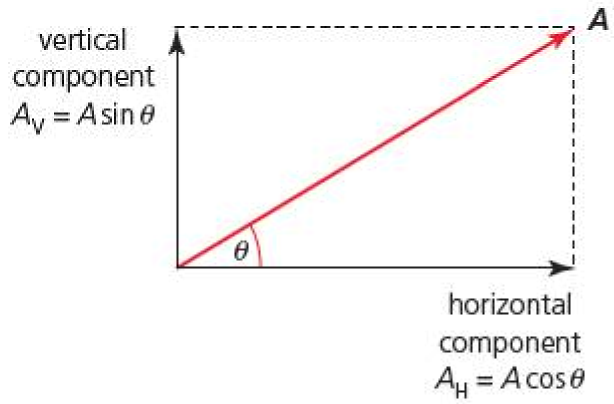


Figure 1.15 Resolving a vector into two perpendicular components.

Figure 1.15 shows a single vector, **A**, acting at an angle θ to the horizontal. If we want to know the effects of this vector in the horizontal and vertical directions, we can resolve it into two components:

$$\cos\theta = \frac{A_H}{A} \quad \text{and} \quad \sin\theta = \frac{A_V}{A}$$

so

$$A_H = A\cos\theta \quad \text{and} \quad A_V = A\sin\theta$$

Both of these equations and the associated diagram are given in the IB *Physics data booklet*.

Worked example

6 Figure 1.16 shows a box resting on a sloping surface (an 'inclined plane'). The box has a weight of 585 N. What are the components of weight:

a down the slope

b perpendicularly into the slope?

a Component down the slope = 585 sin 23° = 230 N

b Component into the slope = 585 cos 23° = 540 N

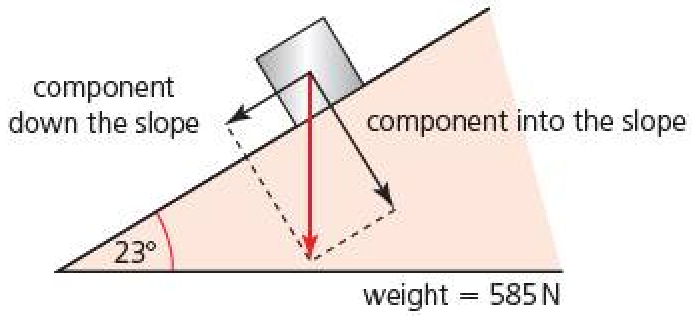


Figure 1.16

SUMMARY OF KNOWLEDGE

1.1 The realm of physics

- When numbers are approximated and quoted to the nearest power of 10, it is called giving them an order of magnitude.
- The mass of the universe is more than 10^{50} kg and the mass of its smallest particle (electron) is about 10^{-30} kg.
- The distance to the edge of the visible universe is about 10^{27} m. The nucleus of an atom has a size of approximately 10^{-15} m.
- The age of the universe is approximately 10^{18} s, but it would take light only about 10^{-23} s to travel across the width of a nucleus.
- It is important to be able to make reasonable estimates of various quantities (and give them to a sensible number of significant figures), and to make comparisons between quantities to the nearest order of magnitude.

1.2 Measurement and uncertainty

- Wherever possible, the units of the SI system should be used.
- The kilogram, metre, second, ampere, mole and kelvin are all fundamental units.
- Derived units are combinations of fundamental units. For example the derived unit of density is $kg\,m^{-3}$. Some derived units have their own name. For example the unit of pressure, $N\,m^{-2}$, is called the pascal, Pa.
- Sometimes physicists use units that are not part of the SI system, for example years or electronvolts. It is important to be able to convert between units for the same quantity.
- The accepted format for writing units is, for example, $W\,m^{-2}$ for watts per square metre, and not W/m^2.
- Values in science are commonly expressed using standard notation, for example 3.9820×10^4, rather than 39 820.
- Units in science are expressed using a range of standard prefixes (for example, kilo-) (see Table 1.4).
- Random errors (uncertainties) occur in all experiments for a range of different reasons, and they result in values scattered around the 'correct' result.
- The effect of random errors can be reduced by calculating the averages of repeated readings, and by taking a range of different readings.
- If the same error occurs in every measurement (made using the same instrument and technique), it is called a systematic error. For example, an instrument may have a zero offset error. This kind of error is not reduced by repeating readings.
- An accurate measurement is one which has a low systematic error. Precise measurements have low random errors.
- The uncertainty of a reading is the range of values within which we would expect any repeated readings to fall. Precise readings have small uncertainties. Uncertainties are usually given as absolute values (with one significant digit in the unit of measurement) or as percentages.
- The number of significant figures used in data represents its precision. A calculated result should not have more significant figures than the least precise data used in the calculation.
- The approximate uncertainty in the results of additions and subtractions is equal to the sum of the absolute uncertainties. For multiplication and division the uncertainty in the result is determined by adding the percentage or fractional uncertainties. For other functions, individual calculations may need to be made.
- Uncertainties in measurements can be represented by error bars on graphs.
- Gradients and intercepts of straight-line graphs can provide important information. The uncertainty in their calculated values can be determined by considering the lines of maximum and minimum gradient that can pass through the first and last error bars.

1.3 Vectors and scalars

- A vector quantity has magnitude and direction. It is represented in a diagram by a straight line in the right direction, with a length proportional to the magnitude. Examples of vectors include force, momentum and gravitational field strength.
- Scalar quantities only have magnitude. Examples include mass, energy and time.
- The resultant of two vectors can be determined in magnitude and direction from the diagonal of a parallelogram drawn to scale.
- The difference between two vectors can be determined by adding the first to the negative of the second.
- A single vector can be resolved into two components at right angles to each other. The two components can then be considered separately and independently. The two components of a force $\boldsymbol{F}$ are $F \sin\theta$ and $F \cos\theta$.

Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 The diameter of a wire was measured three times with an instrument that has a zero offset error. The results were 1.24 mm, 1.26 mm and 1.25 mm. The average of these results is

A accurate but not precise.
B precise but not accurate.
C accurate and precise.
D not accurate and not precise.

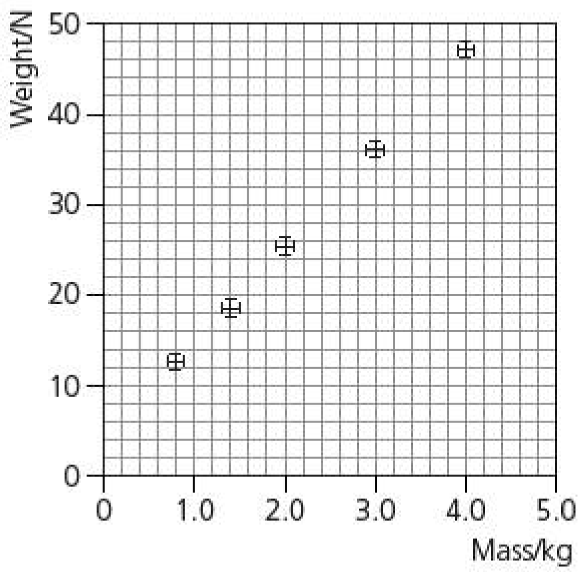
Q2 The approximate thickness of a page in a textbook is

A 0.02 mm.
B 0.08 mm.
C 0.30 mm.
D 1.00 mm.

Q3 Which of the following is the approximate conversion (s) of a time of 1 month into SI units?

A 0.08 y
B 30 d
C 3×10^6 s
D all of the above

Q4 The masses and weights of different objects are independently measured. The graph is a plot of weight versus mass that includes error bars.



These experimental results suggest that

A the measurements show a significant systematic error but small random error.
B the measurements show a significant random error but small systematic error.
C the measurements are precise but not accurate.
D the weight of an object is proportional to its mass.

Standard Level Paper 1, May 09 TZ1, Q2

Q5 Which of the following is a fundamental SI unit?

A newton
B coulomb
C ampere
D joule

Q6 The distance travelled by a car in a certain time was measured with an uncertainty of 6%. If the uncertainty in the time was 2%, what will the uncertainty be in a calculation of the car's speed?

A 3%
B 4%
C 8%
D 12%

Q7 Which of the following quantities is a scalar?

A pressure
B acceleration
C gravitational field strength
D displacement

Q8 The current in a resistor is measured as 2.00 A ± 0.02 A. Which of the following correctly identifies the absolute uncertainty and the percentage uncertainty in the current?

	Absolute uncertainty	Percentage uncertainty
A	± 0.02 A	± 1%
B	± 0.01 A	± 0.5%
C	± 0.02 A	± 0.01%
D	± 0.01 A	± 0.005%

Standard Level Paper 1, May 10 TZ2, Q1

Q9 Which of the following is a reasonable estimate of the order of magnitude of the mass of a large aircraft?

A 10^3 kg
B 10^5 kg
C 10^7 kg
D 10^9 kg

Q10 Which of the following is equivalent to the SI unit of force (the newton)?

A $kg\,m\,s^{-1}$
B $kg\,m^2\,s^{-1}$
C $kg\,m\,s^{-2}$
D $kg\,m^2\,s^2$

2 Mechanics

STARTING POINTS

- Mechanics is the study of motion and the effects of forces on objects.
- A vector quantity has both magnitude (size) and direction. A scalar quantity has only magnitude.
- Vector quantities can be added together to determine a resultant by using scale drawings (or trigonometry) that take account of their directions.
- It is often convenient to resolve a single vector into two components that are perpendicular to each other.

2.1 Kinematics

Kinematics is the study of moving objects. To completely describe the motion of an object at any one moment we need to say where it is, how fast it is moving, and in what direction. For example, we might observe that a car is 20 m to the west of us, and moving northeast at a speed of $8\,m\,s^{-1}$ (see Figure 2.1).

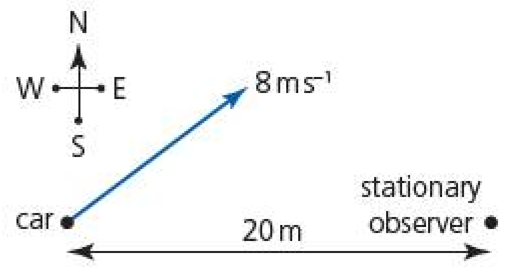


Figure 2.1 Describing the position and motion of a car

Of course, any or all of these quantities may be changing. In real life the movement of many moving objects can be complicated; they do not often move in straight lines and they may even rotate or have different parts moving in different directions.

We will develop an understanding of the basic principles of kinematics by first dealing with single objects moving in straight lines or along circular paths, and calculations will be confined to those objects that have a regular motion.

Describing motion

2.1.1 **Define** *displacement*, *velocity*, *speed* and *acceleration*.
2.1.2 **Explain** the difference between instantaneous and average values of speed, velocity and acceleration.

Distance and displacement

The **displacement** of an object is its position compared to a reference point. For example, the displacement of the car in Figure 2.1 is 20 m to the west of the observer. To specify a displacement we need to state a distance *and* a direction from the reference point. The reference point is often omitted because it is obvious, for example, we might just say that an airport is 50 km to the north. Although a direction may be anywhere in three dimensions, we usually restrict our thinking to one or two dimensions.

Displacement is defined as the distance in a given direction from a fixed reference point.

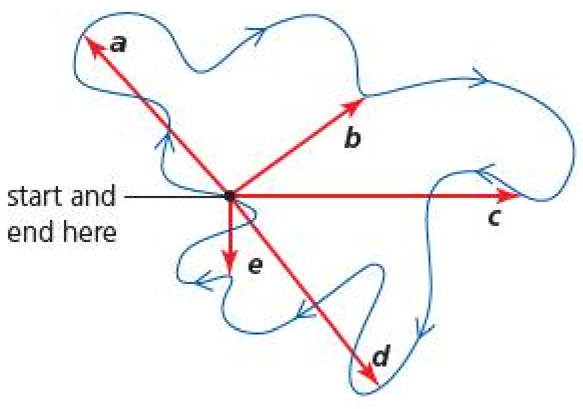


Figure 2.2 A walk in the park

Displacement and distance are both given the symbol s. This should not be confused with the symbol for speed (and velocity), which is v. The symbol h is also widely used for vertical distances (heights). The SI unit for distance is the metre, m, although other units, such as mm, cm and km, are in common use.

Because a direction is specified as well as a magnitude (size), displacement is a **vector** quantity. Distance is a **scalar** quantity because it has magnitude but no direction.

Figure 2.2 shows the route of some people walking around a park. The total *distance* walked was 4 km, but the displacement from the reference point varied and is shown every few minutes by the vector arrows (a to e). The final *displacement* is zero because the walkers returned to their starting place.

Speed and velocity

Speed is defined as the rate of change of distance with time.

Speed is a scalar quantity and is given the symbol v. Its SI unit is metres per second, $m\,s^{-1}$. Speed is calculated from:

$$\text{speed} = \frac{\text{distance travelled}}{\text{time taken}} \left(\text{or, commonly, speed} = \frac{\text{distance}}{\text{time}}\right)$$

The **delta symbol** (Δ) is used wherever we want to represent a (small) change of something, so we can define speed in symbols as follows:

$$v = \frac{\Delta s}{\Delta t}$$

If an object is moving with a constant speed, determining its value is a straightforward calculation. However, the speed of an object often changes during the time we are observing it, and the calculated value is then only an **average speed** during that time. For example, if a car is driven a distance of 120 km in 1.5 h, its average speed is $80\,km\,h^{-1}$, but its actual speed will certainly have varied during the journey. At any one time we may look at the car's speedometer to find out the **instantaneous speed**, that is, the speed at that exact instant (moment). In science we are often much more interested in instantaneous values of speed (and velocity and acceleration) than average values.

Average speeds are calculated over lengths of time that are long enough for the actual speeds to have changed. Instantaneous values have to be calculated from measurements made over very short time intervals (during which time we can assume that the speed was constant).

Speed is therefore calculated using the total distance travelled in the time being considered, regardless of the direction of motion. If the walkers in Figure 2.2 took two hours to walk around the park, their **average speed** would be $\Delta s/\Delta t$ (= 4/2) = $2\,km\,h^{-1}$.

We are often concerned not only with how fast an object is moving, but also with the direction of movement. If speed *and* direction are stated, then the quantity is called **velocity**.

Velocity is defined as the rate of change of displacement with time (speed in a given direction).

$$v = \frac{\Delta s}{\Delta t}$$

Note that Δs in this equation refers to displacement and not to the overall distance. (To avoid confusion, it is generally better to define speed and velocity in words, not symbols.)

Velocity has the same symbol and unit as speed, but the direction should usually be stated as well, since velocity is a vector quantity. However, if the direction of motion does not change, it is common to refer to a speed, of say $4\,m\,s^{-1}$, as its velocity, because the direction is understood from the context.

Returning to the walkers in the park, at the end of their walk their average speed was $2\,km\,h^{-1}$, but their average velocity was zero because the final displacement was zero. This may not be a very useful piece of information; we are more likely to be interested in the instantaneous velocity at various times during the walk.

When the velocity (or speed) of an object changes during a certain time, the symbol u is used for the *initial velocity* and v is used for the *final velocity* during that time. These velocities are not necessarily the beginning and end of the entire motion, just the velocities at the start and end of the time that is being considered.

Acceleration

Any variation from moving at a constant speed in a straight line is called an **acceleration**. It is very important to realize that going faster, going slower and/or changing direction are all different kinds of acceleration.

Acceleration, a, is defined as the rate of change of velocity with time.

$$a = \frac{\Delta v}{\Delta t} = \frac{v - u}{t}$$

The SI unit of acceleration is metres per second squared, $m\,s^{-2}$ (the same as the units of velocity/time, $m\,s^{-1}/s$). Acceleration is a vector quantity.

So, acceleration can be:

- an increase in velocity (positive acceleration)
- a decrease in velocity (negative acceleration – sometimes called a **deceleration**)

and/or

- a change of direction.

Additional Perspectives

Reaction times when timing motions

The delay between seeing something happen and responding with some kind of action is known as our reaction time. A typical value is about 0.20 s, but it can vary considerably depending on the conditions involved. A simple way of measuring a person's reaction time is by measuring how far a metre rule can fall before it can be caught between thumb and finger. The time can then be calculated using the equation $s = 5t^2$.

The measurement can be repeated with the person tested being blindfolded to see if the reaction time changes if the stimulus (to catch the ruler) is either sound or touch, rather than sight. Whatever tests are carried out, our reaction times are likely to be inconsistent. This means that whenever we use stopwatches operated by hand, the results will have an unavoidable uncertainty (see Chapter 1). It is sensible to make time measurements as long as possible to decrease the significance of this problem. (This reduces the percentage uncertainty.) Repeating measurements and calculating an average will also reduce the random errors.

Activity

Use the method described above (or any other) to measure your reaction time when the stimulus is sight. Repeat the measurement 10 times.

a What was the percentage variation between your average result and your quickest reaction time?

b Did your reaction times improve with practice?

Equations of uniformly accelerated motion

2.1.3 **Outline** the conditions under which the equations for uniformly accelerated motion may be applied.

If we know the initial velocity and acceleration of an object, and the acceleration is *uniform* (constant), then we can determine its final velocity v after time t by rearranging the equation for acceleration shown above. This gives:

$$v = u + at$$

Furthermore, the distance travelled in time t can be determined using the equation, distance = average speed × time. For an object with constant acceleration, the average speed = $\frac{1}{2}$(initial speed + final speed). For example, if a car accelerates uniformly from $12\,m\,s^{-1}$ to $16\,m\,s^{-1}$, then its average speed during that time was $14\,m\,s^{-1}$.

In symbols, this is shown as:

$$s = \frac{u+v}{2}t$$

This equation is given in the IB *Physics data booklet*.

The five quantities u, v, a, s and t are all that is needed to fully describe the motion of an object moving with constant/uniform acceleration. If any three of them are known, the other two can be calculated using the two equations above. The equations can be combined mathematically to give two further equations, shown below, which are also found in the IB *Physics data booklet*. These very useful equations do not involve any further physics theory; they just express the same physics principles in a different way.

$$v^2 = u^2 + 2as$$
$$s = ut + \frac{1}{2}at^2$$

Remember, the four equations of motion can only be used if the acceleration is *uniform* during the time interval being considered.

Figure 2.3 Formula One cars ready to go

Worked examples

1 A Formula One racing car (see Figure 2.3) accelerates *from rest* (i.e. it was *stationary* to begin with) at $18\,\mathrm{m\,s^{-2}}$.

a What is its speed after 3.0 s?

b How far does it travel in this time?

c If it continues to accelerate at the same rate, what will its velocity be after it has travelled 200 m from rest?

d Convert the final velocity to $\mathrm{km\,h^{-1}}$.

a $v = u + at$

$v = 0 + (18 \times 3.0)$

$v = 54\,\mathrm{m\,s^{-1}}$

b $s = \frac{u+v}{2}t$

$s = \frac{0+54}{2} \times 3.0$

$s = 81\,\mathrm{m}$

But note that the distance can be calculated directly, without first calculating the final velocity, as follows:

$s = ut + \frac{1}{2}at^2$

$s = (0 \times 3.0) + (0.5 \times 18 \times 3.0^2)$

$s = 81\,\mathrm{m}$

c $v^2 = u^2 + 2as$

$v^2 = 0^2 + (2 \times 18 \times 200)$

$v^2 = 7200$

$v = 85\,\mathrm{m\,s^{-1}}$

d $85\,\mathrm{m\,s^{-1}} = (85 \times 3600) = 3.1 \times 10^5\,\mathrm{m\,h^{-1}}$

$3.1 \times 10^5\,\mathrm{m\,h^{-1}} = \frac{3.1 \times 10^5}{10^3} = 310\,\mathrm{km\,h^{-1}}$

2 A train travelling at $50\,\mathrm{m\,s^{-1}}$ ($180\,\mathrm{km\,h^{-1}}$) needs to decelerate uniformly so that it stops at a station two kilometres away.

a What is the necessary deceleration?

b How long does it take to stop the train?

a $v^2 = u^2 + 2as$

$0^2 = 50^2 + (2 \times a \times 2000)$

$a = \dfrac{-50^2}{2 \times 2000}$

$a = -0.63\,\text{m s}^{-2}$

b $v = u + at$

$0 = 50 + (-0.63) \times t$

$t = \dfrac{50}{0.63} = 80\,\text{s}$

(Alternatively, you could use $s = \dfrac{u+v}{2}t$)

2.1.5 Solve problems involving the equations of uniformly accelerated motion.

Assume all accelerations are constant.

1 A ball rolls down a slope with a constant acceleration. When it passes a point P its velocity is $1.2\,\text{m s}^{-1}$ and a short time later it passes point Q with a velocity of $2.6\,\text{m s}^{-1}$.
   a What was its average velocity between P and Q?
   b If it took 1.4 s to go from P to Q, what is the distance PQ?
   c What is the acceleration of the ball?

2 A plane accelerates from rest along a runway and takes off with a velocity of $86.0\,\text{m s}^{-1}$. Its acceleration during this time is $2.40\,\text{m s}^{-2}$.
   a What distance along the runway does the plane travel before take-off?
   b How long after starting its acceleration does the plane take off?

3 An ocean-going oil tanker can decelerate no faster than $0.0032\,\text{m s}^{-2}$.
   a What is the minimum distance needed to stop if the ship is travelling at 10 knots? (1 knot = $0.514\,\text{m s}^{-1}$)
   b How much time does this deceleration require?

4 An advertisement for a new car states that it can travel 100 m from rest in 8.2 s.
   a What is the average acceleration?
   b What is the speed of the car after this time?

5 A car travelling at a constant velocity of $21\,\text{m s}^{-1}$ (faster than the speed limit of $50\,\text{km h}^{-1}$) passes a stationary police car. The police car accelerates after the other car at $4.0\,\text{m s}^{-2}$ for 8.0 s and then continues with the same velocity until it catches up with the other car.
   a When did the two cars have the same velocity?
   b Has the police car caught up with the other car after 10 s?
   c By equating two equations for the same distance at the same time, determine exactly when the police car overtakes the other car.

6 A car brakes suddenly and stops 2.4 s later, after travelling a distance of 38 m.
   a What was its deceleration?
   b What was the velocity of the car before braking?

7 A spacecraft travelling at $8.00\,\text{km s}^{-1}$ accelerates at $2.00 \times 10^{-3}\,\text{m s}^{-2}$ for 100 hours.
   a What is its final speed?
   b How far does it travel during this acceleration?

Acceleration due to gravity

2.1.4 Identify the acceleration of a body falling in a vacuum near the Earth's surface with the acceleration g of free fall.

We are all familiar with the motion of objects falling towards Earth because of the force of gravity. Figure 2.4 shows an experiment to gather data on distances and times for a falling mass, so that a value for its acceleration can be calculated.

Suppose that when the mass fell 0.84 m the time was measured to be 0.42 s. The acceleration can be calculated as follows:

$s = ut + \frac{1}{2}at^2$

$0.84 = 0 + (0.5 \times a \times 0.42^2)$

$a = \dfrac{0.84}{0.088} = 9.5\,\text{m s}^{-2}$

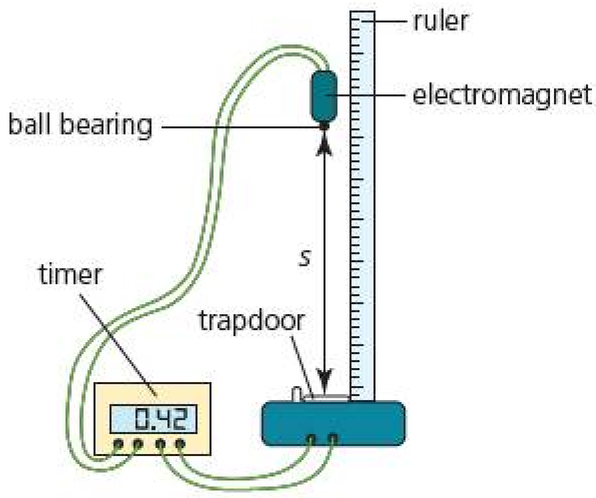


Figure 2.4 An experiment to measure the acceleration due to gravity

Table 2.1 Values of g in some cities around the world

City	g/m s^{-2}
Auckland	9.799
Bangkok	9.783
Buenos Aires	9.797
Cape Town	9.796
Chicago	9.803
Kuwait	9.793
London	9.812
Mexico City	9.779
Tokyo	9.798

If accurate measurements are made in a vacuum (to be sure that there is no air resistance), the results are very similar (but not identical) at all locations on the Earth's surface. Some examples are shown in Table 2.1.

The *acceleration due to gravity* in a vacuum near the Earth's surface is given the symbol g. This is the same as the *acceleration of free fall*. The accepted average value of g is 9.81 m s^{-2}. This value should be used in calculations and it is listed in the IB *Physics data booklet*. Anywhere on the Earth's surface (or in an airplane) can be considered as 'near to the Earth's surface'.

It is very important to remember that *all freely moving* objects close to the Earth's surface experience this same acceleration, g, downwards. This is true whether the object is large or very small, or whether it is moving upwards, downwards, sideways, or in any other direction. 'Freely moving' means that the effects of air resistance can be ignored and that the object is not powered in any way. In reality, however, the effects of air resistance usually cannot be ignored, except for large, dense masses moving short distances from rest. But, as is often the case in science, we need to understand simplified examples first, before we move on to more complicated situations.

Worked examples

3 A coin falls from rest out of an open window 16 m above the ground. Assuming that there is no air resistance:

a what is its velocity when it hits the ground?

b how long did it take to fall that distance?

a $v^2 = u^2 + 2as$

$v^2 = 0^2 + (2 \times 9.81 \times 16) = 314$

$v = 18\text{ m s}^{-1}$

b $v = u + at$

$18 = 0 + 9.81t$

$t = \frac{18}{9.81} = 1.8\text{ s}$

4 A ball is thrown vertically upwards and reaches a maximum height of 21.4 m.

a Calculate the speed with which the ball was released.

b What assumption did you make?

c Where will the ball be 3.05 s after it was released?

d What will its velocity be at this time?

a $v^2 = u^2 + 2as$

When the ball has travelled a distance, $s = 21.4$ m, its speed, v, at the highest point will be zero.

$0^2 = u^2 + (2 \times -9.81 \times 21.4)$

$u^2 = 419.9$

$u = 20.5\text{ m s}^{-1}$

In this example, the vector quantities directed upwards (u, v, s) are considered positive and the quantity directed downwards (a) is negative. The same answer would be obtained by reversing all the signs. Using positive and negative signs to represent vectors (like displacement, velocity and acceleration) in opposite directions is common practice.

b It was assumed that there was no air resistance.

c $s = ut + \frac{1}{2}at^2$

$s = (20.5 \times 3.05) + \left(\frac{1}{2} \times -9.81 \times 3.05^2\right)$

$s = 16.9$ m above the ground

d $v = u + at$

$v = 20.5 + (-9.81 \times 3.05)$

$v = -9.42\text{ m s}^{-1}$ (moving downwards)

2.1.5 Solve problems involving the equations of uniformly accelerated motion.

In all of the following questions, ignore the possible effects of air resistance. Use $g = 9.81\,m\,s^{-2}$.

8 **a** How long does it take a stone dropped from rest from a height of 2.1 m to reach the ground?
b If the stone was thrown downwards with an initial velocity of $4.4\,m\,s^{-1}$, with what speed would it hit the ground?
c If the stone was thrown vertically upwards with an initial velocity of $4.4\,m\,s^{-1}$, with what speed would it hit the ground?

9 A small rock is thrown vertically upwards with an initial velocity of $22\,m\,s^{-1}$. When will its speed be $10\,m\,s^{-1}$? (There are two possible answers.)

10 A falling ball has a velocity of $12.7\,m\,s^{-1}$ as it passes a window 4.81 m above the ground. When will it hit the ground?

11 A ball is thrown vertically upwards with a speed of $18.5\,m\,s^{-1}$ from a window that is 12.5 m above the ground.
a When will it pass the same window moving down?
b With what speed will it hit the ground?
c How far above the ground was the ball after exactly 2 s?

12 Two balls are dropped from rest from the same height. If the second ball is released 0.750 s after the first, and assuming they do not hit the ground, how far apart are the two balls:
a 3.00 s after the second ball was dropped?
b 2.00 s later?

13 A stone is dropped from rest from a height of 34 m. Another stone is thrown downwards 0.5 s later. If they both hit the ground at the same time, what was the initial velocity of the second stone?

14 In Worked example 1 an acceleration of $18\,m\,s^{-2}$ was quoted for a Formula One racing car. The driver of that car could be said to experience a 'g-force' of nearly $2g$, and during the course of a typical race a driver may have to undergo g-forces of nearly $5g$. Explain what you think is meant by a 'g-force' of $2g$.

15 Stone A is dropped from rest from a cliff. After it has fallen 5 m, stone B is dropped.
a How does the distance between the two stones change (if at all) as they fall?
b Explain your answer.

16 **a** A flea accelerates at the enormous average rate of $1500\,m\,s^{-2}$ during a vertical take-off that lasts only about 0.0012 s. What height will the jump reach?
b Measure how high you can jump vertically (standing in the same place), and use the result to calculate your take-off velocity.
c In order to jump up you had to bend your knees and reduce your height. Measure by how much your height was reduced just before jumping, then use the result to calculate your average acceleration during take off.
d What was the duration of your take-off?
e Compare your performance to the flea's.

17 Use the Internet to learn more about the GOCE project (Figure 2.5).

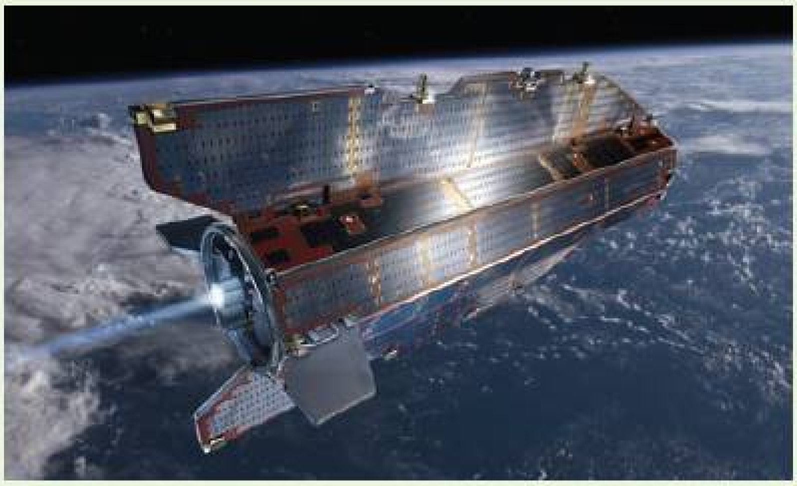


Figure 2.5 The Gravity Field and Steady-State Ocean Circulation Explorer (GOCE) satellite was launched by the European Space Agency in 2009 to measure the Earth's gravitational field in great detail

18 Figure 2.6 shows the tallest building in the world: Burj Khalifa in Dubai.
a How long would it take an object to reach the ground if it was dropped from 828 m (the height of Burj Khalifa)?
b What would the speed of impact with the ground be?

Figure 2.6 Burj Khalifa in Dubai

The effects of air resistance on falling objects

2.1.6 Describe the effects of air resistance on falling objects.

As an object moves through the air, the air is forced to move out of the path of the object. This causes a force opposing the motion called **air resistance** or **drag**.

Figure 2.7 represents the motion of an object falling towards Earth. Line A shows the motion without air resistance and line B shows the motion, more realistically, with air resistance.

When the object first starts to fall, there is no air resistance. The acceleration, g, is the same as if it was in a vacuum. As the object falls faster, the air resistance increases, so that the rate of increase in velocity becomes less. This is shown in the Figure 2.7 by the line B becoming less steep. Eventually the object reaches a constant, maximum speed known as the **terminal speed** or **terminal velocity** ('terminal' means final). The value of an object's terminal speed will depend on its cross-sectional area, shape and weight, as discussed in Section 2.2. The terminal speed of a skydiver is usually quoted at about 200 km h^{-1} (56 m s^{-1}) (Figure 2.8).

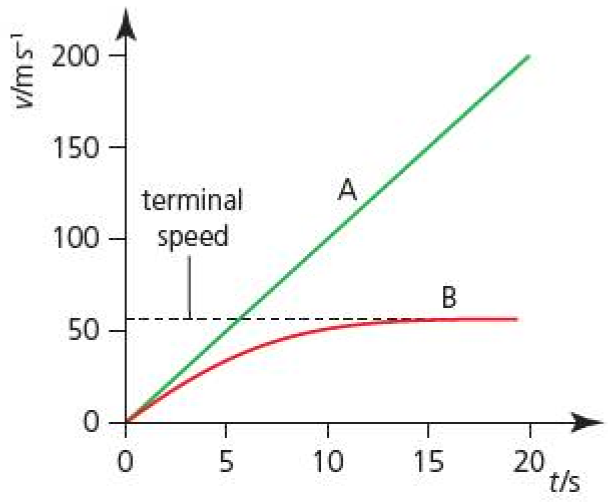


Figure 2.7 An example of a graph of velocity against time for an object falling under the effect of gravity, with and without air resistance

Figure 2.8 Skydivers in free fall

■ Additional Perspectives

Galileo

It is a matter of common observation that 'heavier objects fall to Earth quicker than lighter objects'. This is easily demonstrated by dropping, for example, a ball and a piece of paper side by side. The understandable belief that heavier objects fall faster was a fundamental principle in 'natural philosophy' (the name for early studies of what is now known as science) for more than 2000 years of civilization. In ancient Greece, Aristotle had closely linked the motion of falling objects to the belief that all processes have a purpose and that the Earth was the natural and rightful resting place for everything.

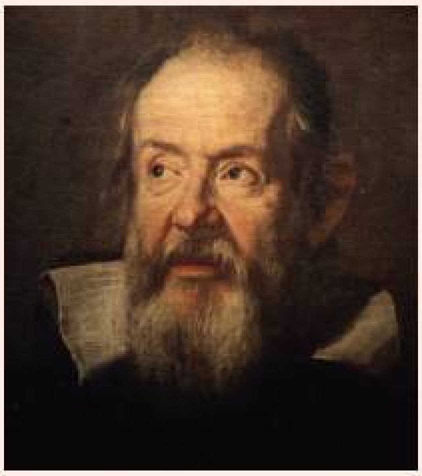


Figure 2.9 Galileo Galilei

In the 16th century the Italian scientist Galileo (Figure 2.9) was among the first to suggest that the reason why various objects fall differently was only because of air resistance. He predicted that, if the experiment could be repeated in a vacuum (without air), all objects would have exactly the same pattern of downwards motion under the effects of gravity.

In one of the most famous stories in science, Galileo dropped different masses off a balcony on the Tower of Pisa in Italy to show to those watching on the ground below that falling objects are acted on equally by gravity. This story may or may not be true, but one of the reasons that Galileo is so respected as a great scientist is that he was one of the first to actually do experiments, rather than just think about them.

It was many years later, after the invention of the first vacuum pumps, that Isaac Newton and others were able to remove the effects of air resistance and demonstrate that a coin (a 'guinea') and a feather fall together.

In 1971 that famous experiment was repeated on the Moon (Figure 2.10) when astronaut David Scott dropped a hammer and feather side by side. Millions of people from all over the world were watching while he reminded them of Galileo's achievements. The strength of gravity is less on the Moon than on the Earth because the Moon is smaller. Objects accelerate towards the Moon at about $\frac{1}{6}$ of the rate that they would on the Earth ($g = 1.6\,\mathrm{m\,s^{-2}}$).

Figure 2.10 Objects falling on the Moon

Question

1 Galileo's achievements were specifically mentioned when the experiment was repeated on the Moon, but do you think that there were there other scientists who were equally deserving of credit for advancing understanding of motion and gravity? Give the names of two such pioneers of science and list their greatest achievements.

TOK Link: What is science?

It has been suggested that Galileo was one of the first practising scientists (in the modern meaning of the word). But what, exactly, is science, and what makes science different from other human activities?

This is not an easy question to answer in a few words, although there are certainly important characteristics that most scientific activities share: science attempts to see some underlying simplicity in the vast complexity around us; science looks for the logical patterns and rules that control events; science seeks to accumulate knowledge and, wherever possible, to build on existing knowledge to make an ever-expanding framework of understanding.

Most importantly, science is based on experimentation and evidence, that is, science relies on 'facts' that are, at the current time, accepted to be 'true'. No good scientist would ever claim that something must be absolutely 'true' for all time, and one of the leading characteristics of science is the constant independent and widespread testing of existing theories by experiment. No fact or theory can ever be proven to be true for all times and all places, so science often advances through experiments that try to disprove new theories or existing knowledge.

The question 'what is science?' is often answered by explaining how scientists work, the so-called 'scientific method', which can be summarized as follows, although any particular scientific process may show variations from this generalized pattern.

- Choose a topic for investigation (for example, the design of golf balls – Figure 2.11).
- Research available information on the chosen topic (maybe use the Internet to find out about the design of golf balls).
- Ask a suitable question for investigation (for example, would a larger golf ball travel further than a smaller golf ball, if struck in the same way?)
- Use theory to predict what you think will happen in the investigation (for example, you might think that a smaller ball has less air resistance and so will go further).
- Design and carry out an investigation to test your prediction.
- Process the results and evaluate their uncertainties.
- Draw conclusions, accepting or rejecting your predictions.
- If the conclusions are unsatisfactory, repeat them and/or redesign the investigation.
- If the conclusions are satisfactory and can be repeated, present your findings to other people.

Figure 2.11 Why are golf balls a certain size?

Question

1 A well-known old Chinese proverb says 'I hear and I forget, I see and I remember, I do and I understand'. Consider your own knowledge of physics. To what extent has doing experimental work improved your understanding? Do you think doing more experimental work (and less theoretical work) would improve: **a** your interest, and **b** your examination results? Explain your answer.

Graphs of motion

2.1.7 **Draw** and analyse distance–time graphs, displacement–time graphs, velocity–time graphs and acceleration–time graphs.
2.1.8 **Calculate** and **interpret** the gradients of displacement–time graphs and velocity–time graphs, and the areas under velocity–time graphs and acceleration–time graphs.

The equations of motion are very useful models for describing and predicting the motion of an object moving with uniform acceleration, but graphs can also be drawn to represent any motion and they can provide an extra understanding and insight (at a glance) that very few of us can get by looking at equations. Furthermore, the gradients of graphs and the areas under graphs often provide additional valuable information.

Displacement–time graphs and distance–time graphs

A displacement–time graph shows how the displacement of an object from a reference point varies with time. All the graphs in Figure 2.12 are straight lines and can be described as representing **linear** relationships.

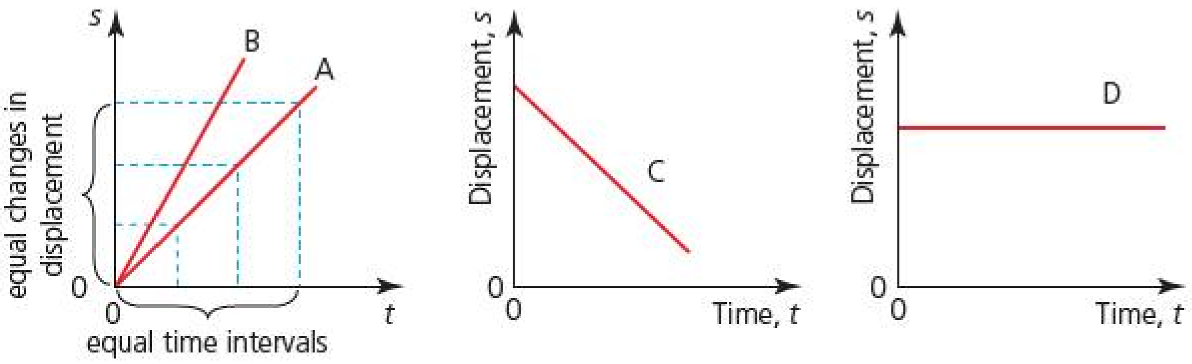


Figure 2.12 Constant velocities on displacement–time graphs (s–t graphs)

- Line A represents an object moving away from a reference point such that equal displacements occur in equal times. That is, the object has a constant velocity. *Any* linear displacement–time graph represents a constant velocity (it does not need to start or end at the origin).
- Line B represents an object moving with a greater velocity than A.
- Line C represents an object that is moving closer to the reference point.
- Line D represents an object that is stationary (at rest). It has zero velocity and stays at the same distance from the reference point.

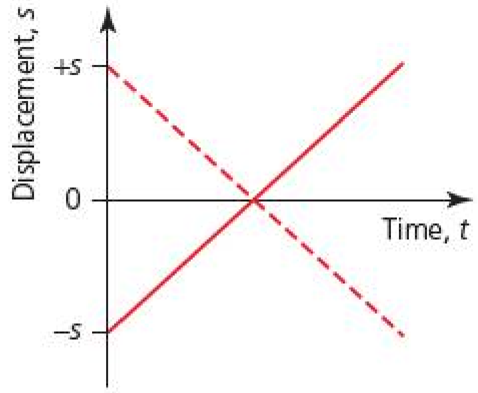


Figure 2.13 Motion in opposite directions represented on a displacement–time graph

Displacement is a vector quantity, but displacement–time graphs like these are usually used in situations where the motion is in a known direction, so that the direction may not need to be stated again. Displacement in opposite directions is represented by the use of positive and negative values. This is shown in Figure 2.13, in which the solid line represents the motion of an object moving with a constant (positive) velocity. The object moves towards a reference point (when the displacement is zero), passes it, and then moves away in the opposite direction with the same velocity. The dotted line represents an identical motion in the opposite direction (or it could also represent the original motion if the directions chosen to be positive and negative were reversed).

Any curved (non-linear) line on a displacement–time graph represents a changing velocity, in other words, an acceleration (or deceleration). This is illustrated in Figure 2.14.

- Figure 2.14a shows motion *away* from a reference point. Line A represents an object accelerating. Line B represents an object decelerating (negative acceleration).
- Figure 2.14b shows motion *towards* a reference point. Line C represents an object accelerating. Line D represents an object decelerating (negative acceleration).

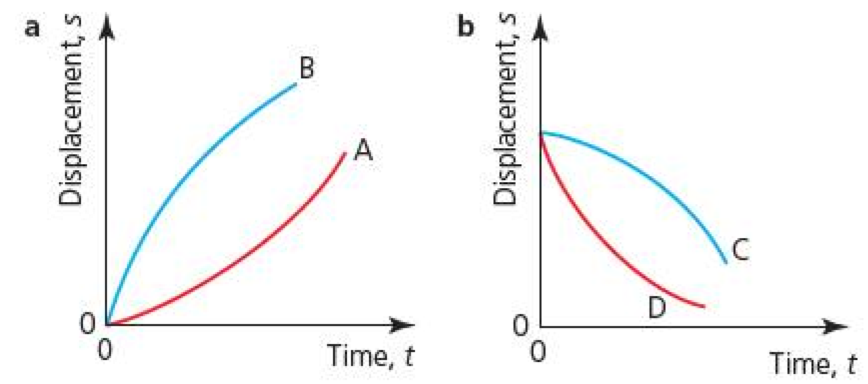


Figure 2.14 Accelerations on displacement–time graphs

The values of the accelerations represented by these graphs may or may not be constant (this cannot be determined without a more detailed analysis).

In physics, we are usually more concerned with displacement–time graphs than distance–time graphs. In order to explain the difference, consider Figure 2.15. Figure 2.15a shows a displacement–time graph for an object thrown vertically upwards with a speed of $20\,\text{m s}^{-1}$ without air resistance. You should be able to confirm by calculation that it takes 2 s to reach a maximum height of 20 m. At that point it has an instantaneous velocity of zero, before returning to where it began after 4 s and regaining its initial speed. Figure 2.15b shows how the same motion would appear on an overall distance–time graph.

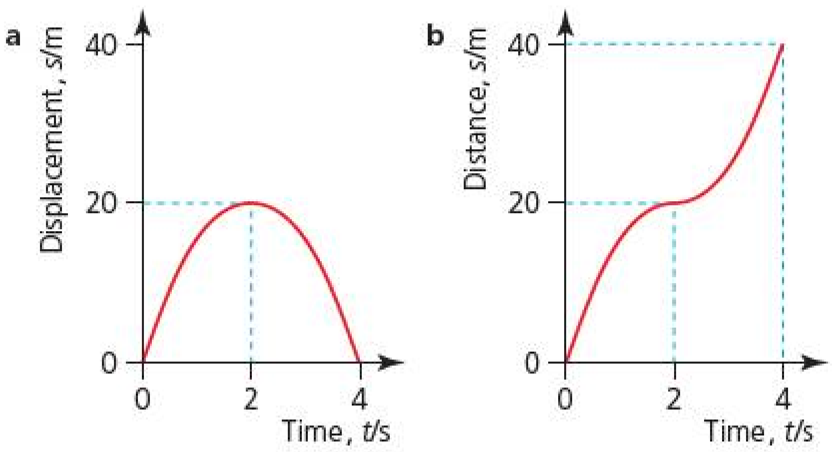


Figure 2.15 Displacement–time and distance–time graphs for an object moving up and then down

Gradients of displacement–time graphs

Consider the motion at constant velocity shown in Figure 2.16.

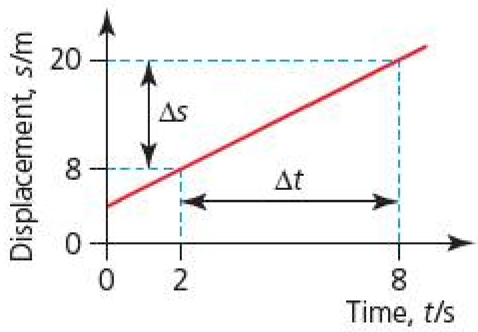


Figure 2.16 Finding a constant velocity from a displacement–time graph

From the graph, the velocity v is given by:

$$v = \frac{\Delta s}{\Delta t} = \frac{20 - 8.0}{8.0 - 2.0} = 2.0\,\text{m s}^{-1}$$

Note that the velocity is numerically equal to the gradient (slope) of the line. This is always true, whatever the shape of the line.

The instantaneous velocity of an object is equal to the gradient of the displacement–time graph at that instant.

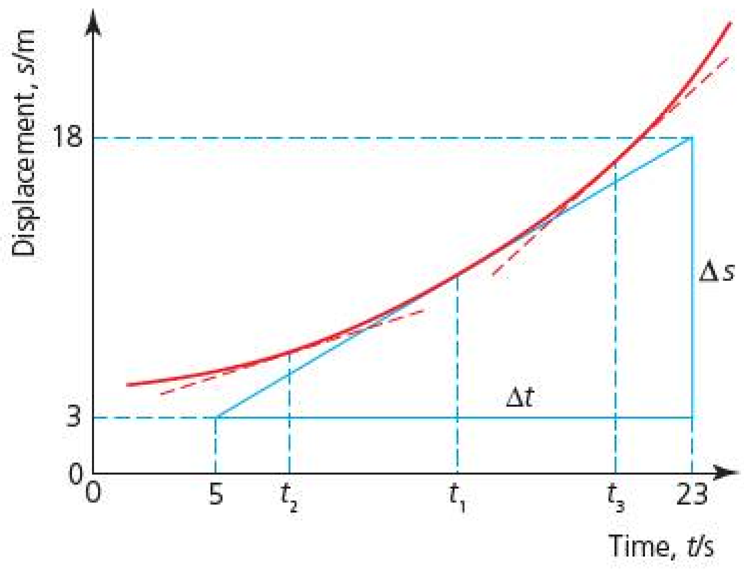


Figure 2.17 Finding an instantaneous velocity from a displacement–time graph

Figure 2.17 shows an object moving with increasing velocity. The velocity at any time (for example t_1) can be determined by calculating the gradient of the tangent to the line at that instant. The triangle used should be large, in order to make this process as accurate as possible. The tangent drawn at time t_2 has a smaller gradient because the velocity is smaller. At time t_3 the velocity is greater and the gradient steeper. So, in this example:

$$\text{velocity at } t_1 = \frac{18 - 3.0}{23 - 5.0} = \frac{15}{18} = 0.83\,\text{m s}^{-1}$$

19 Figure 2.18 represents the motion of a train on a straight track between two stations.
- **a** Describe the motion.
- **b** How far apart are the stations?
- **c** Calculate the maximum speed of the train.
- **d** What was the average speed of the train between the two stations?

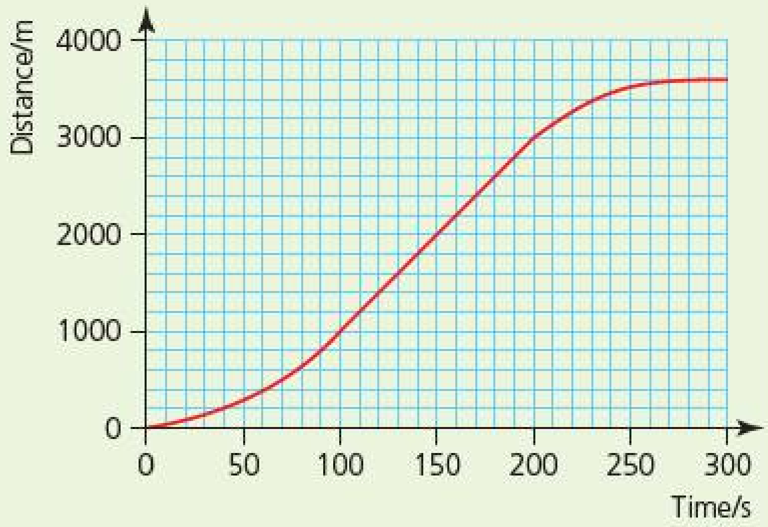


Figure 2.18

20 **a** Draw a displacement–time graph for a swimmer who swims 50 m at a constant speed of $1.0\,\text{m s}^{-1}$ if the swimming pool is 25 m long and the swimmer takes 1 s to turn around half way through the race.
- **b** Find out the average speed of the world freestyle record holder when the 100 m record was last broken.
- **c** The world record for swimming 50 m in a pool of length 25 m is shorter than for swimming in a pool of length 50 m. Suggest why.

21 Draw a displacement–time graph for the following motion: a stationary car is 25 m away; 2 s later it starts to move further away in a straight line from you with a constant acceleration of $1.5\,\text{m s}^{-2}$ for 4 seconds; then it continues with a constant velocity for another 8 s.

22 Describe the motion of the runner shown by the graph in Figure 2.19.

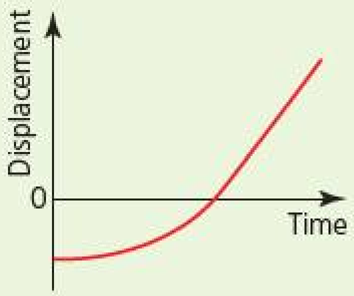


Figure 2.19

23 **a** Describe the motion represented by the graph in Figure 2.20.
- **b** Compare the velocities at points A and B.
- **c** When is the object moving with its maximum and minimum velocities?
- **d** Estimate values for the maximum and minimum velocities.
- **e** Suggest what kind of object could move in this way.

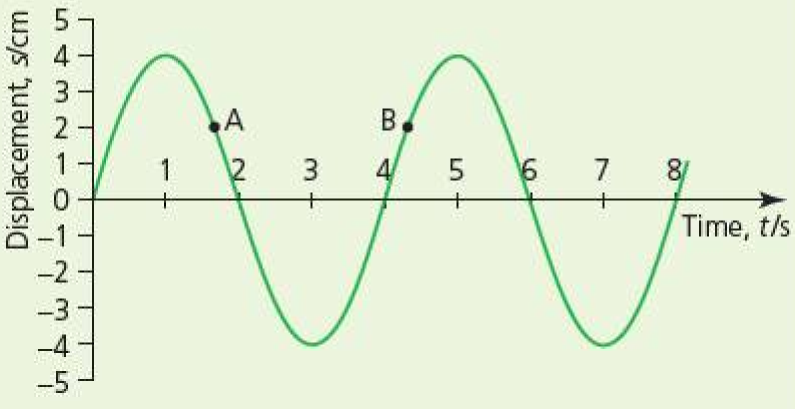


Figure 2.20

Velocity–time graphs

A velocity–time graph shows how the velocity of an object varies with time. *Any* straight (linear) line on any velocity–time graph shows that equal changes of velocity occur in equal times, that is, a constant acceleration.

In Figure 2.21:

- Line A shows an object which has a constant positive acceleration.
- Line B represents an object moving with a greater positive acceleration than A.
- Line C represents an object that is decelerating (negative acceleration).
- Line D represents an object moving with a constant velocity, that is, it has zero acceleration.

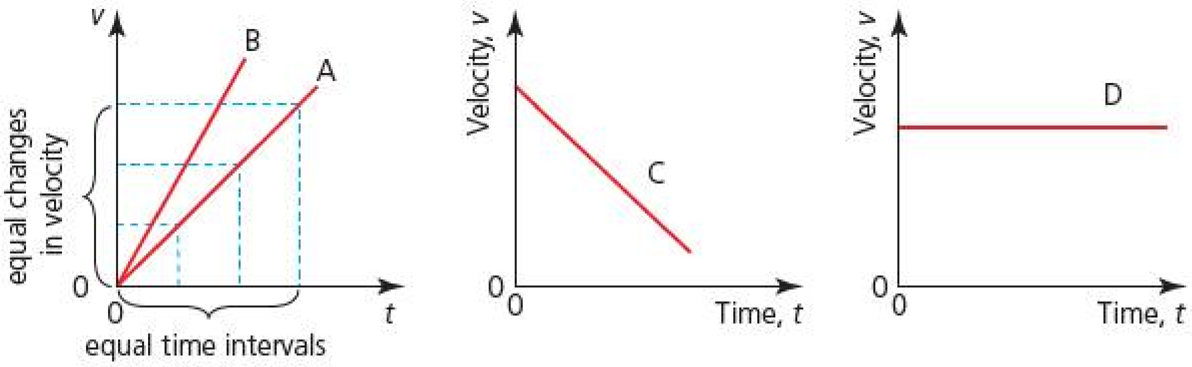


Figure 2.21 Constant accelerations on velocity–time graphs

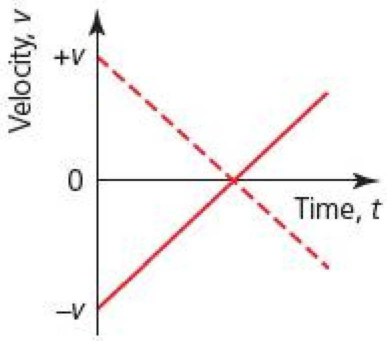


Figure 2.22 Velocities in opposite directions

Curved lines on velocity–time graphs represent changing accelerations. Velocities in opposite directions are represented by positive and negative. The solid line in Figure 2.22 represents an object that decelerates uniformly to zero velocity and then moves in the opposite direction with the same acceleration. This graph could represent the motion of a stone thrown in the air, reaching its maximum height and then falling down again. The acceleration remains the same throughout ($9.81\,m\,s^{-2}$ downwards). In this example velocity and acceleration upwards have been chosen to be negative and velocity and acceleration downwards are positive. The dashed line would represent exactly the same motion if the directions chosen to be positive and negative were reversed.

Gradients of velocity–time graphs

Consider the motion at constant acceleration shown in Figure 2.23.

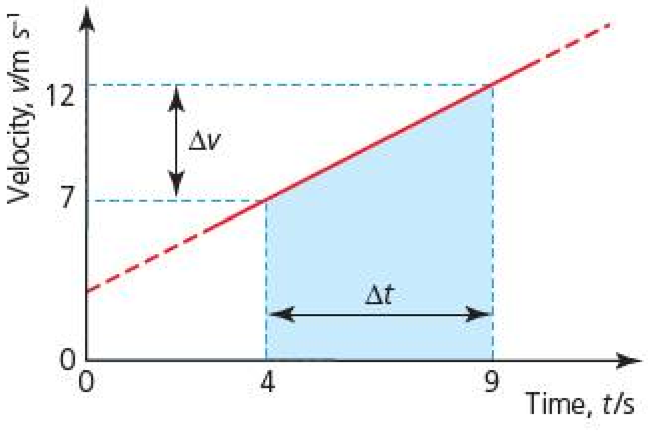


Figure 2.23 Finding the gradient of a velocity–time graph

From the graph:

$$\text{acceleration, } a = \frac{\Delta v}{\Delta t} = \frac{12 - 7.0}{9.0 - 4.0} = 1.0\,m\,s^{-2}$$

Note that the acceleration is numerically equal to the gradient (slope) of the line. This is always true, whatever the shape of the line.

The instantaneous acceleration of an object is equal to the gradient of the velocity–time graph at that instant.

Worked example

5 Figure 2.24 shows an object decelerating (with a decreasing negative acceleration). Use the graph to find the instantaneous acceleration at 10 s.

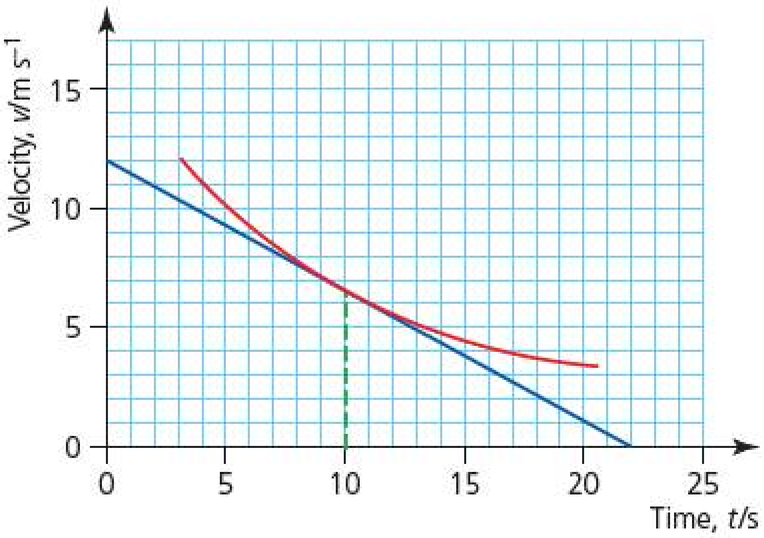


Figure 2.24 Finding an instantaneous acceleration from a velocity–time graph

A tangent drawn at a time of 10 s can be used to determine the value of the acceleration at that instant:

$$a = \frac{\Delta v}{\Delta t} = \frac{0 - 12}{22 - 0} = -0.55\,\text{m}\,\text{s}^{-2}$$

In this example the large triangle used to determine the gradient accurately was extended to the axes for convenience.

Areas under velocity–time graphs

Consider again the motion represented in Figure 2.23. The change of displacement, s, between the fourth and ninth second can be found from (average velocity) × time.

$$s = \left(\frac{12 + 7.0}{2}\right) \times (9.0 - 4.0) = 48\,\text{m}$$

This is numerically equal to the area under the line between $t = 4.0$ s and $t = 9.0$ s (as shaded in Figure 2.25). This is always true, whatever the shape of the line.

The area under a velocity–time graph is equal to the change of displacement.

Worked example

6 Figure 2.25a shows how the velocity of a car changed in the first 5 s after starting. Use the graph to estimate the distance travelled in this time.

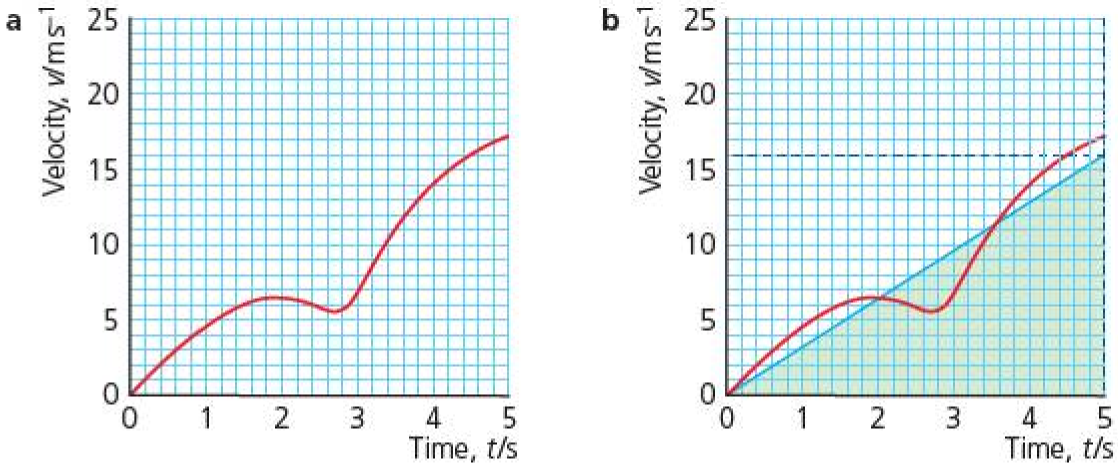


Figure 2.25 Determining the displacement of a car during acceleration

In Figure 2.25b the blue line has been drawn so that the area under it and the area under the original line are the same (as judged by eye).

distance = area under graph $= \frac{1}{2} \times 16 \times 5.0 = 40\,\text{m}$

24 **a** Describe the motion represented by the graph in Figure 2.26.
b Calculate accelerations in the three parts of the journey.
c What was the total distance travelled?
d What was the average speed?

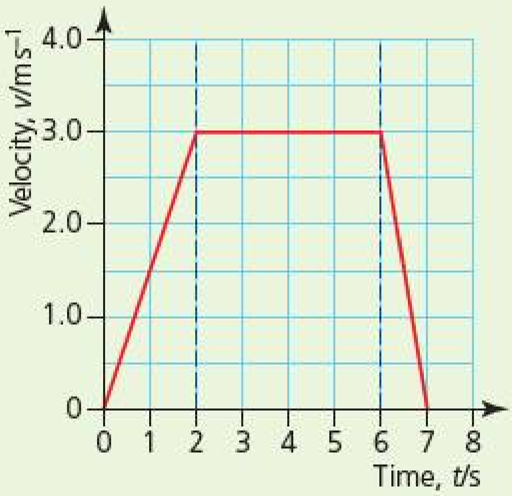


Figure 2.26

25 The velocity of a car was read from its speedometer at the moment it started and every 2 s afterwards. The successive values ($m\,s^{-1}$) were: 0, 1.1, 2.4, 6.9, 12.2, 18.0, 19.9, 21.3 and 21.9. Plot a graph of these readings and use it to estimate the maximum acceleration and the distance covered in 16 s.

26 **a** Describe the motion of the object represented by the graph in Figure 2.27.
b Calculate the acceleration during the first 8 s.
c What was the total distance travelled in 12 s?
d What was the total displacement after 12 s?
e What was the average speed during the 12 s interval?

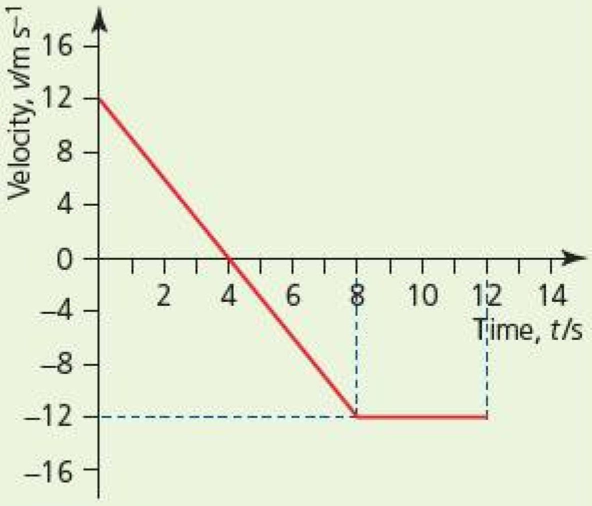


Figure 2.27

27 Sketch a velocity–time graph of the following motion: a car is 100 m away and travelling along a straight road towards you at a constant velocity of $25\,m\,s^{-1}$. Two seconds after passing you, the driver decelerates uniformly and the car stops 62.5 m away from you.

Additional Perspectives

100 m sprinters

Figure 2.28 Usain Bolt broke the world record for 100 m in a time of 9.58 seconds in Berlin in 2009

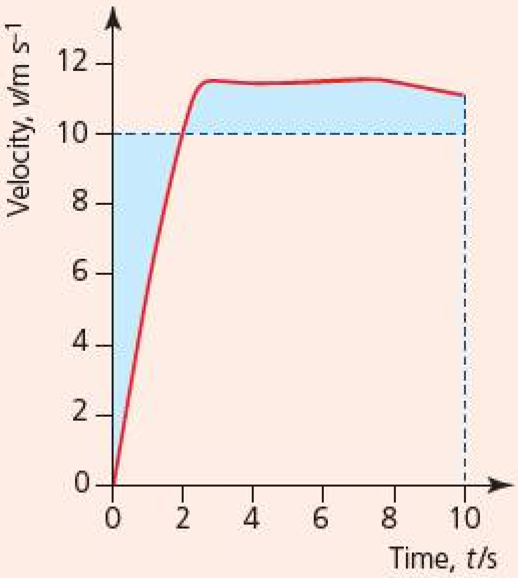


Figure 2.29 Velocity–time graph for an athlete running 100 m

World class sprinters can run 100 m in about 10 s (see Figure 2.28). The average velocity is easy to calculate, $v = 100/10 = 10\,m\,s^{-1}$. Clearly they start from $0\,m\,s^{-1}$, so their greatest velocity must be greater than $10\,m\,s^{-1}$. Coaches use the science of *biomechanics* to improve the athlete's techniques and the latest computerized methods are used to analyse every moment of their races. The acceleration 'off the blocks' at the start of the race is all important, so that the greatest velocity is reached as soon as possible. For the rest of the race the athlete should be able to maintain the same speed, although there may be a slight decrease towards the end of the race. Figure 2.29 shows a typical velocity–time graph for a 100 m race completed in 10 s.

Question

1 a Estimate the greatest acceleration achieved during the race illustrated in Figure 2.29.
b When does the athlete reach their greatest velocity?
c Explain why the two shaded areas on the graph are equal.
d Use the Internet to collect data to draw a graph showing how the world (or Olympic) record for the 100 m has changed in the last 100 years.
e Predict the 100 m record for the year 2040.

Acceleration–time graphs

An acceleration–time (a–t) graph shows how the acceleration of an object changes with time. In this chapter, we are mostly concerned with constant accelerations (it is less common to see motion graphs showing changing acceleration). The graphs in Figure 2.30 show five lines representing constant accelerations.

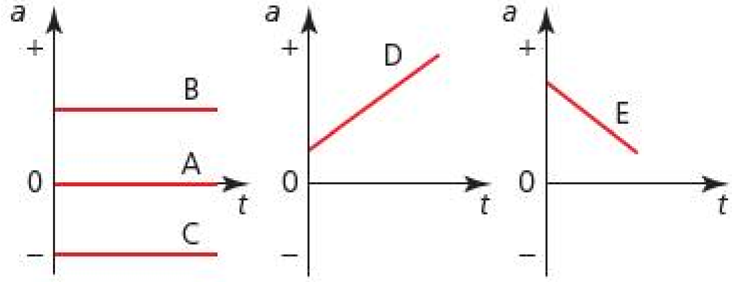


Figure 2.30 Graphs of constant acceleration

- Line A shows zero acceleration, constant velocity.
- Line B shows a constant positive acceleration (uniformly increasing velocity).
- Line C shows the constant negative acceleration (deceleration) of an object that is slowing down at a constant rate.
- Line D shows a (linearly) increasing positive acceleration.
- Line E shows an object that is accelerating, but at a (linearly) decreasing rate.

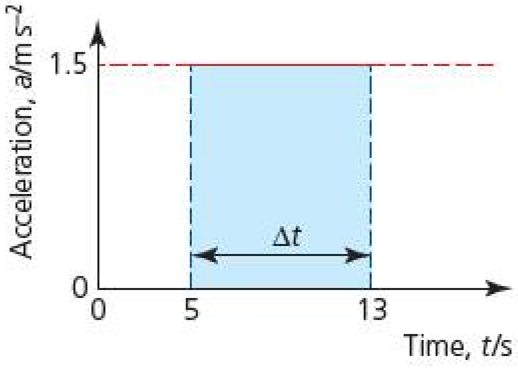


Figure 2.31 Calculating change of velocity from an acceleration–time graph

Areas under acceleration–time graphs

Figure 2.31 shows the constant acceleration of a moving car. Using $a = \Delta v/\Delta t$, between the fifth and thirteenth second the velocity of the car increases by $\Delta v = a\Delta t = 1.5 \times 8.0 = 12\,\text{m s}^{-1}$.

The change in velocity is numerically equal to the area under the line between $t = 5$ s and $t = 13$ s (shaded in Figure 2.31). This is always true, whatever the shape of the line.

The area under an acceleration–time graph is equal to the change of velocity.

28 Draw an acceleration–time graph for a car that starts from rest, accelerates at $2\,\text{m s}^{-2}$ for 5 s, then travels at constant velocity for 8 s, before decelerating uniformly to rest again in a further 2 s.

29 Figure 2.32 shows how the acceleration of a car changed during a 6 s interval. If the car was travelling at $2\,\text{m s}^{-1}$ after 1 s, estimate a suitable area under the graph and determine the approximate speed of the car after 5 s.

30 Figure 2.33 shows a tennis ball being struck by a racquet. Sketch a possible velocity–time graph and an acceleration–time graph from 1 s before impact to 1 s after the impact.

31 Sketch displacement–time, velocity–time and acceleration–time graphs for a bouncing ball dropped from rest. Continue the sketches until the third time that the ball contacts the ground.

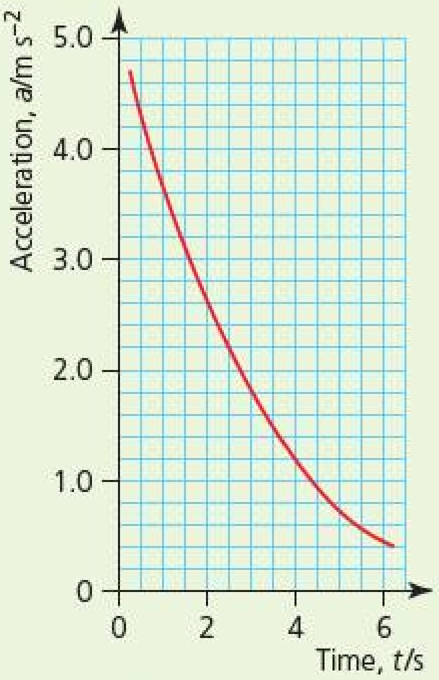


Figure 2.32

Figure 2.33 Striking a tennis ball

Graphs of motion: summary

When discussing the equations of motion for uniform acceleration, it was made clear that if any three of the five quantities u, v, a, s, and t are known, then the motion is completely defined and the two other (unknown) quantities can be calculated. In a similar way, if one graph of motion is plotted (s–t, v–t or a–t), then the motion is fully defined and the other two graphs can be drawn with information about gradients and/or areas taken only from the first graph. This is summarized in Figure 2.34.

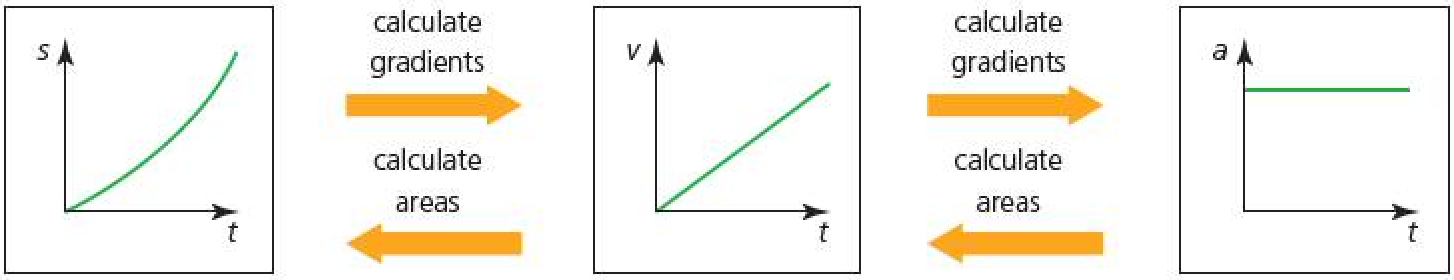


Figure 2.34 The connections between the different graphs of motion

To reproduce one graph from another by hand is a long and repetitive process, because in order to produce accurate graphs a large number of similar measurements and calculations need to be made over short intervals of time. Of course, computers are ideal for this purpose.

In more mathematically advanced work, which is not part of this course, **calculus** can be used to perform these processes using differentiation and integration.

Additional Perspectives

Vehicle braking distances

Figure 2.35 represents how the velocities of two identical cars changed from the moment that their drivers saw dangers in front of them and tried to stop the cars as quickly as possible. It has been assumed that both drivers have the same reaction time (0.7 s) and both cars decelerate at the same rate ($-5.0\,\mathrm{m\,s^{-2}}$).

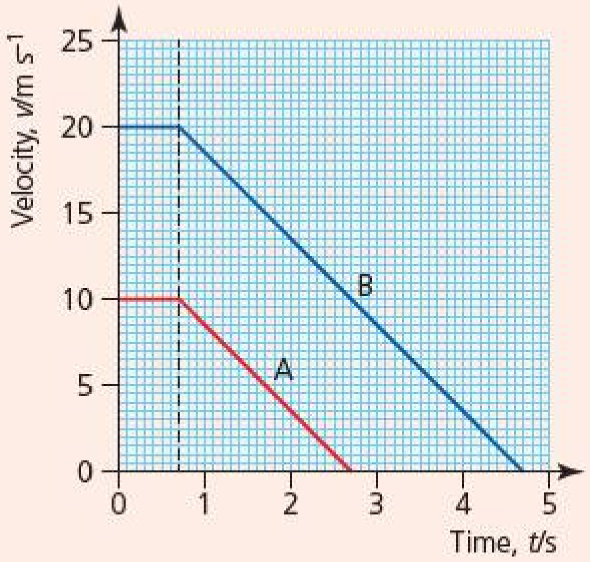


Figure 2.35 Velocity–time graphs for two cars braking

The distance travelled at constant velocity before the driver reacts and presses the brakes is known as the 'thinking' distance. The distance travelled while decelerating is called the 'braking distance'. The total stopping distance is the sum of these two distances.

Car B, travelling at twice the velocity of car A, has twice the thinking distance. That is, the thinking distance is proportional to the velocity of the car. The distance travelled when braking, however, is proportional to the velocity squared. This can be confirmed from the areas under the v–t graphs. The area under graph B is four times the area under graph A (during the deceleration). This has important implications for road safety and most countries make sure that people learning to drive must learn about how stopping distances change with the vehicle's velocity. Some countries measure the reaction times of people before they are given a driving licence.

Reaction times and thinking distances can vary considerably, depending on the condition of the driver and the circumstances inside the vehicle. Taking different kinds of drugs (including alcohol) can greatly increase reaction times, as can tiredness. Older people will generally react more slowly than younger people, and talking to other people and listening to music will normally distract a driver from concentrating on the road ahead.

Activity

Set up a spreadsheet that will calculate the total stopping distance for cars travelling at initial speeds, u, between 0 and $40\,\mathrm{m\,s^{-1}}$ with a deceleration of $-6.5\,\mathrm{m\,s^{-2}}$. (Make calculations every $2\,\mathrm{m\,s^{-1}}$.) The thinking distance can be calculated from $s_t = 0.7u$ (reaction time 0.7 s). In this example the braking time can be calculated from $t_b = u/6.5$ and the braking distance can be calculated from $s_b = (u/2)t_b$. Use the data produced to plot a computer-generated graph of stopping distance (y-axis) against initial speed (x-axis).

Relative velocity

2.1.9 Determine relative velocity in one and in two dimensions.

When we describe the velocity of a train, for example, as $8\,m\,s^{-1}$ to the east, there is no need to make it clear that we are comparing it to the Earth's surface and ignoring the motion of the Earth itself.

But there is no such thing as absolute rest, and *all* motion is relative (compared) to somewhere else, usually the 'observer' of the motion. In most cases any motion that we describe is relative to the Earth's surface, which may be called the 'frame of reference' of the motion.

However, sometimes our frame of reference is changed, most commonly if we are in a moving vehicle. Consider a simple example – if a man inside a moving train walks at $0.5\,m\,s^{-1}$ towards the front of the train, $0.5\,m\,s^{-1}$ is the velocity of the man *relative* to the train and not to the rest of the Earth. A stationary observer outside the train will have a different impression – if the train has a velocity of $8\,m\,s^{-1}$, then the man on the train will have a velocity of $8.5\,m\,s^{-1}$ in the same direction as the train is moving. If the man on the train walks back in the opposite direction, his velocity will be $7.5\,m\,s^{-1}$ as viewed by the observer outside the train.

Relative velocity is the *difference* between two velocities measured in the *same* frame of reference. One velocity is subtracted from the other.

Subtraction of vectors was covered in Chapter 1. Using vector notation, the velocity of A relative to B is therefore given by:

$$\boldsymbol{v}_{AB} = \boldsymbol{v}_A - \boldsymbol{v}_B$$

Figure 2.36 Two vehicles moving with different velocities

In Figure 2.36 a car (A) moving at $22\,m\,s^{-1}$ overtakes a truck (B) moving at $17\,m\,s^{-1}$ (both measured compared to the Earth's surface).

So, in the example illustrated by Figure 2.36, the velocity of A relative to B is:

$$\boldsymbol{v}_{AB} = 22 - 17 = +5\,m\,s^{-1}$$

The velocity of B relative to A is given by:

$$\boldsymbol{v}_{BA} = \boldsymbol{v}_B - \boldsymbol{v}_A$$

So, in the example illustrated by Figure 2.36:

$$\boldsymbol{v}_{BA} = 17 - 22 = -5\,m\,s^{-1}$$

The negative sign indicates that the truck seems to be moving backwards compared to the car.

If the car was travelling in the opposite direction to the truck, the equations would become (still using the direction of the truck to be positive):

$$\boldsymbol{v}_{AB} = -22 - 17 = -39\,m\,s^{-1}$$
$$\boldsymbol{v}_{BA} = 17 - (-22) = +39\,m\,s^{-1}$$

If the velocities are not along the same line, a graphical method can be used to find the relative velocity. As explained in Chapter 1, this is done by adding one vector to the negative of the other. (The negative of a vector has the same magnitude but opposite direction.)

Figure 2.37a shows the velocities of two boats, P and Q, as seen by a stationary external observer.

$$\boldsymbol{v}_{PQ} = \boldsymbol{v}_P - \boldsymbol{v}_Q = \boldsymbol{v}_P + (-\boldsymbol{v}_Q)$$

Figure 2.37b shows how the velocity of P relative to Q can be determined and Figure 2.37c shows the determination of the velocity of Q relative to P.

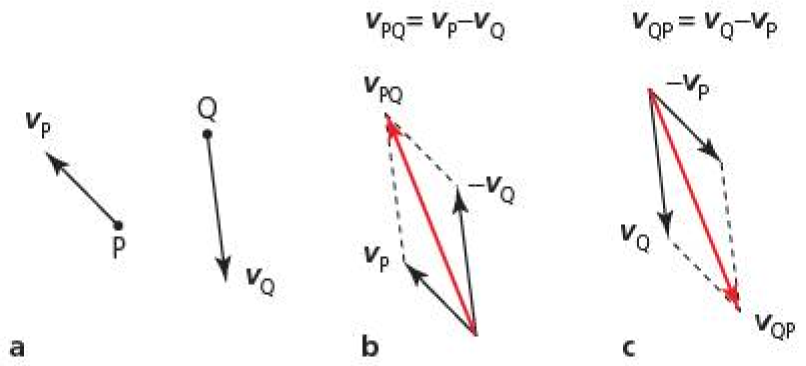


Figure 2.37 Determining relative velocities

If a ball is thrown vertically upwards in a vehicle travelling with constant velocity, such as a fast moving train or plane, it will be observed to fall straight back down to its point of release. This is because the ball and the person throwing it are both moving forward with the same velocity as the vehicle. We say that they are all in the same frame of reference. But an observer from outside, in a different frame of reference, would (also correctly) describe the motion of the ball differently; it would be seen to move forwards, rising and falling in a curved path.

32 Train A is moving at $34\,\text{m s}^{-1}$ to the west when it passes train B travelling at $16\,\text{m s}^{-1}$ in the opposite direction. What is the relative velocity of train B as seen by an observer on train A?

33 A girl runs at $4.2\,\text{m s}^{-1}$ through rain that is falling vertically with a speed of $21\,\text{m s}^{-1}$. Use trigonometry to determine the velocity of the rain relative to the girl.

34 **a** A plane approaching an airport moves through the air with an 'air speed' (speed relative to the air) of $290\,\text{km h}^{-1}$ at an angle of 38° to a headwind, air that is moving south with a speed of $50\,\text{km h}^{-1}$. Use a scale drawing to determine the velocity ('ground speed' and direction) that would be detected by the control tower at the airport.
b For the safe control of the plane in flight, does the pilot need to know the ground speed or the air speed?

35 An apple core is dropped carelessly out of the window of a bus.
a Briefly describe the motion of the apple core as seen by an observer on the bus and an observer at the side of the road.
b Which of these two different descriptions is correct?

2.2 Forces and dynamics

Forces

Figure 2.38 The upwards force on a rocket accelerates it into space

At its simplest, a force is a push or a pull. A force acting on an object (a *body*) may make it start to move (Figure 2.38) or change its motion if it is already moving. In other words, a force can change the velocity of an object. Accelerations are caused by forces.

Forces can also change the shape of the object. That is, a force can make an object become deformed in some way. For example, when we sit on a soft chair the **deformation** is easy to see, but when we sit on a hard chair, or stand on the floor, there is still a deformation, but it is usually too small to see.

Clearly, the effect of a force will depend on the direction in which it acts. Force is a **vector** quantity. Like all vectors, a force can be represented by drawing a (labelled) line of the correct length in the right direction (shown with an arrow).

The length of the line used to show a force should be proportional to the magnitude of the force. For example, in Figure 2.39 vector arrows represent the different weights of two people.

When discussing the forces *acting* on an object, we may alternatively talk about *applying* a force to an object, or *exerting* a force on an object.

The symbol F is used for force and the SI unit of force is the **newton**, N. One newton is defined as the force that makes 1 kg accelerate by $1\,\text{m s}^{-2}$.

Different types of force

2.2.2 (part) **Identify** the forces acting on an object.

Apart from obvious everyday pushes and pulls, we are surrounded by a number of different types of force. In the following section we will discuss the following types of force:

- weight
- tension and compression
- reaction forces
- friction and air resistance
- upthrust
- non-contact forces.

Weight

2.2.1 **Calculate** the weight of a body using the expression $W = mg$.

The **weight**, W, of a mass is the **gravitational force** that pulls it towards the centre of the Earth. Weight is related to mass by the following equation:

$$W = mg$$

W is the weight in newtons, m is the mass of the object in kilograms and g is the acceleration due to gravity in metres per second squared ($\mathrm{m\,s^{-2}}$). The larger the mass of the object, the greater its weight.

An alternative interpretation of g is as the ratio of weight to mass, $g = W/m$. Expressed in this way, it is known as the **gravitational field strength**, with the units newtons per kilogram, $\mathrm{N\,kg^{-1}}$ ($1\,\mathrm{N\,kg^{-1}} = 1\,\mathrm{m\,s^{-2}}$). A mass would weigh less on the Moon because the Moon has a smaller gravitational field strength than the Earth.

The average value of g on, or close to, the Earth's surface is accepted to be $9.81\,\mathrm{m\,s^{-2}}$, although it does vary, as we saw in Table 2.1. For quick approximations, a value of $g = 10\,\mathrm{m\,s^{-2}}$ is often used – which is only a 2% error. The value of g decreases as the distance from the centre of the Earth increases. For example, at a height of 300 km above the Earth's surface the value of g has decreased to $9.67\,\mathrm{m\,s^{-2}}$. This means that objects at that height, such as a satellite, or astronauts in orbit around the Earth, are *not* weightless (as is often believed), but they have a weight only slightly less than on the Earth's surface.

If we want to represent the weight of an object in a diagram, we use a vector arrow of an appropriate length drawn vertically downwards from the **centre of mass** of the object, as shown in Figure 2.39. The centre of mass of an object may be considered as the 'average' position of all of its mass. For symmetrical and uniform objects the centre of mass will be at the geometrical centre.

The mass of an object stays the same, wherever it is in the universe, but the gravitational force on an object (its weight) varies depending on its location. For example, the acceleration due to gravity (gravitational field strength) on the Moon is $1.6\,\mathrm{m\,s^{-2}}$ and on Mars it is $3.7\,\mathrm{m\,s^{-2}}$. The acceleration due to gravity is different on the Moon and Mars because they have different masses and sizes compared with the Earth. In deep space, a very long way from any star or planet, any object would be almost weightless.

Unfortunately, in everyday conversation the word 'weighing' is used for finding the mass (not weight) of an object, and the question 'what is the weight of that?' would usually be answered in kilograms not newtons. This is a common confusion that every student and teacher of physics has to face!

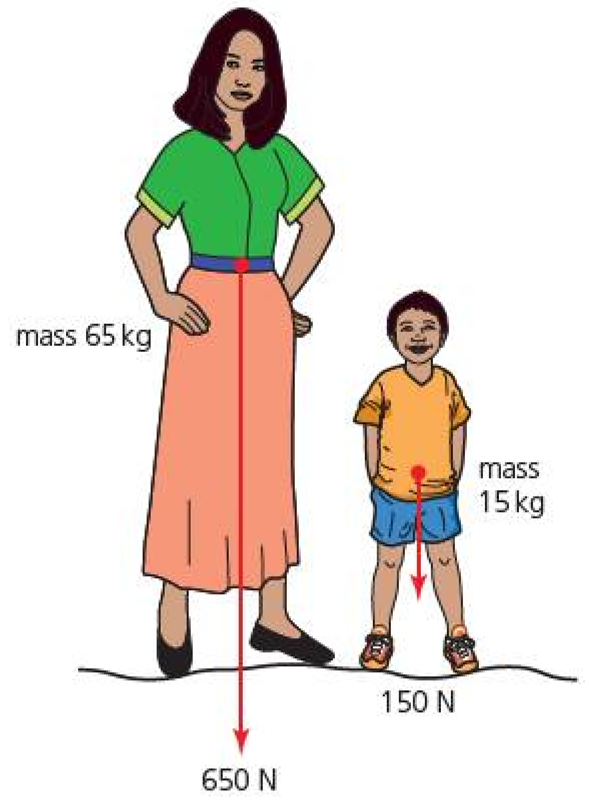


Figure 2.39 Weight acts downwards from the centre of mass

Worked example

7 An astronaut has a mass of 62.2 kg. What would her weight be in the following locations?

a on the Earth's surface
b in a satellite 300 km above the Earth
c on the Moon
d on Mars
e a very, very long way from any planet or star.

a $W = mg = 62.2 \times 9.81 = 610\,\text{N}$
b $W = 62.2 \times 9.67 = 601\,\text{N}$
c $W = 62.2 \times 1.6 = 100\,\text{N}$
d $W = 62.2 \times 3.7 = 230\,\text{N}$
e zero

36 Calculate the weight of the following objects on the surface of the Earth:
a a car of mass 1250 kg
b a new-born baby of mass 3240 g
c one pin in a pile of 500 pins that has a total mass of 124 g.

37 A girl has a mass of 45.9 kg. Use the data given in Table 2.1 to calculate the difference in her weight between Bangkok and London.

38 **a** It said that 'an A380 plane has a maximum take-off weight of 570 tonnes' (Figure 2.40). A tonne is the same as a mass of 1000 kg. What is the maximum weight of the plane (in newtons) during take-off?
b The plane can take a maximum of about 850 passengers. Estimate the total mass of all the passengers and crew. What percentage is this of the total mass of the plane on take-off?
c The maximum landing weight is '390 tonnes'. Suggest a reason why the plane needs to be less massive when landing than when taking off.
d Calculate the difference in mass and explain where the 'missing' mass has gone.

Figure 2.40 The Airbus A380 is the largest passenger airplane in the world

39 The weight of an object decreases very slightly as its distance above the Earth's surface increases. Suggest why the weight of an object may not increase if it is taken closer to the centre of the Earth down a mine shaft.

40 A mass of 50 kg would have a weight of 445 N on the planet Venus. What is the strength of the gravitational field there? Compare it with the value of g on Earth.

41 Consider two solid spheres made of the same metal. Sphere A has twice the radius of sphere B. Calculate the ratio of the two sphere's circumferences, surface areas, volumes, masses and weights.

Additional Perspectives

Weighing

Forces are most easily measured by the changes in length they produce when they squash or stretch a spring (or something similar). Such instruments are called force meters (also called newton meters or spring balances – see Figure 2.41). In this type of instrument the spring usually has a change of length proportional to the applied force. The length of the spring is shown on a linear scale, which can be calibrated (marked) in newtons. The spring goes back to its original shape after it has measured the force.

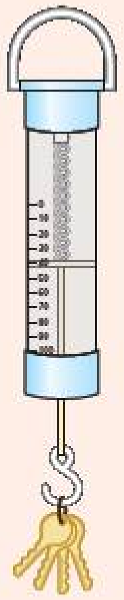


Figure 2.41 Force meter

Such instruments may be used for measuring forces acting in any direction, but they are also widely used for the measurement of weight. The other common way of measuring weight is with some kind of 'balance' (scales). In an equal arm balance, as shown in Figure 2.42, the beam will only balance if the two weights are equal. That is, the unknown weight equals the known weight.

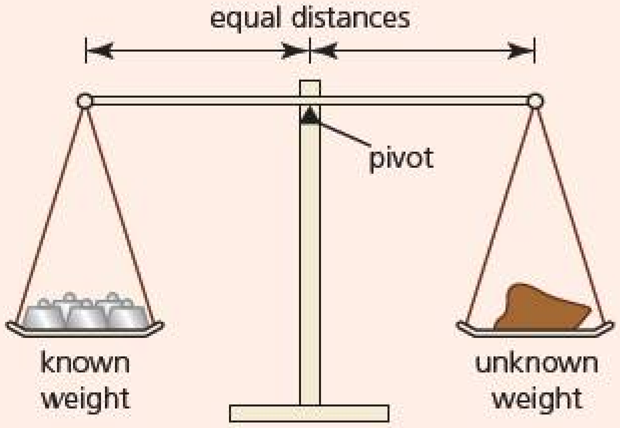


Figure 2.42 An equal arm balance

In this type of balance, the pivot can be moved closer to the unknown weight if it is much heavier than the known weight(s). The balance then has to be calibrated using the *principle of moments*. This principle is not part of the course, but it may be familiar to you from earlier work.

Either of these two methods can be used to determine an unknown weight (N) and they rely on the force of gravity to do this, but such instruments are much more commonly calibrated to indicate mass (kg or g) rather than weight. This is because we are usually more concerned with the 'amount' of something, rather than the effects of gravity on it. We usually assume that mass (kg) = weight (N)/9.81 anywhere on Earth, because any variations in the acceleration due to gravity, *g*, are insignificant for most, but not all, purposes.

Figure 2.43 An expensive ring

Question

1 If you were buying something small and expensive, like gold or diamonds (Figure 2.43), should the amount you are buying be measured as a mass or a weight? Explain your answer.

Tension and compression

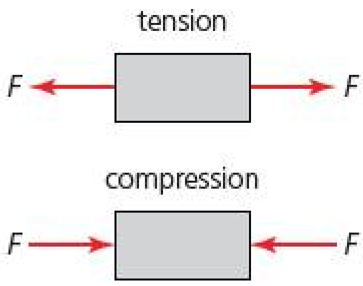


Figure 2.44 Object under **a** tension and **b** compression

When an object is stretched by equal and opposite forces pulling it outwards, we describe it as being **under tension** (Figure 2.44a).

Stretched strings or rubber bands are familiar examples of objects under tension, but **tensile** forces are also very common in more rigid objects, such as the horizontal tie in a stool or chair, where its purpose is to stop the legs from moving outwards.

When an object is squashed by equal and opposite forces pushing it inwards, we describe it as being **under compression** (Figure 2.44b).

All structures have parts that are in tension and parts in compression. The stone pillars at Stonehenge, in the UK (Figure 2.45) are strong under the compression caused by their own weight and the weight of the slabs resting on top.

Figure 2.45 Stonehenge was built over 4000 years ago

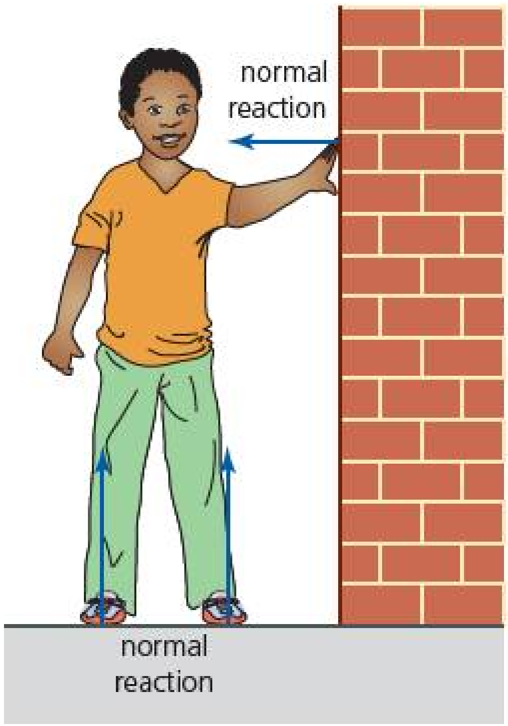


Figure 2.46 Reaction forces

The horizontal slabs on the top bend slightly, so that the top surface is compressed, while the lower surface is in tension and this may result in destructive cracks spreading upwards. Similar principles apply to the construction of all modern buildings, bridges, etc.

Reaction forces

If two objects are touching (in contact with) each other, then each must exert a force on the other. For example, if you push on a wall then the wall must also push on you; when you stand on the floor your weight presses down but the floor must also push up on you. If this was not true, you would fall through the floor and the wall.

In Figure 2.46 the boy's weight is pushing down on the ground and his hand is also pushing the wall. The force of the wall on the boy's hand and the force of the ground on his feet are examples of **contact forces** (also called **reaction forces**). These forces are always perpendicular to the surface and that is why they are often called **normal** reaction forces (the word 'normal' used in this way means perpendicular).

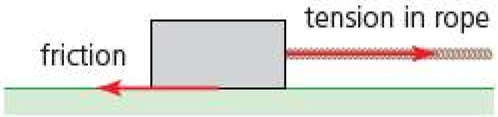


Figure 2.47 Friction opposing motion

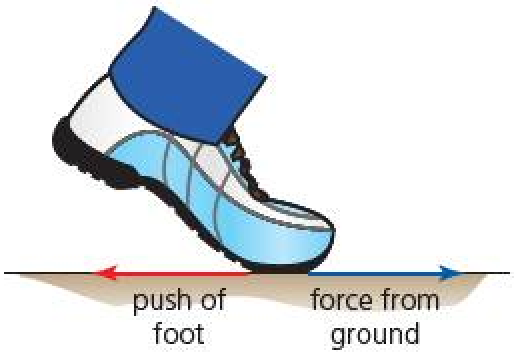


Figure 2.48 We need friction to walk

Friction

When objects are moving (or trying to move) and they are in contact with other surfaces, forces between the surfaces will act in such a way as to *oppose* (try to stop) the motion. This type of force is called **friction**.

There are many ways of trying to reduce the effects of friction in an attempt to make movement easier, but friction can never be completely overcome. Friction between two objects acts parallel to the surfaces of both, in the opposite direction to the motion (or intended motion). This is shown in Figure 2.47, in which a block is being accelerated by being pulled by a rope along the floor.

But without friction, movement would be very difficult. Consider how you walk across the room (Figure 2.48) – in order to take a step, the foot pushes backwards on the ground and, because of friction, the ground pushes forward on the foot. Without friction, walking and most methods of transportation would be impossible.

Additional Perspectives

Friction

If you try to move a stationary piece of furniture (for example) across the floor, to get it moving you have to provide a force that is greater than friction (Figure 2.49). Heavier objects are generally more difficult to move because they press vertically down on the floor with more weight, and this increases the amount of sideways friction. It is usually assumed, under these circumstances, that the frictional force is proportional to the weight of the object being moved.

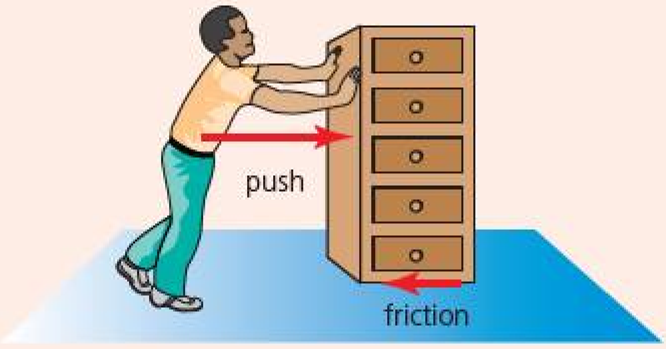


Figure 2.49 Pushing to overcome friction

Apart from the size of the normal contact forces between the two surfaces, there are many other factors that may affect the amount of friction, including the area in contact and the smoothness of the surfaces. Friction is not easy to analyse or predict.

As the force applied to an object (to try and make it move) is increased, the frictional force acting in the opposite direction increases equally. But there is a limit to how great any frictional force can be, and when the applied force becomes greater than the maximum frictional force, the object will start to move. Once movement has started the friction will usually decrease to a lower value and it then becomes easier to move the object. In other words, **static** (stationary) friction is usually greater than **dynamic** (moving) friction.

To reduce the friction between two given surfaces, something should be placed between them. This could be water, oil, air, graphite, small rollers or balls. Substances used in this way are called **lubricants**.

Figure 2.50 Tread on a car tyre

Much of road safety is dependent upon the nature of road surfaces and the tyres on vehicles. Friction between the road and a vehicle provides the forces needed for any change of velocity – speeding up, slowing down or changing direction. The materials used in the manufacture of tyres are chosen with friction in mind, as is the shape of the tyre – the larger the surface area in contact with the road, the greater the friction in any given situation. Smooth tyres will have the most friction in dry conditions, but when the roads are wet, ridges and grooves in the tyres are needed to disperse the water (Figure 2.50).

To make sure road surfaces produce enough friction, they cannot be too smooth and often need to be replaced every few years. This is especially important on sharp corners and hills. Anything that gets between the tyres and the road surface is likely to have a significant effect on friction and road safety, for example oil, water, ice and snow.

Questions

1 How can friction with the road be increased under icy conditions?
2 Why might reducing the pressure in a car tyre increase the friction?
3 Suggest why Formula One racing drivers 'warm up' the tyres on their cars. How is this done?

Air resistance

Air resistance (sometimes called **drag**) is also a force that opposes motion. An object moving through the air has to knock the air out of the way and this produces a force in the opposite direction to motion.

The study of the factors affecting air resistance is of great importance when discussing falling objects, parachutes, all modes of transport (especially for vehicles moving fast) and has many interesting sporting applications.

The amount of air resistance depends on the 'cross-sectional' area of the moving object, but also on the way in which the air flows past the surfaces. Altering the shape of an object and the nature of its surfaces can have a considerable effect on the air resistance it experiences. Changing the shape and/or surface of an object, particularly at the front, in order to reduce air resistance, is called **streamlining**.

Most importantly, the faster an object moves the greater the air resistance becomes opposing its motion. Typically, for a given object, it is often assumed that air resistance is proportional to the speed *squared*. This means that air resistance becomes *much* more important for objects moving very quickly. The air resistance opposing a 100 m sprinter moving at $10\,m\,s^{-1}$ could be 400 times greater than on a casual walker moving at $0.5\,m\,s^{-1}$. The retarding effects of air resistance on a car driving a few blocks to the local shops at an average speed of $30\,km\,h^{-1}$ will be much less than the same car driving at $110\,km\,h^{-1}$ along a motorway (about 13 times greater).

Figure 2.51 The LZR Racer swimsuit is built using NASA technology

Similar ideas apply to the drag effects when an object, person or an animal moves through (or on) water. For example, the amount of resistance experienced by a swimmer can be reduced by about 5% by wearing drag-reducing swimsuits, as shown in Figure 2.51. Of course it is very important that the suit does not affect the swimmer's movement in any way and that the extra weight of the swimsuit is insignificant.

Figure 2.52 Shanghai maglev train

Figure 2.52 shows one of the magnetic levitation ('maglev') trains that run between Shanghai and its main airport, which is about 30 km away. Magnetic forces lift the train above the surface of the track to eliminate friction, and the streamlined shape of the train is designed to reduce air resistance. The train completes the journey in about 7 minutes and reaches a top speed of about 430 $km\,h^{-1}$, although in test runs it exceeded 500 $km\,h^{-1}$.

The effect of air resistance on different objects is often tested in 'wind tunnels', such as shown in Figure 2.53. Instead of the object moving through stationary air, the object is kept still while fast moving air is blown against it.

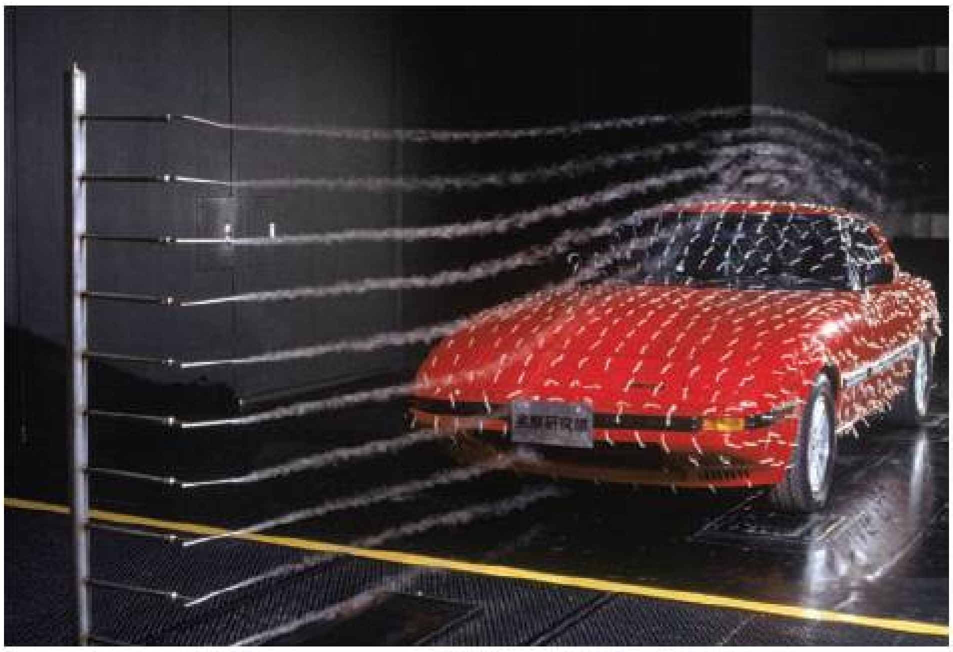


Figure 2.53 Wind tunnel testing

Additional Perspectives

Air travel

Aircraft use a lot of fuel moving passengers and goods from place to place quickly, but we are all becoming more aware of the effects of planes on global warming and air pollution. Some people think that governments should put higher taxes on the use of planes to discourage people from using them too much. Improving railway systems, especially by operating trains at higher speeds, will also attract some passengers away from air travel. Of course, engineers try to make planes more efficient, so that they use less fuel, but the laws of physics cannot be broken, and jet engines, like all other heat engines, cannot be made a lot more efficient than they are already.

Planes will use less fuel if there is less air resistance acting on them. This can be achieved by designing planes with streamlined shapes, and also by flying at greater heights where the air is less dense. Flying more slowly (than their maximum speed) can also reduce the amount of fuel used for a particular trip, as it does with cars, but people generally want to spend as little time travelling as possible.

The pressure of the air outside an aircraft at its typical cruising height is far too low for the comfort and health of the passengers and crew, so the air pressure has to be increased inside the plane, but this is still much less than the air pressure near the Earth's surface. The difference in air pressure between the inside and outside of the plane would cause problems if the plane had not been designed to withstand the extra forces.

Planes generally carry a large mass of fuel, and the weight of a plane decreases during the journey as the fuel is used up. The upwards force supporting the weight of a plane in flight comes from the air that it is flying through and will vary with the speed of the plane and the density of the air. When the plane is lighter towards the end of its journey it can travel higher, where it will have experience less air resistance.

Questions

1 a Find out how much fuel is used on a long-haul flight of, say, 12 hours.
  b Compare your answer to the capacity of the fuel tank on an average sized car.
  c On a short-haul flight as much of 50% of a plane's fuel may be used for taxiing, taking off, climbing and landing, but on longer flights this can reduce to under 15%. Explain the difference.

2 Do you agree that the use of planes should be discouraged in some way? Do the rest of your class agree with you? Does the government have the right to try to change people's behaviour by the use of taxes? (Taxes on alcohol and cigarettes are similar examples.)

Upthrust

Upthrust is a force exerted vertically upwards on *any* object which is in a fluid (gas or liquid). This force arises because the pressure of the fluid on the object is greater at the bottom than at the top. Upthrust acts in the opposite direction to weight and its effect is to reduce the apparent weight of the object. The upthrust from water is a familiar experience for swimmers and divers and it the same force which keeps a boat afloat. The weight of a floating object, or an object buoyant under water, is equal and opposite to the upthrust (see Figure 2.54). Upthrust forces also exist on objects in air, but they are less significant and are only normally noticeable on very light objects like balloons.

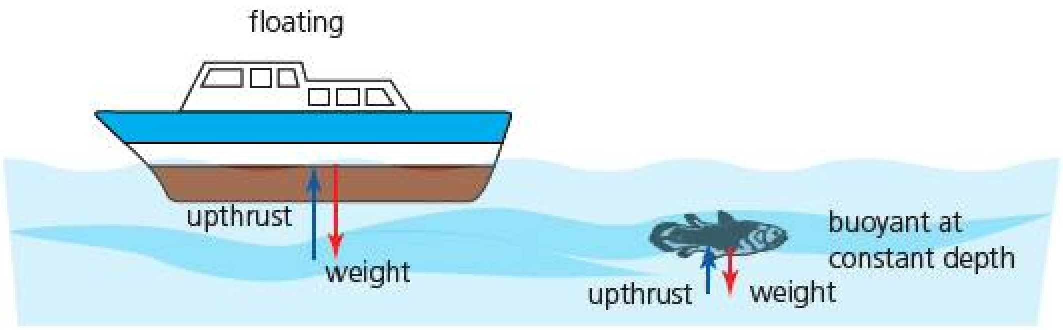


Figure 2.54 Forces on objects which are floating or buoyant under water

Non-contact forces

Of all the forces discussed above, **gravitational force (weight)** is different from the others because it acts across space and there does not need to be any contact (between an object and the Earth). **Magnetic and electric forces** behave in a similar way, and non-contact **nuclear forces** also exist within the atom. Understanding these **fundamental** forces plays a very important part in physics. These forces are all covered in more detail in later chapters of this book.

In order to fully explain all the contact forces mentioned earlier in this chapter, it would be necessary to consider the electromagnetic forces acting between particles in the different objects/substances.

Free-body diagrams

2.2.2 (part) **Draw** free-body diagrams representing the forces acting on an object.

When objects come into contact with each other they exert forces on each other. This means that even the simplest force diagrams can get confusing. To avoid this confusion, we often draw only one object and show only the forces acting *on* that one object. Forces that act from the body onto something else are *not* included.

Drawings that show only one object and the forces acting on it are called **free-body diagrams**. Some simple examples are shown in Figure 2.55.

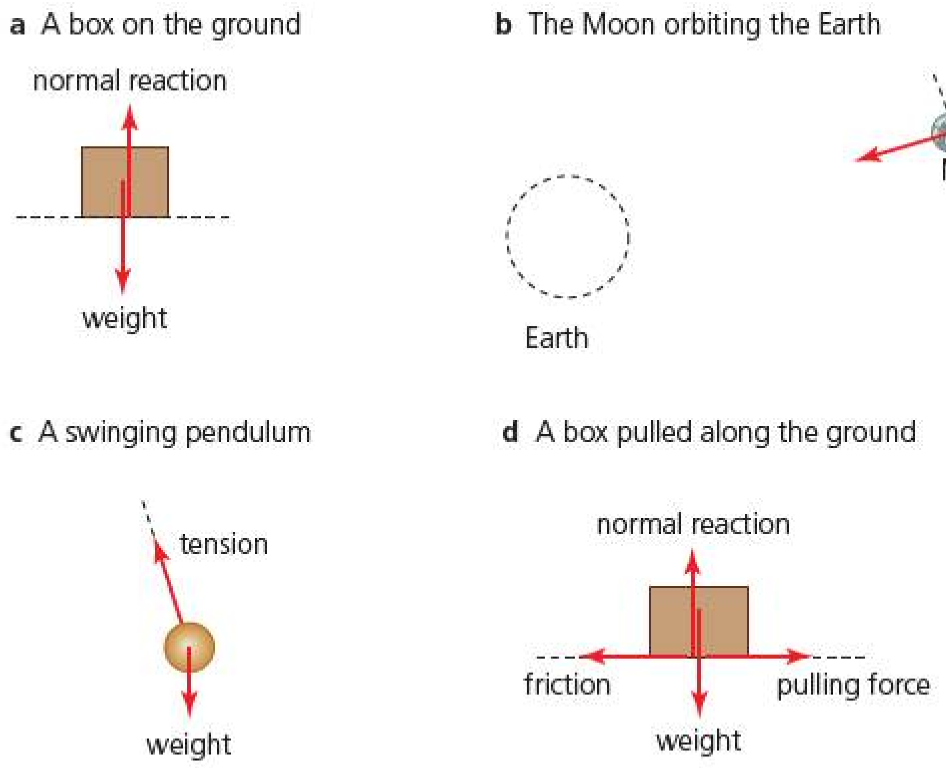


Figure 2.55 Free-body diagrams. The object has a solid outline and the forces are shown in red. The dotted lines would not usually be included

42 Figure 2.56 shows two unequal masses connected by string over a frictionless pulley. Copy the diagram and show the forces acting on the two masses.

43 Figure 2.57 represents a hot air balloon. The two ropes are stopping it from moving vertically away from the ground. Draw a free-body diagram for all the forces acting on the basket.

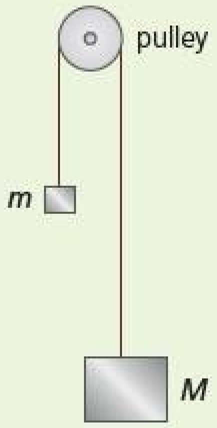


Figure 2.56

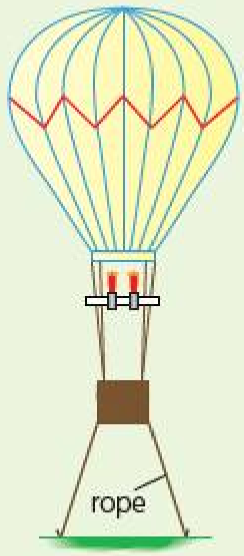


Figure 2.57

Resultant forces and component forces

Resultants

2.2.3 Determine the resultant force in different situations.

There is usually more than one force acting on an object. In order to determine the overall effect of two or more forces on an object, we must add up forces, taking their directions into consideration. This gives the **resultant** (overall, net) force acting on the object.

Determining the sums and differences of vectors, such as forces, was covered in Chapter 1. It may be helpful to review that section before continuing.

44 Use a scale drawing to determine the resultant of two forces of 8.5 N and 12.0 N acting at an angle of 120° to each other.

45 Calculate the resultant of forces of 7.7 N and 4.9 N acting perpendicularly to each other.

46 The resultant of two forces is 74 N to the west. If one force was 18 N to the south, what was the size and direction of the other force?

Components

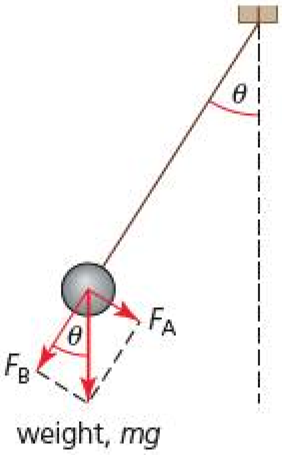


Figure 2.58 Resolving the weight of a pendulum into components

Two or more forces can be combined to give a resultant, but the 'opposite' process is just as important.

A single force can be considered as the sum of two separate forces, which are usually at right angles to each other ($F \cos \theta$ and $F \sin \theta$). This is called **resolving** a force into two **components** (Chapter 1).

Resolving into components is usually done when the actual force is not acting in the direction of motion. As an example, consider the swinging pendulum shown in Figure 2.58.

The single force of the weight, mg, can be resolved into a force F_A, acting in the instantaneous direction of motion and a force F_B, which is at right angles to F_A, acting along the line of the string and is equal and opposite to the tension in the string.

$$F_A = mg \sin \theta$$
$$F_B = mg \cos \theta$$

47 The mass of the pendulum shown in Figure 2.58 is 382 g and the angle θ is 27.4°.
- **a** What is the tension in the string?
- **b** What is the force acting in the direction of motion?

48 The mass shown in Figure 2.59 is stationary on the slope (inclined plane).
- **a** Draw a free-body diagram showing the forces acting on the mass.
- **b** Resolve the weight of the mass into two components that are parallel and perpendicular to the slope.

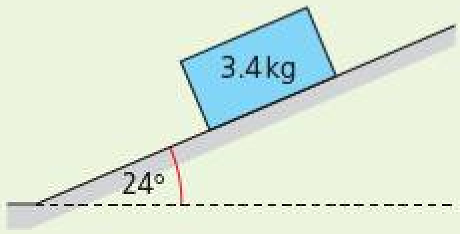


Figure 2.59

49 A resultant forward force of 8.42×10^4 N acts on a train of mass 3.90×10^5 kg accelerates it a rate of $0.216\,m\,s^{-2}$ when it is travelling on a horizontal track.
- **a** If the train starts to climb a slope of angle 1.00° to the horizontal, calculate the component of weight acting down the slope.
- **b** What is the new resultant force acting on the train?
- **c** Predict a possible acceleration of the train as it starts to climb the slope.
- **d** Suggest why it is difficult for trains to travel up steeper slopes.

Newton's first law of motion

2.2.4 **State** Newton's first law of motion.
2.2.6 **State** the condition for translational equilibrium.

If there is a resultant force acting on an object, it will accelerate. This is expressed in **Newton's first law of motion.**

Newton's first law of motion states that an object will remain at rest, or continue to move in a straight line at a constant speed, unless a resultant force acts on it.

An object that has no resultant force acting on it is said to be in **translational equilibrium.**

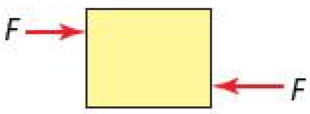


Figure 2.60 The object is in translational equilibrium, but not in rotational equilibrium

The word *translational* refers to movement from place to place. An object which is in translational equilibrium will continue to behave in the same way, i.e. it will remain at rest or it will continue to move in a straight line at constant speed.

We should note that it is possible for equal and opposite forces to act on an object *along different lines* and thereby cause it to rotate, as shown in Figure 2.60. The object will start to rotate under the turning effect of the two forces, but there will be no translational movement. It is not in *rotational* equilibrium.

Examples of Newton's first law

2.2.5 Describe examples of Newton's first law.

It is not possible to have an object on Earth with *no* forces acting on it, because gravity affects all masses. Therefore, any object which is in equilibrium must have at least two forces acting on it, and quite possibly many more. All moving objects, or objects tending to move, will also have frictional forces acting on them. If an object is in equilibrium, then the forces acting on it (in any straight line) are said to be *balanced*, so that the resultant force is zero. Consider the following examples.

- *Objects at rest with no sideways forces* – The box shown in Figure 2.55a is in equilibrium because its weight is equal to the normal reaction force pushing up on it.
- *Horizontal motion at constant velocity* – Consider Figure 2.55d, which also shows an object in translational equilibrium because the forces are balanced. It may be stationary or moving to the right with a constant velocity (we cannot tell from this diagram).
- *Vertical motion of falling objects* – Figure 2.61 shows a falling ball. In part **a** the ball is just starting to move and there is no air resistance. In part **b** the ball has accelerated and has some air resistance acting against its motion, but there is still a resultant force and acceleration downwards. In part **c** the speed of the falling ball has increased to the point where the increasing air resistance has become equal and opposite to the weight. There is then no resultant force and the ball is in translational equilibrium, falling with a constant velocity called its terminal velocity or terminal speed.

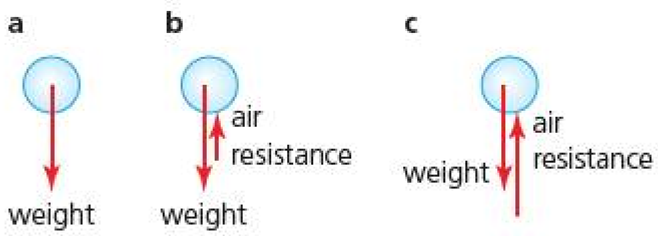


Figure 2.61 The resultant force on a falling object changes as it gains speed

- *Horizontal acceleration* – Figure 2.62 shows the forces acting on a bicycle and rider. Because the force from the road is greater than air resistance the cyclist will accelerate. As the bicycle and rider move faster and faster, they will meet more and more air resistance. Eventually the air resistance becomes equal to the forward force (but opposite in direction) and a top speed is reached. This is similar to the ideas used to explain the terminal speed of a falling object and the same principles apply to the motion of all vehicles.

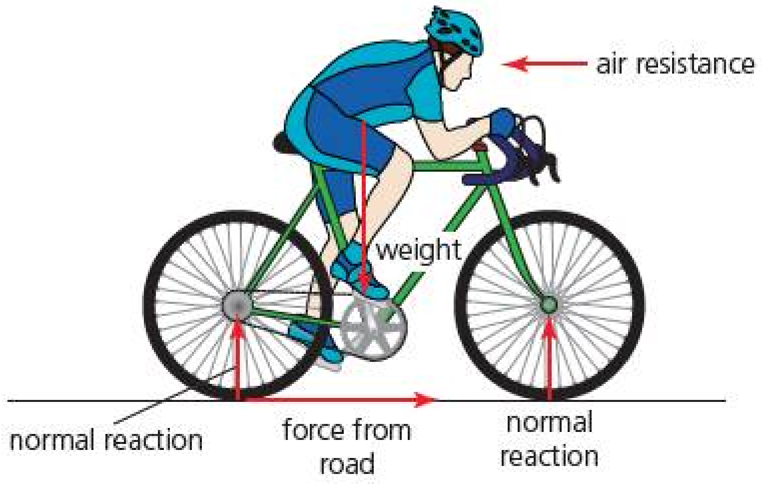


Figure 2.62 A cyclist accelerating

We know that any object that is stationary for some time, like a book placed on a table, is in equilibrium, and an object moving with constant velocity is also in equilibrium. But it is important to realize that a moving object that is only at rest for a moment is *not* in equilibrium. For example, a stone thrown vertically in the air comes to rest for a moment at its highest point, but the resultant force on it is not zero and it is not in equilibrium. Similarly, at the instant that a race is started a sprinter is stationary, but that is the time of greatest resultant force and acceleration.

Three forces in equilibrium

If two forces are acting on a mass such that it is *not* in equilibrium, then to produce equilibrium a third force can be added that is equal in magnitude to the resultant of the other two, but in the opposite direction. All three forces must act through the same point. For example, Figure 2.63 shows a ball on the end of a piece of string kept in equilibrium by a sideways pull which is equal in magnitude to the resultant of the weight and the tension in the string.

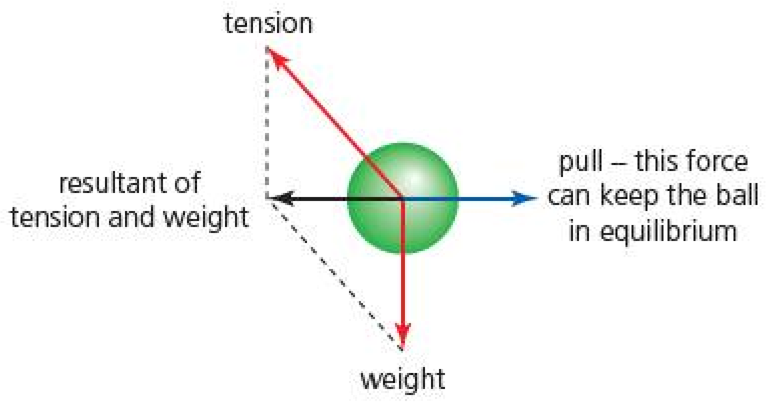


Figure 2.63 Three forces keeping a suspended ball in equilibrium

2.2.7 Solve problems involving translational equilibrium.

50 Draw fully labelled free-body diagrams for:
- **a** a car moving horizontally at constant velocity
- **b** an aircraft moving horizontally with a constant velocity
- **c** a boat decelerating after the engine has been switched off
- **d** a car accelerating up a hill.

51 Figure 2.64 shows the path of an object thrown through the air without air resistance. Make a copy of the diagram and add to it vector arrows to represent the forces acting on the object in each position.

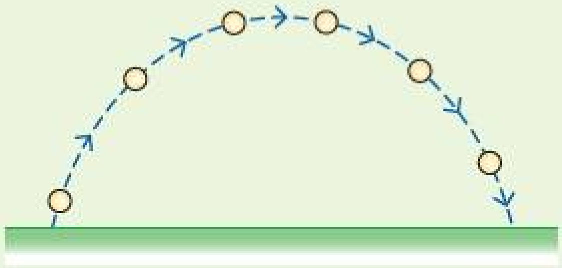


Figure 2.64

52 A heavy suitcase resting on the ground has a mass of 30.6 kg.
- **a** Draw a labelled free-body diagram to show the forces acting on the suitcase.
- **b** Re-draw the sketch to show all the forces acting if someone tries to lift up the case with a vertical force of 150 N.

53 Stand on bathroom scales with a heavy book in your hand. Quickly move the book upwards while watching the reading on the scales. Repeat, but this time move the book quickly downwards. Explain your observations.

54 a If you are in an elevator (lift) with your eyes closed, is it possible to tell if you are stationary, or moving up or moving down? Explain.
- **b** A person in an elevator experiences two forces: their weight downwards and the normal reaction force up from the floor. Sketch a free-body diagram to show the forces acting on a person in an elevator if:
  - **i** they are moving at a constant velocity
  - **ii** they are starting to move downwards
  - **iii** they are starting to move upwards
  - **iv** the elevator decelerates after it has been moving down
  - **v** the elevator decelerates after it has been moving upwards.

55 A skydiver is falling with a terminal velocity of about 200 km h^{-1} when he opens his parachute.
- **a** Draw free-body diagrams showing the forces acting on the skydiver:
  - **i** the moment that the parachute opens
  - **ii** just before the skydiver reaches the ground.
- **b** Sketch a fully labelled graph showing how the velocity of the skydiver changed from the moment he left the plane to the time he landed on the ground.

56 Refer back to Figure 2.59, which shows a box resting on a slope. It can only stay stationary because of the frictional force opposing its movement down the slope. What is the magnitude of the frictional force?

57 Figure 2.65 shows a climber using a rope to get up a mountain. Draw a free-body diagram to represent the forces acting on the climber.

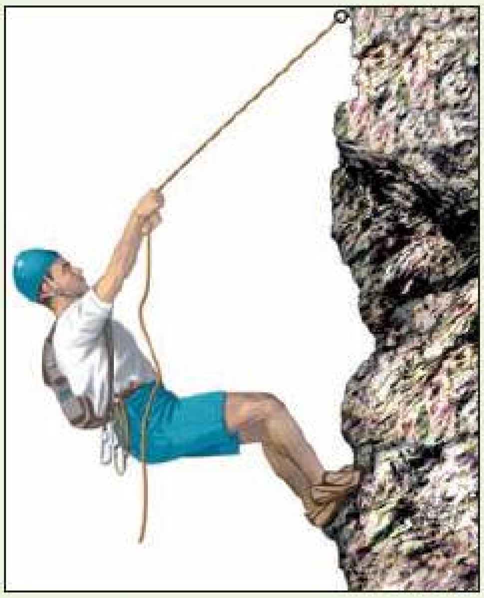


Figure 2.65

58 Figure 2.66 represents two raindrops falling side by side with the same instantaneous velocity. The forces acting on drop A are shown and it has a radius r.
- **a** Copy the diagram and show the forces acting on drop B, which has radius $2r$.
- **b** Describe the immediate motion of the two drops and explain the difference.

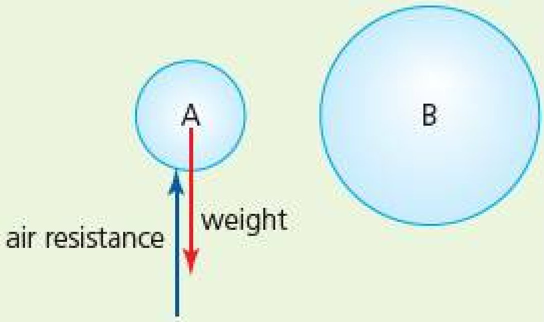


Figure 2.66

TOK Link: Aristotle and natural philosophy

Aristotle (384–322 BCE) was a Greek philosopher and one of the most respected founding figures in the development of human thinking and philosophy.

Aristotle's work covered a very wide range of subjects, including his interpretation of the natural world and the beginnings of what we now call science, although it was called 'natural philosophy' and had a very different approach from modern scientific methods.

Although the 'science' of the time did not involve careful observation, measurements, mathematics or experiments (remember, this was more than 2300 years ago), Aristotle did appreciate the need for universal (all-embracing) explanations of natural events in the world around him.

Aristotle believed that everything in the world was made of a combination of the four elements of earth, fire, air and water. The Earth was the centre of everything and each of the four earthly elements had its natural place. When something was not in its natural place, then it would tend to return – in this way he explained why rain falls, and why flames and bubbles rise, for example.

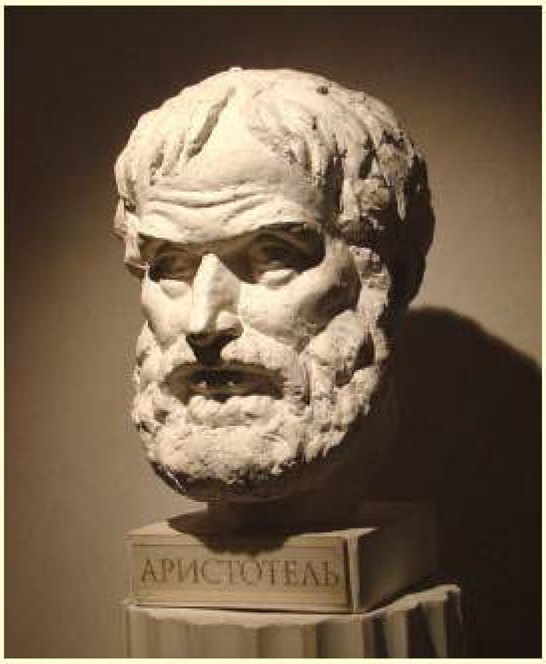


Figure 2.67 Aristotle

With our greatly improved knowledge in the modern world, it may be easy to dismiss Aristotle's work, and point out the inconsistencies, but his basic ideas on motion, for example, were so simple and powerful that they were widely believed for more than 1500 years, until the age of Galileo and Newton.

Questions

1 How would Aristotle have explained the fact that the Moon looks to be solid, but does not fall towards the Earth?

2 Use the Internet to find out about Aristotle's 'four causes'.

Newton's second law of motion

2.2.8 State Newton's second law of motion.

Newton's first law establishes that there is a link between resultant force and acceleration. Newton's second law takes this further and states the mathematical connection – when a resultant force acts on a (constant) mass, the acceleration is proportional to the force.

$$a \propto F$$

Both force and acceleration are vector quantities and the acceleration is in the same direction as the force.

Experiments also show that if the same resultant force is applied to different masses, then the acceleration produced is inversely proportional to the mass, m (that is, doubling the mass results in half the acceleration, etc.).

$$a \propto \frac{1}{m}$$

Combining these results, we see that acceleration is proportional to $\frac{F}{m}$. Newton's second law can be written as:

$$\text{force, } F \propto ma$$

If we define the unit of force, the newton, to be the force which accelerates 1 kg by 1 m s^{-2}, then we can write:

$$\text{force (N)} = \text{mass (kg)} \times \text{acceleration (m s}^{-2}\text{)}$$

$F = ma$ This equation is listed in the IB *Physics data booklet.*

This is just one version of Newton's second law. A second version of the law will be discussed later in the chapter, after the concept of momentum has been introduced.

Worked examples

8 a What resultant force is needed to accelerate 1.8 kg by 3.3 m s^{-2}?
b What acceleration would be produced if the same force were applied to a mass of 800 g?

a $F = ma$
$F = 1.8 \times 3.3$
$F = 5.9\,\text{N}$

b $a = \dfrac{F}{m}$

$a = \dfrac{5.9}{0.800}$

$a = 7.4\,\text{m s}^{-2}$

9 A car engine produces a forward force of 4400 N, but the forces of air resistance and friction opposing the motion of the car are a total of 1900 N.
a What is the resultant force acting on the car?
b If the car accelerates under the action of this force by 1.8 m s^{-2}, what is its total mass?

a Resultant force, $F = 4400 - 1900$
$F = 2500\,\text{N}$ forwards

b $m = \dfrac{F}{a}$

$m = \dfrac{2500}{1.8}$

$m = 1400\,\text{kg}$

10 a What is the resultant force acting on a mass of 764 g falling towards Earth (assume there is no air resistance)?
b What is the common name of this force?

a $F = ma$
$F = 0.764 \times 9.81$
$F = 7.49\,\text{N}$

b This force is usually called the weight of the object and, in this context, the equation $F = ma$ would usually be written as $W = mg$.

Forces acting for a limited time

The resultant forces that change the velocities of various objects do not continue to act for ever. Typically such forces act for a few seconds, and in the case of many **impacts** and **collisions** the duration of the forces involved may be only fractions of a second. (The word **interaction** is often used as a general term to include all possible situations.)

As an example, consider a falling object. If a mass is dropped onto something the force exerted during the impact is *not* equal to its weight (as is often thought), but can be calculated from $F = ma$. For example, if a falling 2.0 kg mass lands in sand and its speed is reduced from 6.0 m s^{-1} to zero in 0.50 s, then the average deceleration is $-12\,\text{m s}^{-2}$ ((0–6.0)/0.50), and the average retarding force is 24 N. If the same mass was dropped onto concrete at the same speed, the time for the impact would be much less and the force much greater. For example, if the impact lasted 0.050 s, then the average retarding force would be 240 N. The forces on the long jumper shown in Figure 2.68 are reduced because the time of impact is increased by landing on sand.

Whenever any moving object is stopped, the longer the time of impact, the smaller the deceleration and the smaller the forces involved. These ideas are very useful when using physics to explain, for example, how forces are reduced in accidents.

Figure 2.68 Impact in a sand-pit reduces force

When the aim is to accelerate an object with a force, the longer the time for which the force is applied, the greater the change of velocity. In many sporting activities a ball is struck with some kind of bat, club or racquet, and the longer the time of contact the better. This is one reason for the advice to 'follow through' with a strike.

2.2.9 Solve problems involving Newton's second law.

59 What resultant force is needed to accelerate a train of mass 3.41×10^5 kg from rest to 15.0 m s^{-1} in exactly 20 s?

60 When a force of 5.6 N was applied to a 4.3 kg mass it accelerated by 0.74 m s^{-2}. Calculate the frictional force acting on the mass.

61 A small plane of mass 12 400 kg is accelerated from rest along a runway by a resultant force of 29 600 N.
- **a** What is the acceleration of the plane?
- **b** If the acceleration remains constant, what distance is needed before the plane reaches its take-off speed of 73.2 m s^{-1}?

62 A car of mass 1200 kg was travelling at 22 m s^{-1} when the brakes were applied. The car came to rest in a distance of 69 m.
- **a** What was the deceleration of the car?
- **b** What was the average resultant force acting on the car?

63 A big box of mass 150 kg is pulled by a horizontal, thin rope in an effort to move it sideways along the ground. The frictional force is 340 N.
- **a** Draw a free-body diagram to show all the forces acting on the box when it just begins to move.
- **b** If the tension in the rope is 380 N, calculate the acceleration of the box.
- **c** Explain what might happen if an attempt was made to accelerate the box at 1 m s^{-2}.

64 A man of mass 82.5 kg is standing still in an elevator that is accelerating upwards at 1.50 m s^{-2}.
- **a** What is the resultant force acting on the man?
- **b** What is the normal reaction force acting upwards on him from the floor?

65 Figure 2.69 shows two masses connected by a light string passing over a pulley.
- **a** Assuming there is no friction, calculate the acceleration of the two blocks.
- **b** What resultant force is needed to accelerate the 2.0 kg mass by this amount?
- **c** Draw fully labelled free-body diagrams for the two masses, showing the magnitude and direction of all forces.

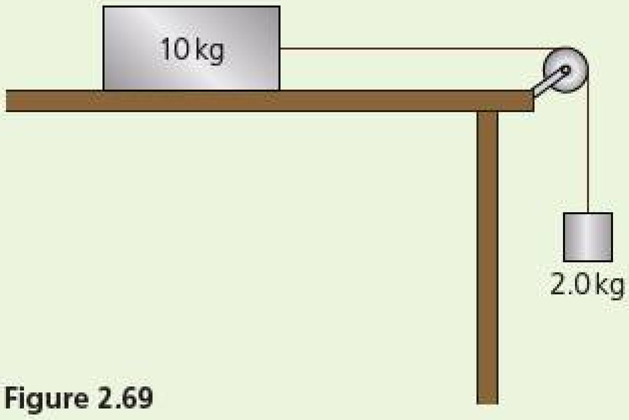


Figure 2.69

66 A high-jumper (Figure 2.70) of mass 75.4 kg raised his centre of mass from 0.980 m to 2.05 m when clearing a high-jump bar set at 2.07 m.
- **a** What vertical take-off speed was necessary in order for him to rise to this height?
- **b** If the athlete took 0.22 s to project himself upwards, what was the average resultant force acting on him?
- **c** What force did the ground exert upwards on him?
- **d** What did the athlete do in order to make the ground push him upwards?
- **e** Explain, using good physics, why it is sensible for him to land on foam when falling back to the ground.

Figure 2.70

67
- **a** Explain why a person jumping down from even a small height should bend their knees as they land.
- **b** Explain how air bags (and/or seat belts) reduce the injuries to passengers in car accidents.

68 A basketball of mass 0.62 kg was thrown downwards with an impact speed of 16 m s^{-1} onto bathroom scales and the maximum force was estimated to be 280 N.
- **a** If the average force exerted on the ball was about half the maximum force, what was the approximate deceleration of the basketball?
- **b** Estimate the length of time that the basketball was in contact with the scales before it bounced away from the surface.

Momentum

2.2.10 Define *linear momentum* and *impulse*.

The product of mass and velocity is an important quantity in physics.

Linear momentum is defined as mass multiplied by velocity.

Momentum is given the symbol p and has the units kilogram metre per second, $kg\,m\,s^{-1}$.

$p = mv$ This equation is listed in the IB *Physics data booklet*.

In all calculations it is very important to remember that momentum is a vector quantity and the direction should always be given. In this course we only deal with momentum in straight lines (*linear* momentum), but similar ideas can be applied elsewhere to the *angular* momentum involved with rotating objects or particles.

Worked example

11 A ball of mass 540 g moving vertically downwards hits the ground with a velocity of $8.0\,m\,s^{-1}$. After impact it bounces upwards with an initial velocity of $5.0\,m\,s^{-1}$.
a Calculate the momentum of the ball immediately before and after impact.
b What was the change of momentum during impact (Δp)?

a Initial momentum before impact, $p = mu = 0.54 \times 8.0 = 4.3\,kg\,m\,s^{-1}$ downwards.
Final momentum after impact, $p = mv = 0.54 \times 5.0 = 2.7\,kg\,m\,s^{-1}$ upwards.
The opposite directions may be represented by positive and negative signs rather than written descriptions. That is, the momentum before impact was $+4.3\,kg\,m\,s^{-1}$ and the momentum afterwards was $-2.7\,kg\,m\,s^{-1}$. (The choice of signs is interchangeable.)
b $\Delta p = (-2.7) - (+4.3) = -7.0\,kg\,m\,s^{-1}$ (upwards)

Expressing Newton's second law in terms of momentum

Returning to Newton's second law of motion ($F = ma$), we can now rewrite it in terms of momentum by using the definition of acceleration ($a = (v - u)/t$).

$$F = \frac{m(v-u)}{t}$$

$$F = \frac{mv - mu}{t}$$

The equation above can be expressed in words as force equals change of momentum divided by time, or *force equals the rate of change of momentum*. This version of Newton's second law is commonly written in the form:

$$F = \frac{\Delta p}{\Delta t}$$ This equation is also listed in the IB *Physics data booklet*.

For most situations the two versions of Newton's second law of motion are equivalent to each other, and the choice of which to use is dependent on the information provided. Because the version above does not require a constant mass, it is a more generalized statement of the law.

Impulse

Rearranging the equation $F = \Delta p/\Delta t$ gives the equation $F\Delta t = \Delta p$. With the equation written in this way we can see that the change of momentum is equal to the product of the force and the time for which it acts. When a resultant force acts for a longer time it has a greater effect.

Impulse is defined as the product of force and the time for which the force acts.

Impulse is numerically equal to the change of momentum it produces ($F\Delta t = \Delta p$).

Since examples of the use of this equation are often limited to changes of velocity of constant masses, it is often written in the following form:

$$\text{impulse} = F\Delta t = m\Delta v$$

This form of the equation is also listed in the IB *Physics data booklet.*

Impulse has the same units as momentum: kg m s^{-1} (or N s).

Worked example

12 Calculate the average force exerted on the bouncing ball by the ground in Worked example 11 if the duration of impact was 0.18 s.

$$F = \frac{\Delta p}{\Delta t}$$

$$F = \frac{-7.0}{0.18} = -39\,\text{N}$$ (The negative sign represents a force upwards)

Alternatively, the same answer can be obtained using $F = ma$:

$$a = \frac{v-u}{t} = \frac{(-5.0)-(+8.0)}{0.18} = -72\,\text{m s}^{-2}$$

Then, $F = ma = 0.54 \times (-72) = -39\,\text{N}$

Force–time graphs

2.2.11 Determine the impulse due to a time-varying force by **interpreting** a force–time graph.

In many simple calculations we may assume that the forces involved are constant, or that the average force is half of the maximum force. For more accurate work this is not good enough and we need to know how the force varied during the interaction. This is commonly represented by a force–time graph.

The area under a force–time graph ($F\Delta t$) for an interaction equals the impulse (change of momentum).

Worked example

13 Figure 2.71 shows how the force on a 57 g tennis ball moving at $24\,\text{m s}^{-1}$ varied when it was struck by a racquet moving in the opposite direction.

a Estimate the impulse given to the ball (from the area under the graph).

b What was the velocity of the ball after being struck by the racquet?

c The ball is struck with the same force with different racquets. Explain why a racquet with looser strings will return the ball with greater speed.

d What is the disadvantage of playing tennis with a tennis racquet with loose strings?

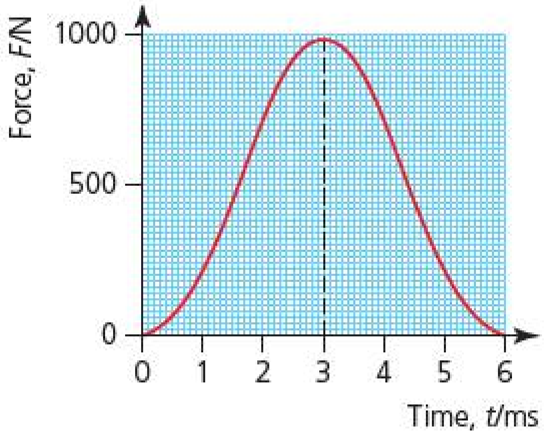


Figure 2.71 Force–time graph for striking a tennis ball

a Approximate area under graph $= 1000 \times (3.0 \times 10^{-3}) = 3.0\,\text{N s}$

b $m\Delta v = 3.0$

$$\Delta v = \frac{3.0}{0.057} = 53\,\text{m s}^{-1}$$

The initial velocity was $24\,\text{m s}^{-1}$ towards the racquet. If the change of velocity was $53\,\text{m s}^{-1}$, then the ball must move away from the racquet with a velocity of $(53 - 24) = 29\,\text{m s}^{-1}$.

c The time of contact with the ball, Δt, will be longer with looser strings, so that the same force will produce a greater impulse (change of momentum).

d There is less control over the direction of the ball.

2.2.13 Solve problems involving momentum and impulse

69 Figure 2.72 shows how the force between two colliding cars changed with time. Both cars were driving in the same direction and after the collision they did not stick together.
- **a** Estimate the impulse.
- **b** Just before the collision the faster car (mass 1200 kg) was travelling at $18\,m\,s^{-1}$. Calculate a value for its speed immediately after the collision.

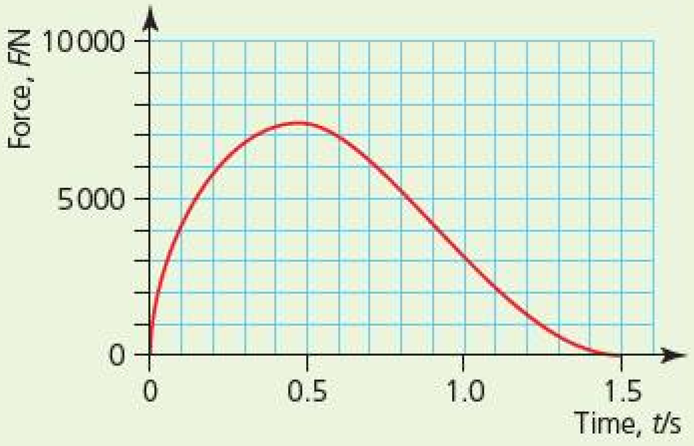


Figure 2.72

70 A soft ball (A) of mass 500 g is moving to the right with a speed of $3.0\,m\,s^{-1}$ when it collides with another soft ball (B) moving to the left. The time of impact was 0.34 s, after which ball A rebounded with a speed of $2.0\,m\,s^{-1}$.
- **a** What was the change of velocity of ball A?
- **b** What was the change of momentum of ball A?
- **c** Calculate the average force exerted on ball A.
- **d** Sketch a force–time graph for the impact.
- **e** Add to your sketch a possible force–time graph for the collision of hard balls of similar masses and velocities.
- **f** Suggest how a force–time graph for ball B would be different (or the same) as for ball A.

Conservation of momentum

2.2.12 State the law of conservation of linear momentum.

Momentum is a key concept in physics because whenever two or more objects interact in any way (for example, in collisions or in 'explosions' – in which two or more masses are propelled off in different directions), the total momentum immediately before the interaction is *always* the same as the total momentum immediately afterwards. This assumes that there are no external forces acting on the objects. We say that momentum is *conserved* ('conservation' means to keep the same). This is always true – there are *no* exceptions to the law of *conservation of momentum*.

The total (linear) momentum of a system is constant provided that no external forces are acting on it.

Figure 2.73 Momentum is conserved in this collision test

The **system** is the interacting objects that are being considered. To make it clear that there can be no forces from outside the system we often refer to an **isolated** system when stating this law.

Since it is *always* true, the law can be applied with confidence to make predictions about interactions in *any* isolated system, for example, in the crash test shown in Figure 2.73.

It is easy to give examples in which the momentum of an object decreases to zero, which may appear to contradict the law of conservation of momentum. This apparent loss of momentum is usually because the system is not isolated – it is acted on by external forces, such as friction. In other examples, some or all momentum may appear to be lost when something collides with an object that has a much greater mass. The motion after impact may be too small to observe or measure. A typical example could be a person jumping down onto the ground. The predicted motion of the person–Earth system after impact is totally insignificant.

The force of gravity commonly increases the momentum of falling objects, but the objects are not isolated systems – there are external forces acting on them. For example, a 3 kg rock experiences a gravitational force towards the Earth of approximately 30 N and therefore gains momentum as it accelerates down. The law of conservation of momentum predicts correctly that the Earth must gain an equal momentum upwards towards the rock. Because the mass of the Earth is so great, its gain of momentum is totally insignificant.

Worked examples

14 Mass A (4.0 kg) is moving at $3.0\,m\,s^{-1}$ to the right when it collides with mass B (6.0 kg) moving in the opposite direction at $5.0\,m\,s^{-1}$.

a If they stick together after the collision, what is their velocity?

b What assumption did you make?

a The total momentum must be the same before and after collision. Choosing velocities and momenta (momentums) to the right to be positive and to the left to be negative:

momentum of A $= m_A u_A = 4.0 \times (+3.0) = +12.0\,kg\,m\,s^{-1}$

momentum of B $= m_B u_B = 6.0 \times (-5.0) = -30\,kg\,m\,s^{-1}$

Therefore, the total momentum (before) the collision is $+12 + (-30) = -18\,kg\,m\,s^{-1}$. The total momentum after the collision must be the same as that before, so:

$m_{AB}v_{AB} = -18\,kg\,m\,s^{-1}$

$(4+6)v_{AB} = -18\,kg\,m\,s^{-1}$

$v_{AB} = \frac{-18}{10} = -1.8\,m\,s^{-1}$

The negative sign means the velocity is to the left.

b The assumption made is that there are no external forces acting on the system. If there are significant frictional forces involved, the calculated answer can be taken to be the instantaneous velocity immediately after collision, and the effects of friction can be considered afterwards.

This explanation has been written out in detail to aid understanding. A more direct way of answering any such question involving two masses interacting is as follows:
momentum before interaction = momentum after interaction.

$m_A u_A + m_B u_B = m_A v_A + m_B v_B$

In this example:

$(4.0 \times 3.0) + (6.0 \times -5.0) = (4.0 + 6.0) \times v_{AB}$

$v_{AB} = -1.8\,m\,s^{-1}$

15 A bus of mass 5800 kg travelling at a steady $24\,m\,s^{-1}$ runs into the back of a car of mass 1200 kg travelling at $18\,m\,s^{-1}$ in the same direction. If the car is pushed forward with a velocity of $20\,m\,s^{-1}$, calculate the velocity of the bus immediately after the collision. (Assume this is an isolated system and ignore the actions of the drivers and engines.)

momentum before interaction = momentum after interaction.

$m_A u_A + m_B u_B = m_A v_A + m_B v_B$

$(5800 \times 24) + (1200 \times 18) = (5800 \times v_{bus}) + (1200 \times 20)$

$v_{bus} = 23.6\,m\,s^{-1}$ in the original direction

16 A bullet of mass 12.0 g is fired at a speed of $550\,m\,s^{-1}$ from a rifle of mass 1.40 kg (Figure 2.74). What happens to the rifle?

Figure 2.74

In this example the total momentum is zero. This means that the momentum of the bullet must be equal and opposite to the momentum of the rifle:
momentum before interaction = momentum after interaction

$m_A u_A + m_B u_B = m_A v_A + m_B v_B$

$0 + 0 = (0.012 \times 550) + (1.40 \times v_{rifle})$

$v_{rifle} = \frac{-6.60}{1.40} = -4.71\,m\,s^{-1}$

The negative sign shows that the rifle moves in the opposite direction to the bullet. (This is often called **recoil**.)

2.2.13 Solve problems involving momentum and impulse.

71 A body of mass 2.3 kg moving to the left at 82 cm s^{-1} collides with a stationary mass of 1.9 kg. If they stick together, what is their velocity after impact?

72 A 50.9 kg bag of cement is dropped from rest onto the ground from a height of 1.46 m.
- **a** What is the maximum speed of the bag as it hits the ground?
- **b** Will the actual speed be much different from the maximum theoretical speed? Explain your answer.
- **c** Assuming that the bag does not bounce, predict the combined speed of the Earth and bag after the impact. (Mass of the Earth = 6.0×10^{24} kg.) Is it possible to measure this speed?

73 In an experiment to find the speed of a 2.4 g bullet, it was fired into a 650 g block of wood at rest on a friction-free surface. If the block (and bullet) moved off with an initial speed of 96 cm s^{-1}, what was the speed of the bullet?

74 A ball moving vertically upwards decelerates and its momentum decreases, although the law of conservation of momentum states that total momentum cannot change. Explain this observation.

75 An astronaut of mass 90 kg pushes a 2.3 kg hammer away from her body with a speed of 80 cm s^{-1}. What happens to the astronaut? How can she stop moving?

76 Two toy cars travel in straight lines towards each other on a friction-free track. Car A has a mass of 432 g and a speed of 83.2 cm s^{-1}. Car B has a mass of 287 g and speed of 68.2 cm s^{-1}. If they stick together after impact, what is their combined velocity?

77 A steel ball of mass 1.2 kg moving at 2.7 m s^{-1} collides head-on with another steel ball of mass 0.54 kg moving in the opposite direction at 3.9 m s^{-1}. The balls bounce off each other, each returning back in the direction it came from.
- **a** If the smaller ball had a speed after the collision of 6.0 m s^{-1}, use the law of conservation of momentum to predict the speed of the larger ball.
- **b** In fact, this result is not possible. Suggest a reason why not.

78 Figure 2.75 shows two trolleys on a friction-free surface joined together by a thin rubber cord in tension. When the trolleys are released, they accelerate towards each other and the cord quickly becomes loose. Where would you expect the two trolleys to collide with each other?

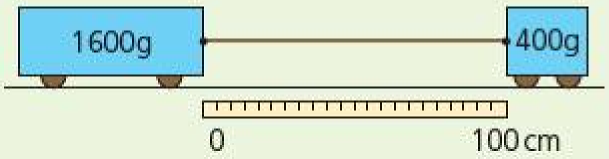


Figure 2.75

Newton's third law of motion

2.2.14 State Newton's third law of motion.
2.2.15 Discuss examples of Newton's third law.

From the law of conservation of linear momentum, we know that when any two objects (A and B) interact, the change of momentum of one must be equal and opposite to the change of momentum of the other. We can write this as:

$$\Delta p_A = -\Delta p_B$$

The time of interaction, Δt, must be the same for both, so that:

$$\frac{\Delta p_A}{\Delta t} = -\frac{\Delta p_B}{\Delta t}$$

Since, $F = \Delta p/\Delta t$, we can write:

$$F_A = -F_B$$

The force that A exerts on B is equal and opposite to the force that B exerts on A (see Figure 2.76). This statement is equivalent to the law of conservation of momentum, but expressed in this way it is known as Newton's third law of motion.

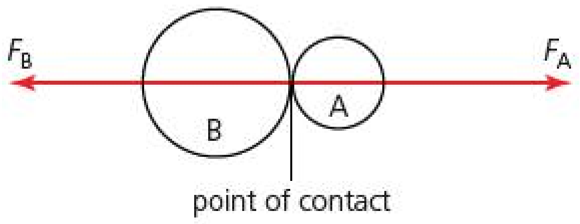


Figure 2.76 When two bodies interact $F_A = -F_B$

Newton's third law of motion states that whenever one body exerts a force on another body, the second body exerts exactly the same force on the first body but in the opposite direction.

Essentially this law means that forces must *always* occur in equal pairs, although it is important to realize that the two forces must act on *different* bodies and in opposite directions. Sometimes the law is quoted in the form used by Newton, 'to every action there is an equal and opposite reaction'. In everyday terms, it is simply not possible to push something that does not push back on you. Here are some examples.

- If you pull a rope, the rope pulls you.
- If the Earth pulls a person, the person pulls the Earth (Figure 2.77).
- If a fist hits a cheek, the cheek hits the fist (Figure 2.78).
- If you push on the ground, the ground pushes on you.
- If a boat pushes down on the water, the water pushes up on the boat.
- If the Sun attracts the Earth, the Earth attracts the Sun.
- If a plane pushes down on the air, the air pushes up on the plane.

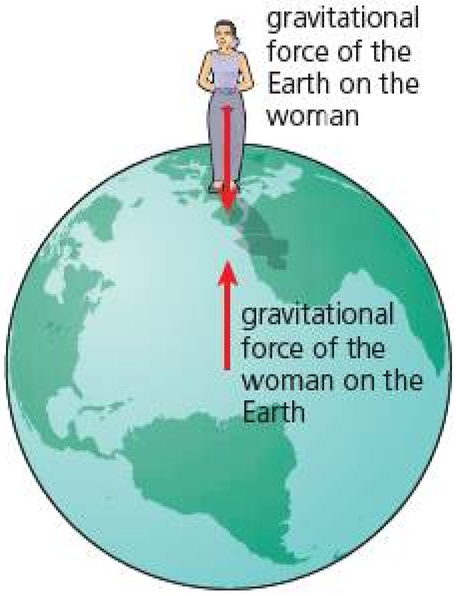


Figure 2.77

Figure 2.78 The force on the glove is equal and opposite to the force on the cheek

Figure 2.79 The force pushing the man forward is equal to the force pushing the boat backwards

Vehicle propulsion

Figure 2.80 The propeller at the back of a boat exerts a force on the water and the water pushes back

In order to accelerate a vehicle, or keep it moving at a constant velocity, a force is needed to overcome friction and/or air resistance. This can be achieved by the vehicle pushing backwards on a road or track. But, there is another way, which may be best explained by considering the momentum given to something in the opposite direction to the intended vehicle motion.

When the propeller in Figure 2.80 spins, a force pushes the water backwards and an equal and opposite force therefore pushes the boat forwards. The backwards momentum of the high-speed water is equal and opposite to the forward momentum of the slower-moving boat.

Figure 2.81 Helicopter blades give downwards momentum to the air

The use of propellers on planes and helicopters (Figure 2.81) involves similar principles, as does the more basic use of oars and paddles in boats.

The jet engines on the plane shown in Figure 2.40 burn fuel combined with oxygen from the air that is taken in at the front of the engine. The resulting gases are ejected (thrown out) at the back of the engine with high speed. The gain in backwards momentum of the gases must be equal and opposite to the forward momentum of the engine and plane. Newton's third law offers an alternative and equivalent explanation – the force pushing the gases backwards is equal and opposite to the force pushing the plane forwards.

79 Figure 2.82 shows a suggestion to make a sailing boat move when there is no wind. Discuss how effective this method could be.

80 A rocket engine on a spacecraft of total mass 10 000 kg ejected 1.4 kg of hot gases every second at an average speed of $240\,\text{m s}^{-1}$.
a What was the force on the spacecraft?
b What was the acceleration of the spacecraft?
c If the engines were used for 30 s, what was the impulse given to the spacecraft?
d What was the change of velocity?

Figure 2.82

81 A book has a weight of 2 N and is at rest on a table. The table exerts a normal reaction force upwards on the book. Explain why these two forces are not an example of the Newton's third law.

2.3 Work, energy and power

Energy is probably the most widely used concept in the whole of science and the word is also in common use in everyday language. Despite that, the abstract idea of energy is not easy to explain or define, although it is still a very useful concept.

When a battery is placed in a child's toy dog (see Figure 2.83), it moves, jumps up in the air and barks. After a certain length of time, the toy stops working. In order to try to *explain* this we will almost certainly need to use the concept of energy. Chemical energy in the battery is transferred to electrical energy, which produces motion energy in a small electric motor. Some energy is also transferred from electricity to sound in a loudspeaker. Eventually all the useful energy in the battery is transferred to the surroundings. The concept of 'energy' makes this easier to explain.


Figure 2.83 This toy dog is powered by batteries

We may talk about the energy 'in' the gasoline (petrol) we put in the tanks in our cars (for example), and go on to describe that energy being transferred to the movement of the car. But nothing has actually flowed out of the gasoline into the car, and all this is just a convenient way of expressing the idea that the controlled combustion of gasoline with oxygen in the air can do something which we consider to be useful.

Perhaps the easiest way to understand the concept of energy is this: energy is needed to make things happen. Whenever *anything* changes, energy is transferred from place to place or from one form to another. Most importantly, energy transfers can be *calculated* and this provides the basic 'accounting system' for science. Any event will require a certain amount of energy for it to happen and if there is not enough energy available, it cannot happen. For example, if you do not get enough energy (from your food), you will not be able to climb a 500 m hill; if you do not put enough gasoline in your car, you will not get to where you want to go; if energy is not transferred quickly enough from an electrical heater, it will not keep your room warm enough in cold weather, etc.

Calculating energy transfers is a very important part of physics and we start with the very common situations in which energy is transferred to objects to make them move.

Work

2.3.1 Outline what is meant by work.

One very common kind of energy transfer occurs when an object is moved (displaced) by using a force. This is called doing **work** (symbol *W*). The word 'work' has a very precise meaning in physics, which is different from its use in everyday language.

work done = force × displacement in the direction of the force

If the force is constant and in the same direction as movement, we can write:

$W = Fs$

Figure 2.84 This weightlifter is not doing any work

However, in many examples these assumptions may not be true.

The unit of work is the **joule, J**. 1 J is the work done when a force of 1 N displaces an object 1 m in the direction of the force. The same unit is used for measuring *all* forms of energy. Kilojoule, kJ and megajoule, MJ, are also in common use. Work and energy are scalar quantities.

The weightlifter in Figure 2.84 is *not* doing any work on the weights over her head while they are being held still, because there is no movement. But energy *is* being transferred in her muscles and, no doubt, they hurt because of that.

If movement is perpendicular to a force, no work is done by the force. For example the force of gravity keeps the Moon in orbit around the Earth (Figure 2.85), but no work is being done.

Worked example

17 How much work is done when a 1.5 kg mass is raised 80 cm vertically upwards?

The force needed to raise an object (at constant velocity) is equal to its weight (mg) and the symbol h is widely used for vertical distances. (To avoid confusion, W will normally be used to represent work and not weight.)

$W = Fs$

$W = mg \times h$

$W = 1.5 \times 9.81 \times 0.80$

$W = 12\,\text{J}$

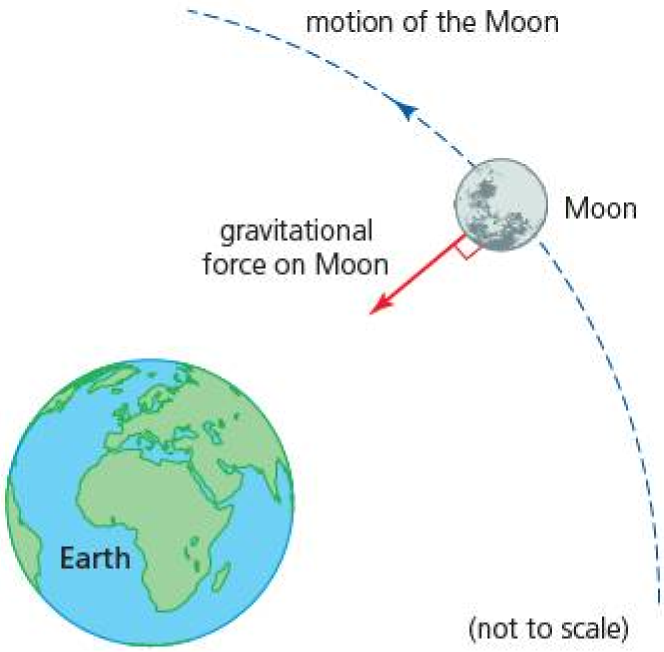


Figure 2.85 No work is done as the Moon orbits the Earth.

Calculating the work done if the force is not in the same direction as the displacement

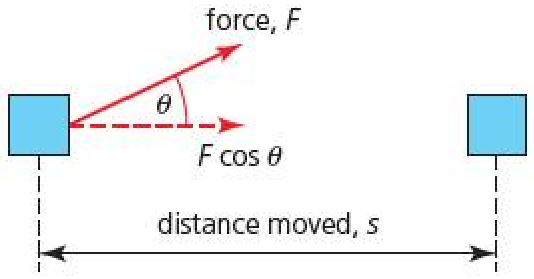


Figure 2.86 Movement which is not in the same direction as the force

Sometimes a displacement is produced that is not in the same direction as the force (Figure 2.86). For whatever reason, the object shown can only be displaced in the direction shown by the dashed line, but the force is acting at an angle θ to that line.

We know that work done = force × displacement in the direction of the force. In this case the force used to calculate the work done will be the *component* of F in the direction of movement, that is, $F\cos\theta$, so that:

$$W = Fs\cos\theta$$

This equation is listed in the IB *Physics data booklet*.

If force and movement are in the same direction, $\cos\theta$ equals 1 and the equation becomes $W = Fs$, as before. If the force is perpendicular to motion, then $\cos\theta$ equals zero and no work is done.

Worked example

18 The 150 kg box in Figure 2.87 was pulled 2.27 m across horizontal ground by a force of 248 N as shown.

a How much work was done by the force?

b If the frictional force was 132 N, what was the acceleration of the box?

c Suggest why it may make it easier to move the box if it is pulled in the direction shown by the dashed line.

d When the box was pulled at an angle of 20° to the horizontal, the force used was 204 N. Calculate the work done by this force in moving the box horizontally the same distance.

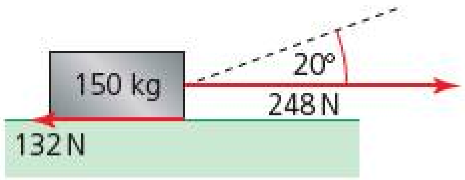


Figure 2.87

a $W = Fs$

$W = 248 \times 2.27 = 563\,\text{J}$

b $a = \dfrac{F}{m}$

$a = \dfrac{248 - 132}{150}$

$a = 0.773\,\text{m s}^{-2}$

c When the box is pulled in this direction the force has a vertical component that helps reduce the normal reaction force between the box and the ground. This will reduce the friction opposing horizontal movement.

d The force is not acting in the same direction as the movement. To calculate the work done we need to use the horizontal component of the 204 N force.

$W = F\cos 20° \times s$

$W = 204 \times 0.940 \times 2.27$

$W = 435\,\text{J}$

2.3.3 Solve problems involving the work done by a force.

82 Calculate the work done when a 12 kg suitcase is:
a pushed 1.1 m across the floor with a force of 54 N
b lifted 1.1 m upwards.

83 The pram in Figure 2.88 is being pulled with a force that is not in the same direction as its movement. Calculate the work done when the pram is pulled 90 cm to the right with a force of 36 N at an angle of 28° to the horizontal.

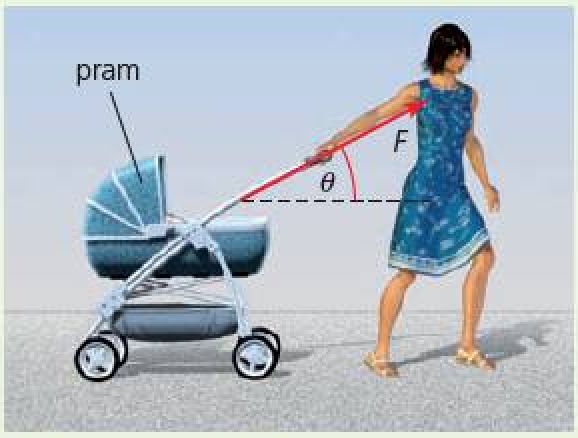


Figure 2.88

84 In Figure 2.89 a gardener is pushing a lawnmower at a constant speed of $0.85\,\mathrm{m\,s^{-1}}$ with a force, P, of 70 N at an angle of 40° to the ground.

a Calculate the magnitude of the frictional force, F.

b How much work is done in moving the lawnmower for 3 s?

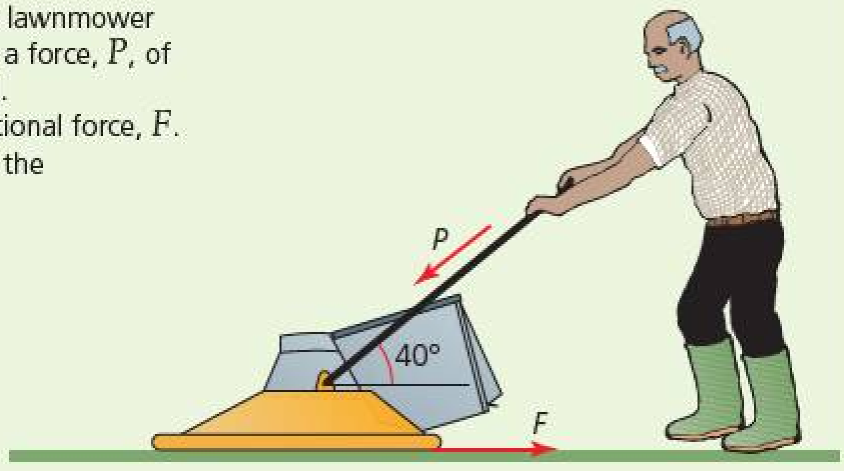


Figure 2.89

Work done by varying forces

2.3.2 **Determine** the work done by a non-constant force by **interpreting** a force–displacement graph.

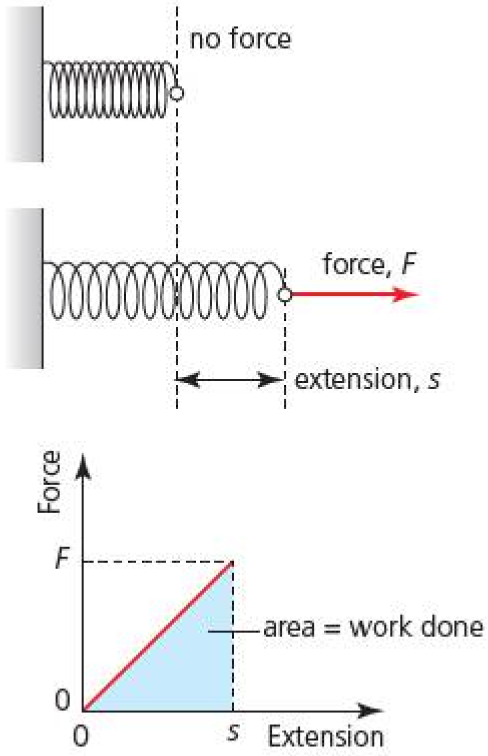


Figure 2.90 Force and extension when stretching a spring

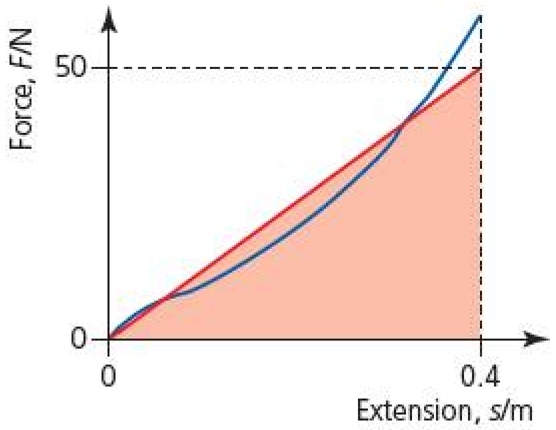


Figure 2.91 Force–extension (displacement) graph for rubber under tension

When making calculations with the equation $W = Fs\cos\theta$ we must use a single value for the force, but, in reality, forces are rarely constant. In order to calculate the work done by a *varying* force we have to make an estimate of the *average* force involved. In some situations we can assume that the force varies in a predictable and linear way, so that the average force is halfway between the starting force and the final force. A simple example would be calculating the work done in stretching a spring, as shown in Figure 2.90. Provided that the spring is not over-stretched, the force is proportional to the **extension** (displacement of the end of the spring), as shown in the force–extension graph. (This is known as *Hooke's law*.)

Using $W = Fs$ in this situation involves using the *average* force during the extension of the spring. As the spring was stretched, the force increased linearly from zero to F and the average was $F/2$. The work done, W, was therefore $(F/2) \times s$.

This calculation is identical to a calculation to determine the shaded area under the graph. This is *always* true – if the variation of force with displacement is known in detail, a force–displacement graph can be drawn.

The area under a force–displacement graph equals the work done.

The usefulness of determining the work done from the area under a graph becomes more obvious when we consider a material like rubber, for which the displacement is *not* proportional to the force. Typically, rubber becomes stiffer the more it is stretched. Figure 2.91 shows a possible force–extension (displacement) graph for rubber.

The work done, W, in stretching the rubber by 40 cm (for example) is found from the area under the curved graph. This can be estimated from the shaded triangle, which is judged by eye to have a similar area.

$$W = 0.5 \times 50 \times 0.40 = 10\,\mathrm{J}$$

Worked example

19 a Calculate the work done in raising the centre of gravity of a trampolinist of mass 73 kg through a vertical height of 2.48 m (see Figure 2.92).

b When he lands on the trampoline he is brought to rest for a moment before being pushed up in the air again. If the maximum displacement of the trampoline is 0.8 m, sketch a force–displacement graph for the surface of the trampoline.

Figure 2.92 The more the trampolinist stretches, the higher the trampolinist can jump

a $W = Fs = \text{weight} \times \text{height} = mg \times h$
$W = 73 \times 9.81 \times 2.48 = 1800\,\text{J}$

b The shape of the graph is not known, so it has been drawn linearly for simplicity (Figure 2.93). The area under the graph must be approximately equal to 1800 J (ignoring any energy transferred to other forms). This means that the maximum force must be about 4500 N, so that $0.5 \times 0.8 \times 4500 = 1800\,\text{J}$.

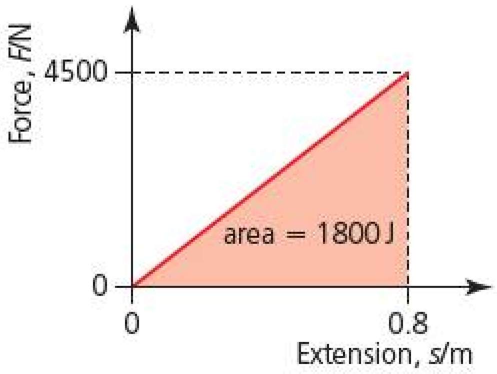


Figure 2.93

2.3.3 Solve problems involving the work done by a force.

85 A spring was 48 mm long with no force applied. When a force of 8.3 N was applied, its overall length extended to 74 mm. How much work was done on the spring?

86 Some springs are designed to be compressed (squashed) as well as stretched. The suspension system of a car is a good example of this.

a Estimate the maximum downwards force you can exert on the side of a car and how far down it will move because of your force.

b Calculate an approximate value for the work done on the spring in this process.

87 A steeper gradient on a force–extension (F–s) graph means that a spring requires a greater force to stretch it by a certain amount. In other words, the gradient of a force–extension graph represents the *stiffness* of the spring. $\Delta F/\Delta s$ is called the *spring constant*, but a similar concept can be used when stretching materials other than springs.

a Calculate the spring constant for the spring represented in Figure 2.94. Its original length was 27.0 cm.

b How much work was done when the spring was extended by 8.0 cm?

c What was the length of the spring when the force was 19 N?

d Explain why it would be unwise to use information from this graph to predict the length of the spring if a mass of 10 kg was hung from its end.

Figure 2.94

■ Additional Perspectives

Stress and strain

Force–extension graphs can be drawn for specimens of different materials (often in the shape of wires). Such graphs provide important basic information about the *stiffness* and *strength* of the actual specimens tested, whether they remain out of shape, and their ability to store energy when stretched. Of course, all this depends not only on what the specimen is made of, but also on its original length and cross-sectional area. However, if we draw stress–strain graphs (rather than force–displacement graphs), they can be applied to specimens of *any* shape of the same material.

$$\text{stress} = \frac{\text{force}}{\text{area}} \qquad \text{strain} = \frac{\text{change in length}}{\text{original length}}$$

As an example, consider Figure 2.95a, which shows the stress–strain relationship for a metal wire up to the point where it breaks. From this graph we can make the following conclusions about this metal.

- For strains up to about 0.05 (5%), any extension produced will be proportional to the force applied (stress is proportional to strain).
- For greater stresses, the metal becomes less stiff.
- The wire will be approximately twice its original length when it breaks.
- The work done in stretching the material can be found from the area under the graph if the dimensions of the wire are known.

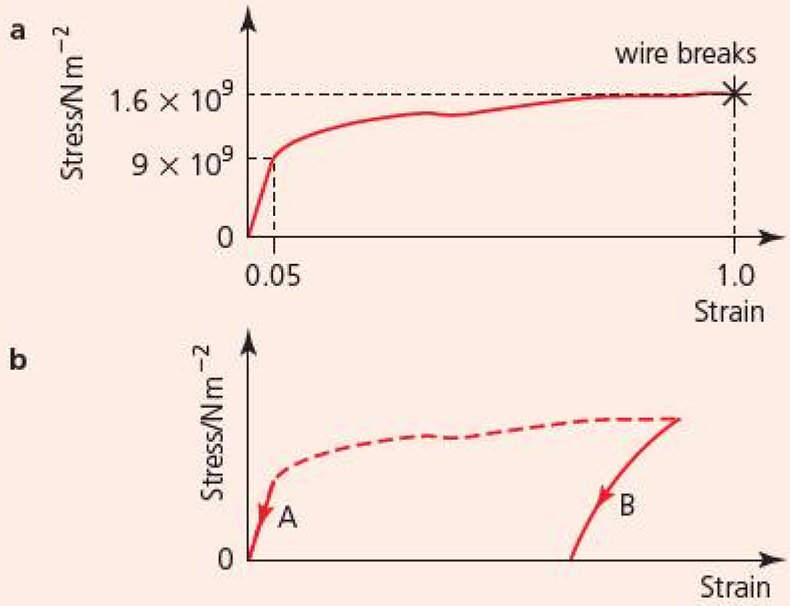


Figure 2.95 Stress–strain graph

Figure 2.95b represents the same metal, but shows what happens as the force is *removed* in two different situations. For A the maximum force on the wire did not take it beyond its limit of proportionality and the wire returned to its original shape. This is called 'elastic' behaviour. For B the wire stretched a lot more before the force was removed and the wire was permanently deformed from its original shape. This is known as 'plastic' behaviour.

Questions

1 a If the original length of the wire was 1.20 m and its cross-sectional area was $3.8 \times 10^{-8}\,\text{m}^2$, calculate the force required to break the wire and its length just before it broke.
  b Estimate the force needed to break another wire of the same metal, but of length 2.4 m
2 Calculate the ratio stress/strain for the linear section of the graph. This is the stiffness of the metal. What are the units of stiffness?
3 Approximately how much work was done in stretching this wire to its breaking point?

Different forms of energy

2.3.7 (part) **List** different forms of energy.

When anything happens (i.e. something changes) it is because energy has been transferred from one form to another or from one place to another. It is important to have an overview of the most important forms of energy. In this section we will discuss the following:

- gravitational (potential) energy
- nuclear (potential) energy
- chemical (potential) energy
- electric (potential) energy, including energy transferred by an electric current
- elastic strain (potential) energy
- thermal energy (heat)
- radiant energy
- kinetic energy (including wind energy)
- internal energy
- mechanical wave energy (including sound and waves on water).

Potential energies

2.3.5 **Outline** what is meant by change in gravitational potential energy.

When work is done to stretch a spring, we say that energy has been transferred to it and that the energy is then in the form of elastic strain energy. (**Elastic** behaviour means that when the force is removed the material will return to its original shape. When we say that a material is **strained**, we mean that it has changed shape.) While the spring is stretched we may say that energy is 'stored' in it and that the energy is still available for later use, possibly to do something useful like running a toy or a wind-up watch. Elastic strain energy is just one of several forms of potential energy.

Potential energies are energies that are stored. In general, potential energies exist where there are forces between objects. The following is a list of the important potential energies.

- When a mass is lifted off the ground against the force of gravity, we say that it has gained **gravitational potential energy**. This energy is stored in the mass. If the mass is allowed to fall, it loses some of its gravitational potential energy and this energy could be transferred to do something useful.
- **Nuclear (potential) energy** is associated with the forces between particles in the nuclei of all atoms.
- **Chemical (potential) energy** is associated with the forces between electrons and other particles in atoms and molecules.
- **Electric (potential) energy** is associated with the forces between electric charges.
- **Elastic strain (potential)** energy is associated with the forces needed to change the shape of objects.

Energy of motion

2.3.4 **Outline** what is meant by kinetic energy.

Work has to be done on any object to make it move from rest or to make it move faster. This means that all moving objects have a form of energy. This is called **kinetic energy**. ('Kinetic' means related to motion.) Wind energy is the kinetic energy of moving air.

Energy being transferred

While energy is being transferred from place to place, we may refer to the following forms of energy:

- **electrical energy** carried by an electric current
- **thermal energy** (heat) transferred because of a difference in temperature
- **radiant energy** transferred by electromagnetic waves ('light energy' or 'solar energy' are specific examples of electromagnetic radiant energy)
- **mechanical waves** (like water and sound waves), involve oscillating masses. They involve a combination of kinetic energy and potential energy (see Chapter 4).

Internal energy

There is a large amount of energy *inside* substances because of the random motions of the particles they contain and the forces between them. The total random kinetic energy and electric potential energy of all the particles inside a substance is called its *internal energy*.

Whenever objects move in everyday situations, some or all of their kinetic energy will be transferred to internal energy because of frictional forces (and then thermal energy spreads out into the surroundings).

Examples of energy transfers

2.3.7 (part) **Describe** examples of the transformation of energy from one form to another

Everything that happens in our lives involves transfers of energy, so we do not need to look far for examples. When we switch on an electric fan, energy transferred from the electric current is changed into kinetic energy of the fan and then kinetic energy of the moving air. After the fan is turned off, all of that energy will have spread out into the environment as thermal energy and then internal energy. But where did the current get its energy from? Maybe in a nearby power station the chemical energy in oil was transferred to internal energy and kinetic energy of steam, which was then converted to kinetic energy in the turbines that generated the electricity.

The Sun provides most of the energy we use on this planet. Nuclear energy in the hydrogen nuclei in the Sun is converted to internal energy and radiation. The radiation that reaches Earth is transferred to chemical energy in plants (and then animals). Oil comes from what remains of plants and small organisms that died millions of years ago and have then decayed in the absence of oxygen.

88 Make a flow chart showing all the energy transfers listed in the two paragraphs above.

89 Name devices whose main use is to perform the following energy transfers:
- **a** electricity to sound
- **b** chemical to electricity
- **c** nuclear to electricity
- **d** sound to electricity
- **e** chemical to electromagnetic radiation
- **f** chemical to kinetic
- **g** elastic strain to kinetic
- **h** kinetic to electricity
- **i** chemical to internal
- **j** electromagnetic radiation to electricity.

Conservation of energy

2.3.6 **State** the principle of conservation of energy.

The study of the different forms of energy and the transfers between them is the central theme of physics. The main reason that the concept of energy is so important lies in the **principle of conservation of energy**.

The principle of conservation of energy states that energy cannot be created or destroyed; it can only be transferred from one form to another.

In other words, the total amount of energy in the universe is constant. This is one of the most important principles in the whole of science. If, in any particular situation, energy seems to have just appeared or disappeared, then we know for absolute certain that there must be an explanation for where it has come from or where it has gone. It is worth repeating at this point that *all* everyday processes spread (**dissipate**) some energy into the surroundings, and this dissipated energy is then of no further use to us.

In order to use the principle of conservation of energy, we must be able to measure and calculate energy transfers. Equations for calculating quantities of energy in its different forms are essential knowledge for all students of physics.

Calculating changes of mechanical energies

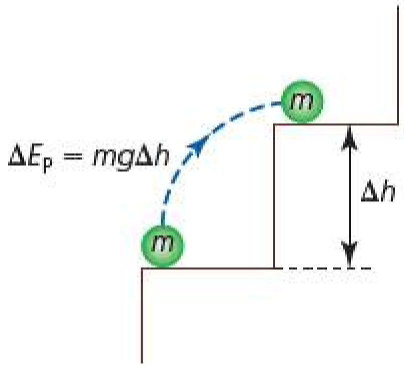


Figure 2.96 A mass gaining gravitational potential energy

Gravitational potential energy

When we lift a mass we transfer energy to it and, because we are using a force to move it, this kind of energy transfer can be described as **doing work**. We have already seen (in Worked example 17) that we can calculate the work done from $W = mgh$. The difference in energy the mass has because of the work done to move it up or down to a new position is called the **change in gravitational potential energy**.

The symbol E is used to represent energy in general and E_p is used to represent gravitational potential energy in particular. When a mass m is raised a vertical height Δh, as shown in Figure 2.96, the change in gravitational potential energy, ΔE_p, can be calculated from:

$$\Delta E_p = mg\Delta h$$

This equation is included in the IB *Physics data booklet.*

Note that this equation enables us to calculate *changes* in gravitational potential energy between various positions. (If we wanted to calculate the 'total' gravitational energy of a mass, we would have to ask the question, gravitational potential energy compared to where? This might be the floor, sea level or the centre of the Earth, for example. This is discussed further in Chapter 9.)

Worked example

20 How much gravitational potential energy is gained in each situation?

a An 818 g book is raised 91 cm from the floor to a desk.

b A 56 kg boy walks up eighteen 14 cm steps from the second to the third floor of a building.

c The same boy walks down to the first floor from the second floor.

a $\Delta E_p = mg\Delta h$
$\Delta E_p = 0.818 \times 9.81 \times 0.91$
$\Delta E_p = 7.3\,\text{J}$

b $\Delta E_p = mg\Delta h$
$\Delta E_p = 56 \times 9.81 \times (18 \times 0.14)$
$\Delta E_p = 1400\,\text{J}$

c $-1400\,\text{J}$. The negative sign means that the gravitational potential energy is less than on the second floor.

2.3.11 Solve problems involving energy.

90 **a** What is the difference in gravitational potential energy of a 74 kg mountain climber between the top and bottom of a 2800 m mountain?

b Is the value of the gravitational field strength the same at the top and bottom of the mountain? If not, does it affect your answer to part **a** of this question?

91 A cable car rises a vertical height of 700 m in a total distance travelled of 6 km (Figure 2.97).

a How much gravitational potential energy must be transferred to a car of mass 1800 kg during the journey if it has eight passengers with an average mass of 47 kg?

b Suggest reasons why a lot more energy (than your answer to part **a**) has to be transferred in making this journey.

Figure 2.97 Ngong Ping cable car in Hong Kong

92 **a** How much gravitational energy must be transferred *to* an elevator of total mass 2340 kg when it rises 72 floors, each of average height 3.1 m?

b The same amount of gravitational potential energy must be transferred *from* the elevator when it comes down again. Rather than waste this energy, elevators use counterweights (see Figure 2.98). As the elevator comes down the counterweight goes up, and when the elevator goes up, the counterweight comes down. Explain the advantage of this system.

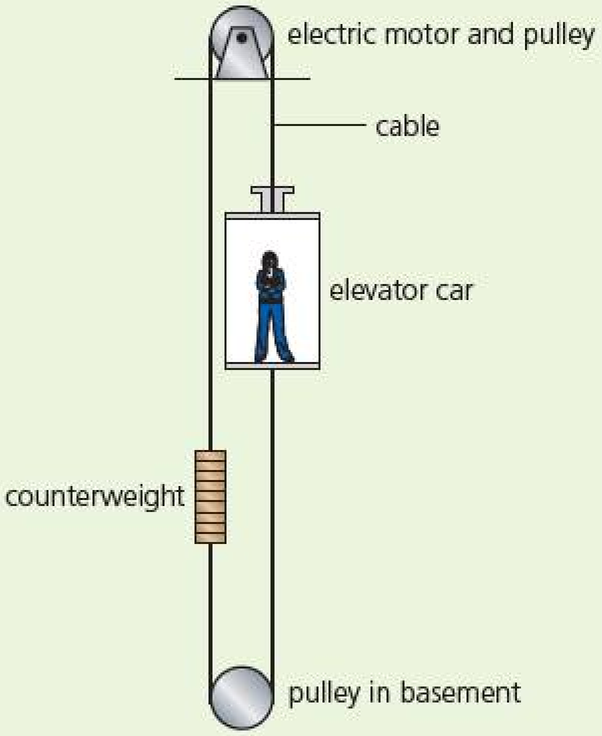


Figure 2.98 An elevator and its counterweight

Kinetic energy

Consider a mass m accelerated uniformly from rest by a constant force F. The work done can be calculated from $W = Fs$. But $F = ma$ and $v^2 = u^2 + 2as$ (with $u = 0$), so the equation can be re-written as:

$$W = Fs = (ma) \times \left(\frac{v^2}{2a}\right) = \frac{1}{2}mv^2$$

That is, the kinetic energy of a moving mass, E_K, can be calculated from:

$$E_K = \frac{1}{2}mv^2$$

This equation is included in the IB *Physics data booklet*.

Worked example

21 Calculate the kinetic energy of car of mass 1320 kg travelling at 27 m s^{-1}.

$$E_K = \frac{1}{2}mv^2$$

$$E_K = \frac{1}{2} \times 1320 \times 27^2$$

$$E_K = 4.8 \times 10^5\,\text{J}$$

Since momentum is given by $p = mv$, the equation for kinetic energy can be re-written and linked to momentum in the following form. This equation is particularly useful when dealing with the motion of atomic particles.

$$E_K = \frac{p^2}{2m}$$

This equation is included in the IB *Physics data booklet*.

2.3.11 Solve problems involving momentum and energy.

93 **a** Calculate the kinetic energy of:

i a soccer ball (mass 440 g) kicked with a speed of 24 m s^{-1}

ii a 24 g bullet moving at 420 m s^{-1}.

b Estimate the kinetic energy of a woman sprinter in an Olympic 100 m race.

94 A mass of 2.4 kg in motion has kinetic energy of 278 J. What is its speed?

95 An electron has a mass of 9.1×10^{-31} kg. If it moves at 1% of the speed of light, what is its kinetic energy? (Speed of light = 3.00×10^8 m s^{-1}.)

96 A car travelling at $10\,m\,s^{-1}$ has kinetic energy of 100 kJ. If its speed increases to $30\,m\,s^{-1}$, what is its new kinetic energy?

97 **a** What is the momentum of a 1340 kg car that has 4.30×10^5 J of kinetic energy?
b What is the kinetic energy of a 340 g mass that has a momentum of $8.3\,kg\,m\,s^{-1}$?
c What is the mass of a sub-atomic particle that has a momentum of $2.50 \times 10^{-23}\,kg\,m\,s^{-1}$ when its kinetic energy is 3.43×10^{-16} J?

Additional Perspectives

Making work easier

When lifting a heavy object, the amount of (gravitational potential) energy that we need to transfer to it is decided only by its weight and the vertical height (mgh). For example, to lift a 100 kg mass a height of 1 m requires about 1000 J of work to be done on it. Although 1000 J is not a lot of work, that does not mean that we can do this job easily.

There are two main reasons why this job could be difficult. Firstly, we may not be able to transfer that amount of energy in the time required to do the work. Another way of saying this is that we may not be **powerful** enough. (Power is discussed later in this chapter.) Secondly, we may not be **strong** enough because we are not able to provide the required upwards force of about 1000 N. Power and strength are often confused with each other in everyday language.

Lifting (heavy) weights is a common human activity and many types of simple 'machines' were invented many years ago to make this type of work easier, by reducing the force needed. These include the ramp (inclined plane), the lever and the pulley, as shown in Figure 2.99.

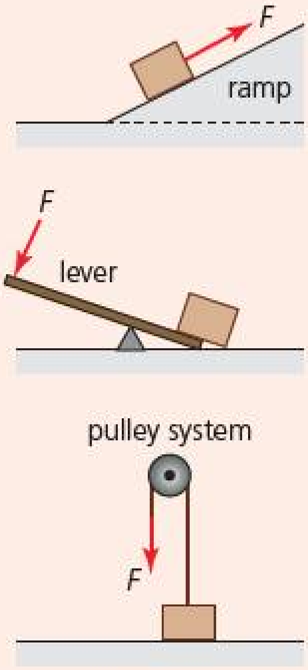


Figure 2.99 Simple machines

In each of these simple machines the force needed to do the job is reduced, but the distance moved by the force is increased. If there were no energy losses (any energy losses would be mainly due to friction), the work done by the force (Fs) would equal the useful energy transferred to the object being raised (mgh). In practice, because of energy losses, we would transfer more energy using a machine than if we lifted the load directly, without the machine. This is not a problem because we are usually much more concerned about how easy it is to do a job, rather than the total energy needed or the efficiency of the process.

Question

1 Figure 2.100 shows a car jack being used to raise one side of a car.

a Estimate how much gravitational energy must be transferred to the car so that a tyre can be changed.

b By changing the design of the jack it is possible in theory to raise the car with almost any sized force that we choose. Estimate a convenient force that most people would be happy to use with a car jack.

c Use your answers to **a** and **b** to estimate how far the force must move in order to raise the car.

d Explain how it can be possible for the force to move such a large distance.

e Discuss why it might be useful for there to be a lot of friction when using a car jack.

Figure 2.100 Changing a car tyre using a simple machine (car jack)

Transfers between gravitational energy and kinetic energy

When an object falls towards the ground it transfers gravitational potential energy to kinetic energy. When we throw or project an object vertically upwards we have to give it kinetic energy, which is then transferred to gravitational potential energy. At its highest point the kinetic energy of the object will be zero and its gravitational potential energy will be greatest.

For an object of mass m that falls from rest moving through a vertical height Δh *without air resistance*, we can write:

$$mg\Delta h = \frac{1}{2}mv_{max}^2$$

It follows that:

$$v_{max}^2 = 2g\Delta h$$

This useful equation is equivalent to the equation of motion: $v^2 = u^2 + 2as$, with $u = 0$, a written as g and s written as Δh. The same equation can be used for objects moving upwards. The equation can also be used for objects moving up or down slopes, but only if we assume that the friction is negligible.

Worked examples

22 A ball is dropped from a height of 18.3 m. With what speed does it hit the ground? What assumption did you have to make in order to do this calculation?

$$mg\Delta h = \frac{1}{2}mv_{max}^2$$
$$9.81 \times 18.3 = \frac{1}{2} \times v^2$$
$$v = 18.9\,\mathrm{m\,s^{-1}}$$

This assumes that there was no energy dissipated because of air resistance.

23 The ball shown in Figure 2.101 was released from rest in position A. It accelerated down the slope and had its greatest speed at the lowest point. It then moved up the slope on the other side, reaching its highest point at B.

a Explain why B is lower than A.

b Describe the motion of the ball after leaving position B, explaining the energy transfers until it finally comes to rest at F.

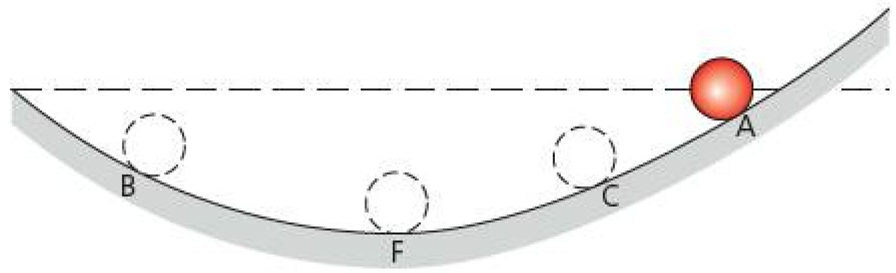


Figure 2.101

a As the ball rolls backwards and forwards there is a continuous transfer of energy between gravitational potential energy and kinetic energy. The ball would only be able to reach the same height as A, and regain all of its gravitational potential energy, if there was no friction. Because of friction some kinetic energy of the moving ball is transferred to internal energy in the ball and the slope, so it cannot reach the same height.

b The ball accelerates as it moves down the slope. When it reaches F, all of the extra gravitational potential it had because of its position at B has been transferred, and the ball has its greatest speed. The ball decelerates as it moves up to C, as kinetic energy is transferred back to gravitational potential energy. At all times that the ball is moving, some energy is being transferred to internal energy, so that the maximum height and maximum speed of the ball gets less and less. Eventually the ball stops at F.

2.3.11 Solve problems involving energy.

98 From what height would you have to drop a mass for it to hit the ground with a speed of exactly $5\,\mathrm{m\,s^{-1}}$? (Assume no air resistance.)

99 A stone is thrown vertically up with a speed of $14\,\mathrm{m\,s^{-1}}$. What is the maximum possible height that it could rise to?

100 In order to estimate the height of a building, a ball is repeatedly thrown vertically upwards and its speed changed until it just reaches the top of the building (see Figure 2.102).

a If the ball takes 4.2 s from the moment it is released to the time it takes to hit the ground, estimate the height of the building.

b What are the main sources of error in this experiment?

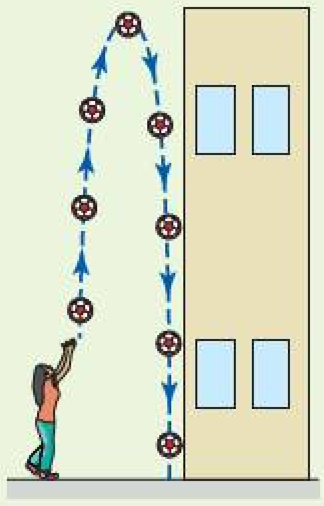


Figure 2.102

101 A ball is allowed to roll down the slope shown in Figure 2.103a.

a If the surface is friction-free, what is its speed at point P?

b The same ball is then allowed to roll down the friction-free slope shown in Figure 2.103b. What is its speed at Q?

c Explain why the ball reaches point P more quickly than Q.

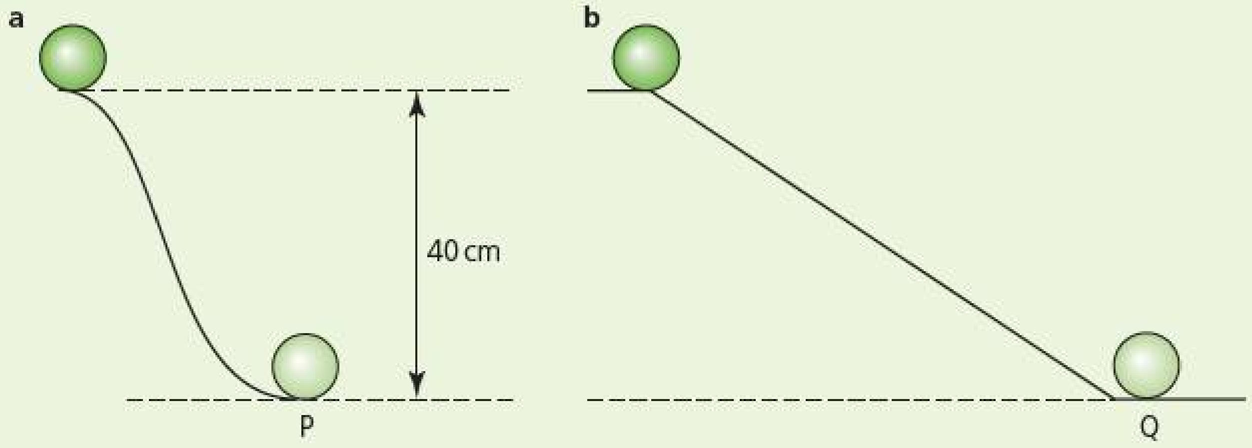


Figure 2.103

102 A rubber ball was dropped vertically onto the floor from a height of 1.0 m. After colliding with the floor it bounced up to a height of 60 cm. After the next bounce it only reached a height of 36 cm.

a In what form is the energy when the ball is in contact with the ground?

b List the energy transfers that take place from the release of the ball until the top of the first bounce.

c Calculate the speed of the ball just before and just after it touches the floor during the first bounce.

d Predict the height of the third bounce. What assumption did you make?

103 Figure 2.104 shows a pendulum. Calculate the maximum velocity of the pendulum when released from rest.

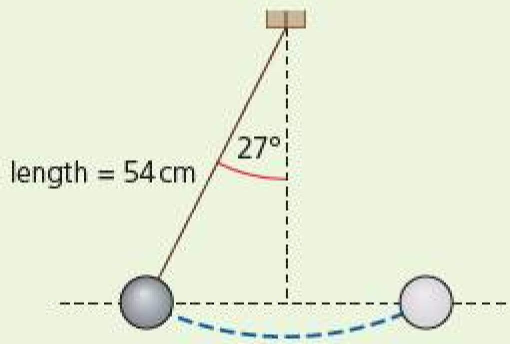


Figure 2.104

Doing work to change motion

Using forces to accelerate or decelerate objects is a very common human activity. The relationship, work done equals change of kinetic energy, therefore has many uses. The equation is:

$$Fs = \frac{1}{2}mv^2 - \frac{1}{2}mu^2$$

The force F is assumed to be constant and in the same direction as the motion. Alternatively, an average force may be used. Where the force does not vary linearly, the area under a force–distance graph can be used to determine the change in kinetic energy.

This equation can be used for objects which are accelerating or decelerating. If the motion starts or ends with an object at rest (zero kinetic energy), we can write:

$$Fs = \frac{1}{2}mv^2$$

This equation shows us that when a force is used to stop a moving object, the smaller the force used, the greater the stopping distance of the object. (Or, the greater the force used, the smaller the stopping distance.) To reduce the forces in an impact, the stopping distance (and time) should be made as large as possible. For example, jumping into sand involves less force on our bodies than jumping onto solid ground.

If the speed of an object is decreasing because of work done by friction and/or air resistance, its kinetic energy will be transferred to internal energy, thermal energy and maybe some sound. We say that the energy has been **dissipated**, which means that it has spread out into the surroundings. This very common process is happening around us all the time.

Moving objects may also slow down because they collide with other things. As with friction, this will result in the dissipation of energy, but usually some energy is also used to **deform** the shape of the objects. The change of shape may be temporary, but it is often permanent.

Worked examples

24 A constant resultant force of 2420 N acts on a 980 kg car at rest. What is its speed after the car has moved 100 m? (Ignore resistive forces.)

$$Fs = \frac{1}{2}mv^2$$
$$2420 \times 100 = \frac{1}{2} \times 980 \times v^2$$
$$v = 22\,\text{m s}^{-1}$$

25 What average force is needed to stop a car of mass 1350 kg travelling at 28 m s^{-1} in a distance of 65 m?

$$Fs = \frac{1}{2}mv^2$$
$$F \times 65 = \frac{1}{2} \times 1350 \times 28^2$$
$$F = 8100\,\text{N}$$

26 In a car accident, the driver must be stopped from moving forward (relative to the car) and hitting the steering wheel in a distance of less than 25 cm.

a If he has a mass of 80 kg, what force is needed to stop him in this distance if he and the car were travelling at 10 m s^{-1} immediately before the impact?

b How is this force applied to him?

a
$$Fs = \frac{1}{2}mv^2$$
$$F \times 0.25 = \frac{1}{2} \times 80 \times 10^2$$
$$F = 16\,000\,\text{N}$$

b The force is applied mainly by an inflated air bag (Figure 2.105) or from a seat belt. It is a very large force, but considerably less than the force that would act on him if he collided with the steering wheel. The effect of the force is also reduced because it is spread over a large area. (This reduces the pressure.)

Figure 2.105 An air bag greatly reduces the forces in an accident

2.3.11 Solve problems involving work and energy.

104 What average force is needed to reduce the speed of a 400 000 kg train from 40 m s^{-1} to 10 m s^{-1} in a distance of 1.0 km?

105 'Crumple zones' are a design feature of most vehicles (Figure 2.106). They are designed to compress and deform permanently if they are in a collision. Use the equation $Fs = \frac{1}{2}mv^2$ to help explain why a vehicle should not be too stiff and rigid.

Figure 2.106 The front of the car is deformed but the passenger compartment is intact

106 A bungee jumper (Figure 2.107) of mass 61 kg is moving at 23 m s^{-1} when the rubber bungee cord begins to become stretched.
a Calculate her kinetic energy at that moment.
Figure 2.108 shows how the extension of the cord varies with the applied force.
b What quantity is represented by the area under this graph?
c Describe the relationship between force and extension shown by this graph.
d Use the graph to estimate how much the cord has extended by the time it has brought the jumper to a stop.

Figure 2.107 Bungee jumping in Taupo, New Zealand

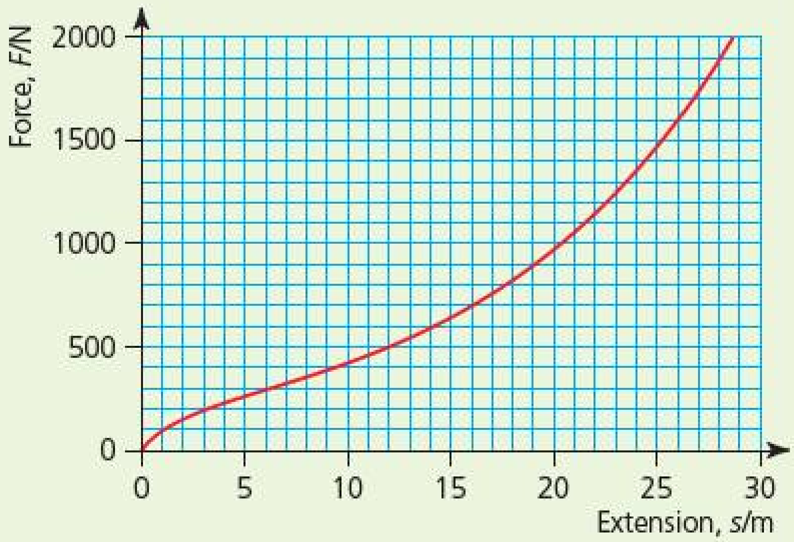


Figure 2.108 Force–extension graph for a bungee cord

107 A pole-vaulter of mass 59.7 kg falls from a height of 4.76 m onto foam.
a Calculate the maximum kinetic energy on impact.
b Will air resistance have had a significant effect in reducing the velocity of impact? Explain your answer.
c If the foam deforms by 81 cm, calculate the average force exerted on the pole-vaulter.

108 Explain why you should always bend your legs when jumping down onto the ground.

109 **a** A nail in a piece of wood is struck once by a 1.24 kg hammer moving at 12.7 m s^{-1}. If the nail is pushed 15 mm into the wood, what was the average force exerted on the nail?
b What assumption did you make?

■ Additional Perspectives

Kinetic energy of trains

The kinetic energy that must be given to a long, fast-moving train is considerable. Values of 10^8 J or more would not be unusual. When the train stops all of that energy has to be transferred to other forms and, unless the energy can be recovered, the same amount of energy then has to be transferred to accelerate the train again. This is very wasteful, so the train and its operation should be designed to keep the energy wasted to a minimum. One way of doing this is to make sure that big, fast trains stop at as few stations as possible, perhaps only at their origin and their final destination.

Most ways of stopping moving vehicles involve braking systems in which the kinetic energy of the vehicle is transferred to internal energy because of friction. The internal energy is dissipated into the surroundings as thermal energy and cannot be recovered. A lot of research has gone into designing efficient 'regenerative braking systems' in recent years, usually involving the generation of an electric current, which can be used to transfer energy to chemical changes in batteries.

Figure 2.109 The Light Rail Transit trains on the SBS network in Singapore have regenerative braking

Small electric trains, which are often operated underground or on overhead tracks, are a feature of most large cities around the world (see Figure 2.109). Such trains usually have stations every few kilometres or less, so regenerative braking systems and other energy saving policies are very important. When designing a new urban train system, it has been suggested that energy could be saved by having a track shaped as shown in Figure 2.110.

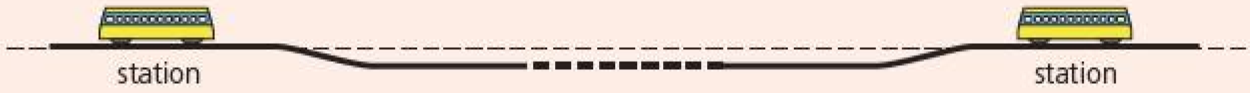


Figure 2.110 Possible track profile

Questions

1 Explain exactly what is meant by 'saving energy'.
2 Explain the reasoning behind the proposal shown in Figure 2.110.
3 Discuss other features of an urban train system that could 'save energy'.
4 Discuss which kind of engine would be the best choice for such trains.

Elastic and inelastic collisions

2.3.8 Distinguish between elastic and inelastic collisions.

When objects come into contact (or otherwise interact) with each other and exert forces on each other for relatively short periods of time, the interactions can be described as 'collisions'. We normally expect some of the kinetic energy of the objects to be dissipated in any collision, but it is important to define the *extreme* case.

A collision in which the total kinetic energy of the masses is the same before and after the collision is known as an **elastic collision**.

In our everyday, large-scale world, elastic collisions are not possible because some energy is *always* dissipated into the surroundings. However, on the small scale, collisions between particles, such as molecules in a gas, are usually elastic. It should be noted that the concepts of internal energy, sound and deformation cannot be used sensibly to describe individual molecules.

Momentum is conserved in *all* collisions and other types of interaction. There are no exceptions to this law. In the theoretical extreme of an elastic collision, kinetic energy is also conserved and the two equations representing this can be combined together to predict what would happen if an elastic collision occurred. For example, if a mass collides elastically with another identical mass at rest, the moving mass will stop and the other mass will move off with the same velocity as the first mass.

Collisions in which some or all kinetic energy is transferred to other forms of energy are called **inelastic collisions**. All collisions of everyday objects are inelastic.

A collision in which the objects stick together is called a **totally inelastic** collision.

In an '**explosion**', masses which were originally at rest with respect to each other are propelled in different directions, so that there is greater kinetic energy after the explosion than before. By definition this type of interaction clearly cannot be described as *elastic* and it may be considered to be similar to a totally *inelastic* collision in reverse.

Power and efficiency

When energy is transferred by people, animals or machines to do something useful, we are often concerned about how long it takes for the change to take place. If the same amount of useful work is done by two people (or machines), the one that does it faster is said to be more *powerful*. (In everyday use the word 'power' is used more vaguely, often related to strength and with no connection to time.) If the same amount of useful work is done by two people (or machines), the one that does it by using the least overall energy is said to be more *efficient*.

Power

2.3.9 **Define** *power*

Power is the rate of transferring energy. It is defined as:

$$\text{power} = \frac{\text{energy transferred}}{\text{time taken}}$$

The symbol for power is P and it has the SI unit the watt, W ($1\,\text{W} = 1\,\text{J}\,\text{s}^{-1}$). The units mW, kW, MW and GW are also in common use. The following are some examples of values of power in everyday life.

- A woman walking up stairs transfers chemical energy to gravitational energy at a rate of about 300 W.
- A 18 W light bulb transfers energy from electricity to light and thermal energy at a rate of 18 joules every second.
- A 0.0001 W calculator transfers energy at a rate of 0.0001 joules every second.
- A 2 kW water heater transfers energy from electricity to internal energy at a rate of 2000 joules every second.
- A typical family car may have a maximum output power of 100 kW.
- A 500 MW oil-fired power station transfers chemical energy to electrical energy at a rate of 500 000 000 joules every second.
- In many countries homes use electrical energy at an average rate of about 1 kW.

In this chapter, the energy transfer we are concerned with is that of doing work, W, so that we can write:

$$\text{power} = \frac{\text{work done}}{\text{time taken}}$$

$$P = \frac{\Delta W}{\Delta t}$$

Worked examples

27 Calculate the average power of a 75 kg climber moving up a height 30 m in two minutes.

$$P = \frac{\Delta W}{\Delta t}$$

$$P = \frac{mg\Delta h}{\Delta t}$$

$$P = \frac{75 \times 9.81 \times 30}{2 \times 60}$$

$$P = 180\,\text{W}$$

28 What average power is needed to accelerate a 1200 kg car from rest to 20 m s^{-1} in 8 s?

$$P = \frac{\Delta W}{\Delta t}$$

$$P = \frac{\text{kinetic energy gained}}{\text{time taken}}$$

$$P = \frac{\frac{1}{2}mv^2}{\Delta t}$$

$$P = \frac{0.5 \times 1200 \times 20^2}{8}$$

$$P = 30\,000\,\text{W} \ (= 30\,\text{kW})$$

Power transferred when travelling at constant velocity

The calculation for Worked example 28 ignored the large amount of work done in overcoming air resistance. When any vehicle travels at a constant velocity, *all* of the work done is used to overcome resistance, rather than produce an acceleration. If we replace ΔW with $F\Delta s$ in the equation for power (where F is the resistive force, which, at constant velocity, is equal and opposite to the forward force), we get:

$$P = \frac{F\Delta s}{\Delta t}$$

and, since $\frac{s}{t} = v$ it follows that:

power to maintain constant velocity = resistive force × speed

$$P = Fv$$

This equation is given in the IB *Physics data booklet*.

Worked example

29 If the maximum forward force of propulsion on a boat is 2750 N, what is the maximum speed for a boat powered by a 4.00 kW engine?

$$P = Fv$$

$$4000 = 2750 \times v$$

$$v = 1.45\,\text{m s}^{-1}$$

Efficiency: comparing input and output powers

2.3.10 Define and **apply** the concept of *efficiency*.

It is an ever-present theme of physics that, whatever we do, some of the energy transferred is 'wasted' because it is transferred to less 'useful' forms. In mechanics this is usually because friction or air resistance transfer kinetic energy to internal energy and thermal energy. The *useful* energy we get out of any energy transfer is *always* less than the *total* energy transferred.

When an electrical water heater is used, nearly all of the energy transferred makes the water hotter and it can therefore be described as 'useful', but when a light bulb is used only some of the energy is transferred to useful light (most of the rest is transferred to thermal energy). Driving a car involves transferring chemical energy from the fuel and the useful energy is considered to be the kinetic energy of the vehicle, although at the end of the journey there is no kinetic energy remaining.

A process that results in more useful energy output (for a given energy input) is described as being more efficient. **Efficiency** is defined as follows:

$$\text{efficiency} = \frac{\text{useful energy output}}{\text{total energy input}}$$

Because it is a ratio of two energies, efficiency has no units. It is often expressed as a percentage. It should be clear that efficiency will *always* be less than 1 (or 100%).

An alternative definition of efficiency is obtained by dividing energy by time to get power:

$$\text{efficiency} = \frac{\text{useful power output}}{\text{total power input}}$$

It is possible to discuss the efficiency of *any* energy transfer, such as the efficiency with which our bodies transfer the chemical energy in our food to other forms. However, the concept of efficiency is most commonly used when referring to machines and engines of various kinds, especially those in which the input energy or power is easily calculated. Sometimes we need to make it clear exactly what we are talking about. For example, when discussing the efficiency of a car, do we mean only the engine, or the whole car in motion along a road with all the energy dissipation due to resistive forces?

The efficiencies of machines and engines usually change with the operating conditions. For example, there will be a certain load at which an electric motor operates with maximum efficiency; if it is used to raise a very small or a very large mass it will probably be inefficient. Similarly, cars are designed to have their greatest efficiency at a certain speed, usually about $60\,\text{km}\,\text{h}^{-1}$. If a car is driven faster (or slower), then its efficiency decreases, which means that more fuel is used for every kilometre travelled.

Car engines, like all other engines that rely on burning fuels to transfer energy, are inefficient because of fundament physics principles (see Chapter 10). There is nothing that we can do to change that, although better engine design and maintenance can make some improvements to efficiency.

Worked example

30 a In an experiment, energy is provided to an electric motor at a rate of 0.80 W (Figure 2.111). If it raises a 20 g load at 80 cm in 1.3 s, what is its efficiency?

b What happens to all the energy that is not transferred usefully to the load?

a Power used to raise mass, $P = \frac{mg\Delta h}{\Delta t}$

$$P = \frac{0.02 \times 9.81 \times 0.80}{1.3}$$

$$P = 0.12\,\text{W}$$

$$\text{efficiency} = \frac{\text{useful power output}}{\text{total power input}}$$

$$\text{efficiency} = \frac{0.12}{0.80} = 0.15 \text{ (or 15\%)}$$

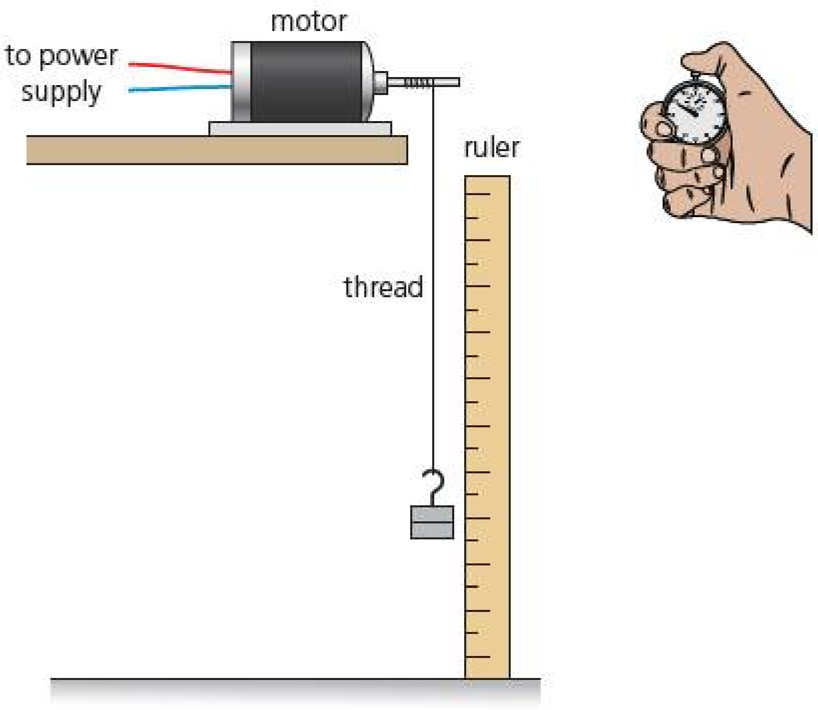


Figure 2.111 Experiment to measure the efficiency of a motor

b 85% of the energy transferred by electricity to the motor does not go usefully to the increased gravitational energy of the load. The wasted energy goes mostly to internal energy in the motor, which is then transferred as thermal energy to the surroundings. In addition, some energy will have been transferred to sound and some energy was used to stretch the string that connected the load to the motor (elastic strain energy).

2.3.11 Solve problems involving energy and power.

110 **a** How much energy must be transferred to lift twelve 1.7 kg bottles from the ground to a shelf that is 1.2 m higher?
b If this task takes 18 s, what was the average useful power involved?

111 Estimate the output power of a motor that can raise an elevator of mass 800 kg and six passengers 38 floors in 52 s. (Assume there is no counterweight.)

112 How large are the resistive forces opposing the motion of a car that is operating with an output power of 23 kW at a constant speed of 17 m s^{-1}?

113 **a** What is the output power of a jet aircraft that has a forward thrust of 660 000 N when travelling at its top speed of 950 km h^{-1} (264 m s^{-1}) through still air?
b If homes use on average 1 kW of electrical power, how many homes could this same amount of power supply?
c Discuss whether the rapidly increasing use of aircraft for travel around the world should be encouraged or discouraged.

114 The output from a power station is 325 MW. What is the power input if it is 36% efficient?

115 A man uses a ramp to push a 60.0 kg box onto the back of a truck as shown in Figure 2.112. To lift the box directly would require a vertical force of 590 N, but by using the ramp this force is reduced.

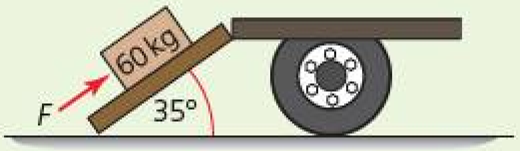


Figure 2.112

a Calculate the minimum force needed to push the box up the ramp if there is no friction.
b In practice the force needed was 392 N. What was the size of the frictional force? Explain why this force can be useful.
c The length of the ramp is 2.11 m. Calculate the work done in pushing the box along the ramp from the ground onto the truck.
d The useful energy transferred to the box is its gravitational potential energy. Calculate the efficiency of using the ramp.

2.4 Uniform circular motion

An object moving along a circular path with a constant speed has a continuously changing velocity because its direction of motion is changing all the time. From Newton's first law, we know that any object that is not moving in a straight line must be accelerating and therefore it must have a resultant force acting on it, even if it is moving with a constant speed.

Perfectly uniform motion in complete circles may not be a common everyday observation, but the theory for circular motion can also be used with objects, such as people or vehicles, moving along arcs of circles and around corners. This theory is also very useful when discussing the orbits of planets, moons and satellites. It is also needed to explain the motion of subatomic particles in magnetic fields, which is covered in later chapters.

Imagine yourself to be standing in a train on a slippery floor, holding onto a post (Figure 2.113). While you and the train are travelling in a straight line with a constant speed (constant velocity) there is no resultant force acting on you and you do not need to hold onto the post, but as soon as the train changes its motion (accelerates) there needs to be a resultant force on you to keep you in the same place in the train. If there is little or no friction with the floor, the post is the only thing that can exert a force on you to change your motion. The directions of these forces (from the post) are shown in the diagram for different types of acceleration. If the post pushes or pulls on you, then by Newton's third law, you must be pushing or pulling on the post, and that is the force you would be most aware of.

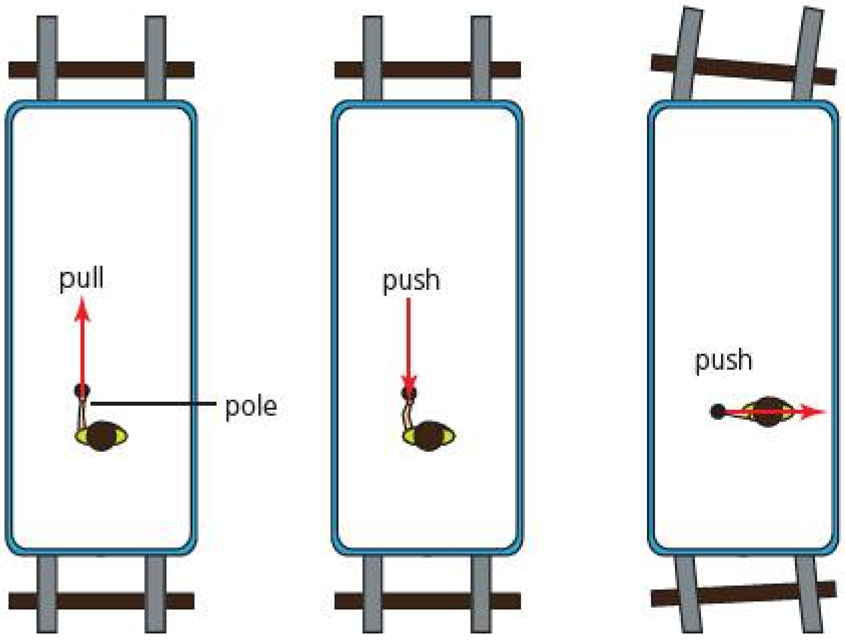


Figure 2.113 Forces which make a passenger accelerate in a train

In particular, note the direction of the force needed to produce a curved, circular path, perpendicular to motion.

Vector diagrams for circular motion

2.4.1 **Draw** a vector diagram to illustrate that the acceleration of a particle moving with constant speed in a circle is directed towards the centre of the circle.

Figure 2.114 shows four random positions of the same mass, m, moving in a circle with a constant speed. The blue arrows represent the instantaneous velocity, v, of the mass at various places around the circle. The velocities are always directed along *tangents* to the circle. A force, F, causing circular motion is called a **centripetal force** and it always acts perpendicularly to the velocity, inwards towards the centre of the circle, as shown by the red arrows in Figure 2.114. This resultant force acting on the mass always has the same magnitude but it is continuously changing direction.

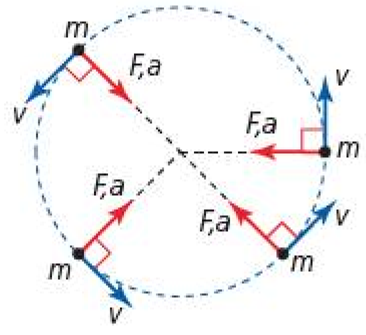


Figure 2.114 The directions of three vectors (velocity, force and acceleration) during circular motion

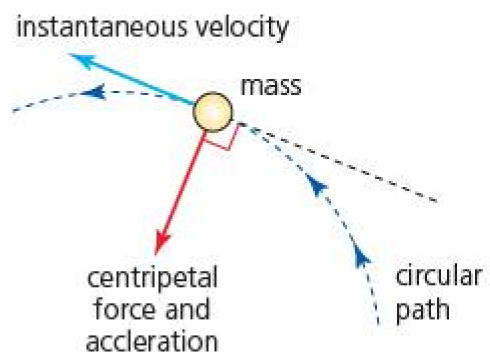


Figure 2.115 Centripetal force and acceleration

As we know, a resultant force causes an acceleration, *a*. Therefore, a centripetal force towards the centre results in a **centripetal acceleration**, also towards the centre of the circle, but there is no movement in that direction. This is shown more clearly in Figure 2.115.

Because the force is always perpendicular to the motion, no work is done by any centripetal force which makes an object move in a circle.

Examples of forces producing circular motion

2.4.3 Identify the force producing circular motion in various situations.

The following list gives some examples of circular motion and the forces causing them.

- If an object is whirled around on the end of a string in a nearly horizontal circle, the centripetal force is provided by the tension in the string (see Figure 2.120).
- Gravity provides the centripetal force that keeps planets, moons and satellites in orbit.
- Clothes in a washing machine or drier are kept moving in circles by the sides of their container pushing inwards.
- A car can only move around a corner because of the force of friction between the road and the tyre (Figure 2.116).
- A person walking around a corner needs the force of friction between their feet and the ground to change their direction.
- A passenger in a car can move in a circular path because of the force between them and the seat. If the car moves faster, the side of the car may also need to push inwards on them, although the passenger will feel that they are pushing outwards.
- An aircraft can change direction because of the force provided on the wings from the air when the plane tilts.
- A train on a track can go around bends because of the normal reaction force of the track on the wheel (Figure 2.117).

Figure 2.116

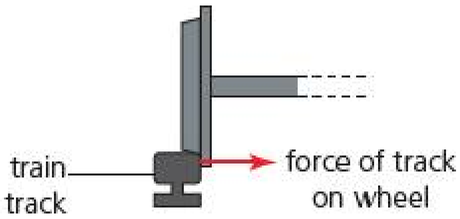


Figure 2.117 The train track pushes inwards on the wheel of a train moving in a circular path

Calculating centripetal accelerations

2.4.2 Apply the expression for centripetal acceleration.

When an object moving in a circle has a greater speed v or smaller radius r, the centripetal acceleration (and force) needs to be larger. This is represented in the equation for calculating the magnitude of the centripetal acceleration:

$$a = \frac{v^2}{r}$$

This equation is listed in the IB *Physics data booklet*.

Since the speed in a circle equals the circumference divided by the time taken to complete one circle, a period T, speed $v = 2\pi r/T$. The equation for centripetal acceleration can then be rewritten as:

$$a = \frac{v^2}{r} = \frac{(2\pi r/T)^2}{r}$$

$$a = \frac{4\pi^2 r}{T^2}$$

This equation is also included in the IB *Physics data booklet*.

Since $F = ma$, the expression for the centripetal force needed to make a mass m move in a circle of radius r at a constant speed v is:

$$F = \frac{mv^2}{r}$$

In Figure 2.118, the bigger child needs a much greater centripetal force because she is more massive and travelling with a greater speed.

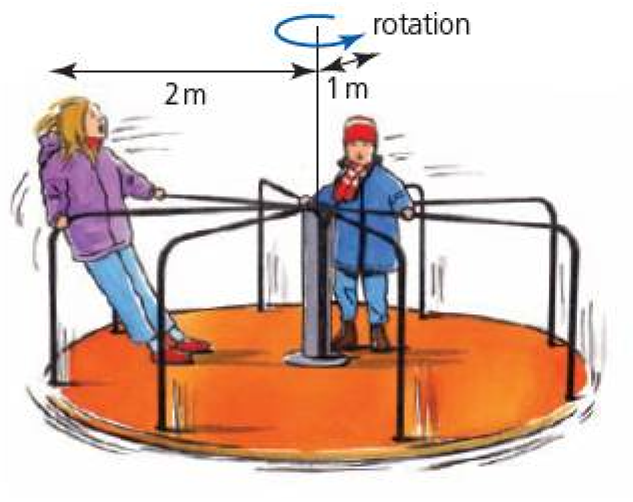


Figure 2.118 Children on a playground ride

Additional Perspectives

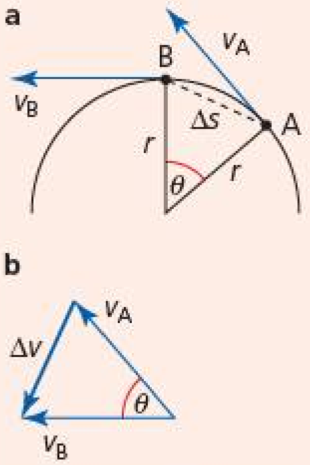


Figure 2.119 Deriving an equation for centripetal acceleration

Deriving the expression $a = \frac{v^2}{r}$

Consider a mass moving in a circular path of radius r, as shown in Figure 2.119a. It moves a through an angle θ, and a distance Δs along the circumference as it moves from A to B, while its velocity changes from v_A to v_B.

To calculate acceleration we need to know the change of velocity, Δv. This is done using the vector diagram shown in Figure 2.119b. Note that the direction of the change of velocity (and therefore the acceleration) is towards the centre of motion.

The two triangles are similar and, if the angle is small enough that Δs can be approximated to a straight line, we can write:

$$\theta = \frac{\Delta v}{v} = \frac{\Delta s}{r}$$

(The magnitudes of v_A and v_B are equal and represented by the speed v.)

Dividing both sides of the equation by Δt we get:

$$\frac{\Delta v}{\Delta t \times v} = \frac{\Delta s}{\Delta t \times r}$$

Then, since $a = \Delta v/\Delta t$ and $\Delta s/\Delta t = v$:

$$a = \frac{v^2}{r}$$

Question

1 The orbit of the Earth around the Sun is approximately circular, with an average radius of about 150 million kilometres. Calculate the centripetal acceleration of the Earth towards the Sun.

Worked examples

31 Consider a ball of mass 72 g whirled with a constant speed of $3.4\,\mathrm{m\,s^{-1}}$ around in a (nearly) horizontal circle of radius 65 cm on the end of a thin piece of string, as shown in Figure 2.120.

a Calculate the centripetal acceleration and force.

b Explain why the force provided by the string cannot act horizontally.

c Explain a probable reason why the string breaks when the speed is increased to $5.0\,\mathrm{m\,s^{-1}}$.

d In which direction does the ball move immediately after the string breaks?

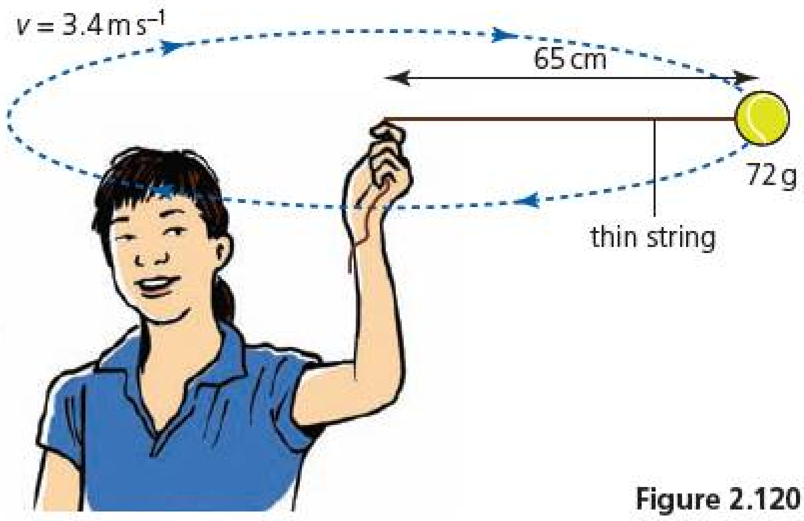


Figure 2.120

a $a = \frac{v^2}{r}$

$a = \frac{3.4^2}{0.65}$

$a = 18\,\text{m s}^{-2}$

$F = ma$

$F = 0.072 \times 18.0 = 1.3\,\text{N}$

b If the force is horizontal, it cannot have a vertical component with which to support the weight of the ball (see Figure 2.121).

c As the speed of the ball is increased, a greater centripetal force is needed for the same radius. If this force is greater than can be provided by the string, the string will break.

d The ball will continue its instantaneous velocity in a straight line after the string breaks. It will move at a tangent to the circle, but gravity will also affect its motion.

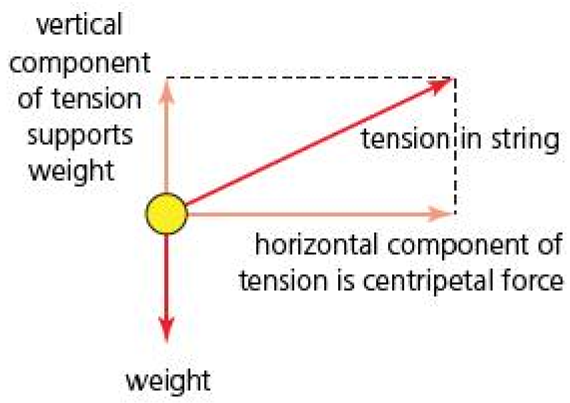


Figure 2.121 Free-body diagram for a ball whirled in a circle

32 Explain how it is possible in a fairground ride for the passengers to be upside-down without falling out of the carriage (Figure 2.122).

The passengers do not fall out because their speed is sufficiently fast that the centripetal force required to keep them moving around the curved path is greater than their weight.

At the top of the circle the downwards forces acting on a passenger are their weight (constant) and the reaction force from the seat (variable) – *if* they are in contact with it. The resultant force is found by adding these two forces. If the centripetal force required is greater than their weight they will stay in contact with the seat and the extra force needed is provided by the seat. If their weight is greater than the centripetal force needed (because the speed is too slow), then they will fall out of the carriage.

Figure 2.122 Upside-down on a fairground ride

2.4.4 Solve problems involving circular motion.

116 A toy train of mass 432 g is moving around a circular track of radius 67 cm at a steady speed of $25\,\text{cm s}^{-1}$.

a Calculate the centripetal acceleration of the train.

b What is the resultant force acting on the train (in magnitude and direction)?

c What is the time taken for the toy train to go around the track once?

117 The hammer being thrown in Figure 2.123 completed two full circles of radius 1.60 m at constant speed in 0.824 s just before it was released.

a What was its centripetal acceleration?

b What force did the thrower need to exert on the hammer if its mass was 4.0 kg?

Figure 2.123

118 The Moon's distance from Earth varies, but averages about 380 000 km. The Moon orbits the Earth in an approximately circular path every 27.3 days.
a Calculate the Moon's orbital speed in $m\,s^{-1}$.
b What is the centripetal acceleration of the Moon towards the Earth?

119 **a** Calculate the centripetal force needed to keep a 1200 kg car moving at $15\,m\,s^{-1}$ around a curved road of radius 60 m.
b What provides this force?
c Explain why it might be dangerous for the car to try to travel at twice this speed around the same bend.
d If the road is wet or icy, why should the driver go even slower?
e Discuss the possible effects on safety if the car was carrying passengers and luggage that increased the mass by 500 kg.

120 A bucket of water can be whirled in a vertical circle over your head without the water coming out if the centripetal force required is greater than the weight.
a Calculate the minimum speed of the bucket of water if the radius of the circle is 90 cm.
b How many revolutions is that every second?

121 **a** A boy of mass 65 kg standing on the equator is spinning with the Earth at a speed of approximately $460\,m\,s^{-1}$. What resultant centripetal force is required to keep him moving in a circle? (The Earth's radius is 6.4×10^6 m.)
b What is the weight of the boy?
c Draw a fully labelled free-body diagram of the boy.

122 Figure 2.124 shows a pendulum of mass 120 g being swung in a horizontal circle.
a Draw a free-body diagram of the mass, m.
b Calculate the centripetal force acting on the mass.
c If the radius of the circle is 28.5 cm, what is the speed of the pendulum and how long does it take to complete one circle?

123 Explain why a satellite orbiting the Earth in a circular path does not need an engine to keep it moving.

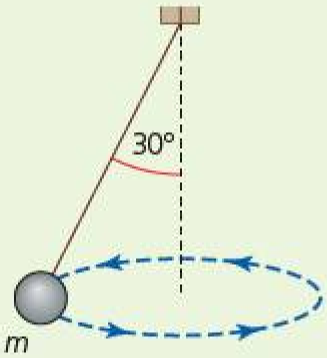


Figure 2.124

SUMMARY OF KNOWLEDGE

2.1 Kinematics

- Displacement, s, is defined as the distance in a given direction from a fixed reference point. Speed, v, is defined as rate of change of distance with time. $v = \Delta s/\Delta t$ (unit: $m\,s^{-1}$). Velocity, v, is defined as rate of change of displacement with time (speed in a given direction).
- Speed is a scalar quantity; velocity is a vector quantity.
- Any object that is changing the way it is moving (changing its velocity) is accelerating. This includes going faster, going slower and changing direction. Acceleration, a, is defined as the rate of change of velocity with change in time. $a = \Delta v/\Delta t$ (unit: $m\,s^{-2}$).
- When we calculate a velocity from change of displacement/time, we are determining the average velocity during that time. However, in physics we are usually more concerned with the precise values of quantities such as velocity and acceleration at an exact moment. These are called instantaneous values (rather than average values) and, for changing velocities, they can be calculated from measurements made over very short intervals of time, or from the gradients of tangents to displacement–time graphs.
- One common kind of motion is that of objects moving under the effects of gravity. If air resistance is negligible, all masses near the Earth's surface accelerate downwards at $9.81\,m\,s^{-2}$. This acceleration is given the symbol g.
- The simplest kind of motion to study is that of objects moving with uniform acceleration. The symbol u is used for the velocity at the start of observation, and the symbol v for the velocity after time t. Problems can be solved using the equations of motion for uniformly accelerated motion:

$$v = u + at \qquad s = \frac{u + v}{2}t \qquad v^2 = u^2 + 2as \qquad s = ut + \tfrac{1}{2}at^2$$

- Drawing and interpreting distance–time, velocity–time and acceleration–time graphs are important skills. The gradient of a displacement–time graph equals velocity. The gradient of a velocity–time graph equals acceleration and the area under the graph represents change of displacement. The area under an acceleration–time graph represents change of velocity.
- The motion of objects through the air is opposed by the force of air resistance. The amount of air resistance depends on the speed of the object, its cross-sectional area and its shape. As an object's speed increases, so does the air resistance, so that the resultant force and acceleration decrease. When air resistance becomes equal and opposite to the weight, a falling object will reach a constant (terminal) speed.
- All motion is relative to something else. We usually assume that an 'observer' of movement is at rest in a stationary frame of reference.
- Relative velocity is the difference between two velocity vectors. The velocity of A relative to B, $\boldsymbol{v}_{AB} = \boldsymbol{v}_A - \boldsymbol{v}_B$. If the two velocities are not acting along the same line, then $\boldsymbol{v}_{AB}$ can be determined by graphically adding $\boldsymbol{v}_A$ to the negative of $\boldsymbol{v}_B$, using a parallelogram.

2.2 Forces and dynamics

- Forces can change the shapes or velocities of objects.
- Force is a vector quantity and can be represented by an arrow of length proportional to the magnitude of the force.
- Common types of contact force include reaction forces, tension, compression, friction, upthrust, air resistance and drag.
- Non-contact forces are an important part of the study of physics: gravitational, electrical, magnetic and nuclear forces are all studied in this course.
- The force of gravity pulling a mass towards a planet is called weight. Weight can be calculated from mg, where g is the gravitational field strength, with units N kg^{-1} (g is also called the acceleration due to gravity).
- Free-body diagrams show all the forces acting on an object, but without showing the surroundings.
- The resultant of two or more forces can be found using vector addition. This may involve scale drawing using a parallelogram. If three forces are in equilibrium, the resultant of any two forces is equal and opposite to the third force.
- A single force, F, can be resolved into two components at right angles to each other: $F\cos\theta$ and $F\sin\theta$. The effects of these two components can then be considered independently.
- Newton's first law of motion states that an object will remain at rest, or continue to move in a straight line at a constant speed, unless a resultant force acts on it. An object that has no resultant force acting on it is said to be in translational equilibrium.
- Newton's second law can be expressed in two ways: $F = ma$ or $F = \frac{\Delta p}{\Delta t}$.
- When objects collide with each other, the forces involved can be reduced by increasing the length of time of the impact.
- The linear momentum of a moving mass is defined as its mass × velocity (symbol p and unit kg m s^{-1}): $p = mv$. Momentum is a vector quantity.
- The law of conservation of linear momentum states that the total (linear) momentum of a system is constant, provided that there are no external forces acting on it. There are no exceptions to this law – it is always true because of Newton's third law (see below).
- For masses A and B interacting with each other: $m_A u_A + m_B u_B = m_A v_A + m_B v_B$
- If a resultant force acts for a longer time it has a greater effect in changing the motion. Impulse is defined as $F\Delta t$. From Newton's second law, impulse = change of momentum, Δp ($= m\Delta v$, for a constant mass).
- The magnitude of an impulse can be determined from the area under a force–time graph.

- Newton's third law of motion states that whenever one body exerts a force on another body, the second body exerts exactly same force on the first body, but in the opposite direction. All forces occur in pairs.
- Newton's third law (or the law of the conservation of momentum) can be very useful when explaining the propulsion of vehicles.

2.3 Work, energy and power

- Work, W, is the name that we give to the transfer of energy when a force moves (its point of application). $W = Fs$, assuming that the force is constant and in the same direction as the movement. The unit of work is the joule. 1 J of work is done when a force of 1 N moves 1 m.
- If the movement and the force are at an angle θ to each other, the work done, $W = Fs\cos\theta$. If an object is moving perpendicularly to a force, the force is not doing any work on the object.
- If the force is not constant, the work done can be determined from the area under the force–displacement graph.
- To make anything happen, energy has to be transferred. Energy is measured in joules.
- Energy exists in different forms. Stored energy can be very useful and is called potential energy. Gravitational, elastic strain, chemical, electric and nuclear are all examples of potential energies.
- The energy of motion, kinetic energy, is another important form of energy. It can be calculated from $E_K = \frac{1}{2}mv^2$ $(= p^2/2m)$.
- Mechanical waves, including sound, are a combination of kinetic energy and potential energy.
- Thermal energy flows from place to place because of a difference in temperature, and electric currents transfer energy (because of a difference in potential). Electromagnetic waves, like light, also transfer energy from place to place.
- The energy of the molecules inside everything is called internal energy.
- The principle of conservation of energy states that energy cannot be created or destroyed, it can only be transferred.
- The main energy transformations occurring in a wide variety of examples can be identified. Useful energy is dissipated into the surroundings in all macroscopic processes, resulting in an increase in internal energy, and the spread of thermal energy.
- When a mass is raised away from the Earth, work has to be done as the mass gains gravitational potential energy: $\Delta E_P = mg\Delta h$ (weight × height).
- For masses moving from rest up or down freely under the effects of gravity (without air resistance), the change in gravitational potential energy equals the change in kinetic energy: $mg\Delta h = \frac{1}{2}mv^2$.
- A collision in which the total kinetic energy of the masses was the same before the collision as afterwards is known as an elastic collision. Perfectly elastic collisions are not possible in the macroscopic world. In inelastic collisions some or all of the kinetic energy is transferred to other forms of energy.
- When moving objects are accelerated (or decelerated) by a constant force, the work done equals the change in kinetic energy: $Fs = \frac{1}{2}mv^2 - \frac{1}{2}mu^2$. If the force varies, an average value must be used in the calculation, or the area under a force–distance graph considered.
- Power is the rate of transferring energy. In the example of doing work, $P = \Delta W/\Delta t$. The unit of power is the watt, W ($1\,W = 1\,J\,s^{-1}$). The power needed to keep an object moving at constant speed is given by $P = Fv$, where F is the force opposing motion.
- When energy is transferred, the energy output that is useful to us is always less than the total energy input, because some energy will have been dissipated to the surroundings. The efficiency of an energy transfer is defined as:

$$\text{efficiency} = \frac{\text{useful energy (or power) output}}{\text{total energy (or power) input}}$$

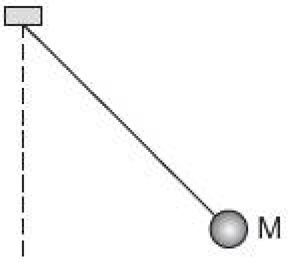
2.4 Uniform circular motion

- Any object which is moving in a circular path has a continuously changing velocity, even if it has a constant speed (because velocity is a vector quantity). This means that there is an acceleration, which is directed towards the centre of the circle, even though the object keeps moving with the same radius from the centre. This is called a centripetal acceleration.
- Centripetal acceleration can be calculated from $a = v^2/r$. Since $v = 2\pi r/T$ (where T is the time period taken to travel in one circle), this can also be written as: $a = 4\pi^2 r/T^2$.
- In order to have a centripetal acceleration and move in a circular path, an object must have a resultant force acting on it, which also acts towards the centre of the circle. This is called a centripetal force. $F = ma = mv^2/r$.
- A force which acts continuously at right angles to a mass moving at constant speed produces circular motion. The origin of the centripetal forces making a variety of objects move along circular paths should be identified.
- Vector diagrams can be drawn showing the instantaneous velocity, acceleration and force on an object moving in a circle at a constant speed.

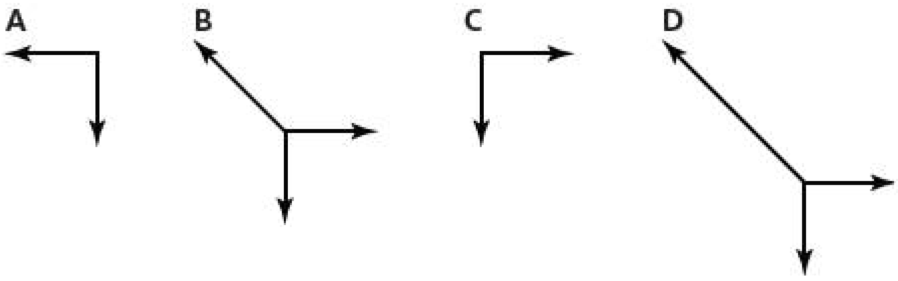
Examination questions – a selection

Paper 1 IB questions and IB style questions

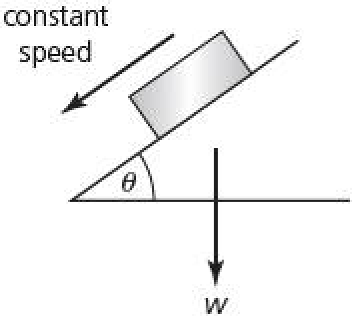
Q1 M is a small mass on the end of a light-weight string. It has been pulled to one side with a force that keeps it stationary.



Which of the following four diagrams is the correct free-body representation of the forces acting on the mass?



Q2 An object of weight W is slipping down a slope (inclined plane) which makes an angle of θ with the horizontal. If the object is moving with constant speed, what is the magnitude of the frictional force up the slope?



A W **B** $W \sin\theta$ **C** $W \cos\theta$ **D** $W/2$

Q3 The work done when a constant force acts on a mass is always equal to

A the magnitude of the force multiplied by the distance moved by the mass.
B the magnitude of the force multiplied by the displacement perpendicular to the force.
C the magnitude of the force multiplied by the displacement in the direction of the force.
D the vector sum of the force and the distance moved by the mass.

Q4 A rocket is travelling across space when its engine ejects gases of total mass m in a time t. The speed of the gases relative to the rocket is v.
Which of the following is the correct expression for the force exerted by the gases on the rocket?

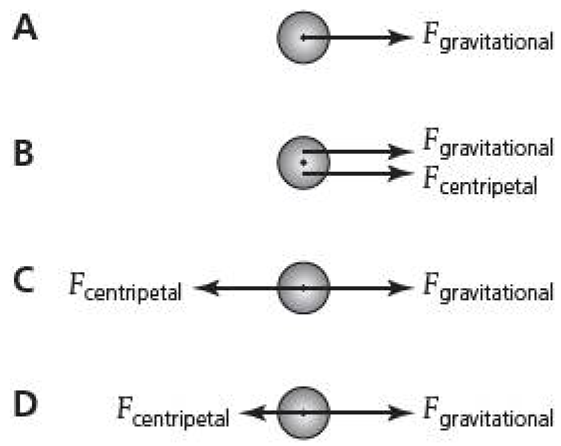
A mv **B** $\frac{mv}{2}$ **C** mvt **D** $\frac{mv}{t}$

Q5 Which of the following is a correct definition of the instantaneous velocity of a moving object at time t?

A $\frac{\text{distance moved}}{\text{time taken}}$
B $\frac{\text{displacement}}{\text{time taken}}$
C rate of change of displacement at time t
D rate of change of distance at time t

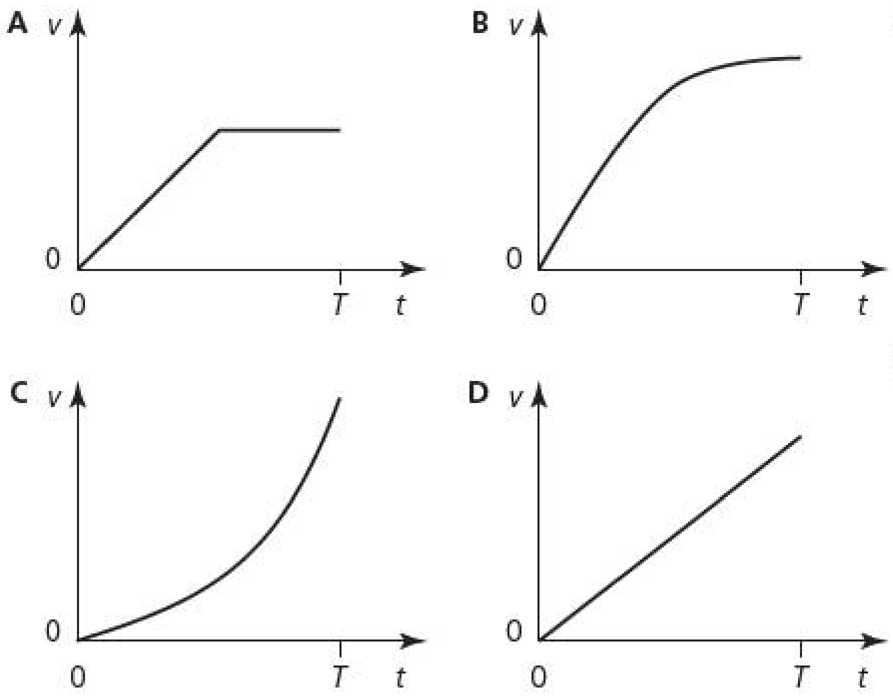
Q6 The Moon orbits the Earth.

Which of the following diagrams correctly represents the force(s) acting on the Moon?



Higher Level Paper 1, May 09 TZ2, Q5

Q7 A big object is dropped from a large height. It hits the ground at time T after being dropped. Which of the following graphs best represents how the speed, v, of the object varies with time, t, until just before it hits the ground?



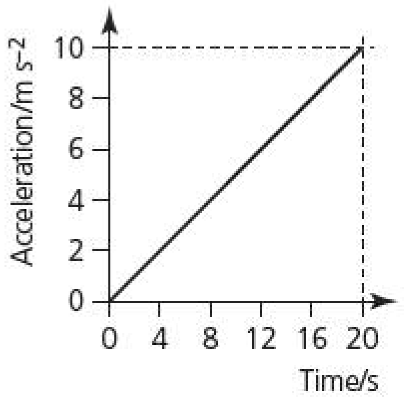
Q8 When the motion of two cars was compared, it was found that car A was more powerful than car B. Which of the following statements must be true?

A Car A produces more useful energy than car B.
B Car A produces a greater force than car B.
C In the same time, car A does more useful work than car B.
D In the same time, car A moves a greater distance than car B.

Q9 A mass of 5 kg is pulled up a slope at a constant speed of 2 m s^{-1}. After rising a vertical height of 4 m, the total work done was 1200 J. The work done in overcoming friction was

A 1000 J
B 200 J
C 2400 J
D 1400 J

Q10 An increasing force acts on an object and its acceleration increases as shown in the graph.



If the object was initially at rest, what is the speed of the object after 20 seconds?

A 0.5 m s^{-1}
B 2.0 m s^{-1}
C 100 m s^{-1}
D 200 m s^{-1}

Q11 If there is no resultant force acting on an object, which of the following quantities must also be zero?

A speed
B velocity
C acceleration
D momentum

Q12 An electric motor raises a 2.5 kg mass a distance of 12 m in a time of 6 s. If the efficiency of the process was 20%, what was the input power to the motor?

A 10 W
B 25 W
C 250 W
D 500 W

Q13 A mass is moving at a constant speed with a kinetic energy E_K. What is the kinetic energy of another mass which has twice the mass and half the speed of the first mass?

A $E_K/2$
B E_K
C $2E_K$
D $4E_K$

Q14 A vehicle is driven up a hill at constant speed. Which of the following best describes the energy changes involved?

A Chemical energy is converted into gravitational potential energy.
B Chemical energy is converted into gravitational potential energy, sound and thermal energy.
C Gravitational potential energy is converted into chemical energy.
D Gravitational potential energy is converted into chemical energy, sound and thermal energy.

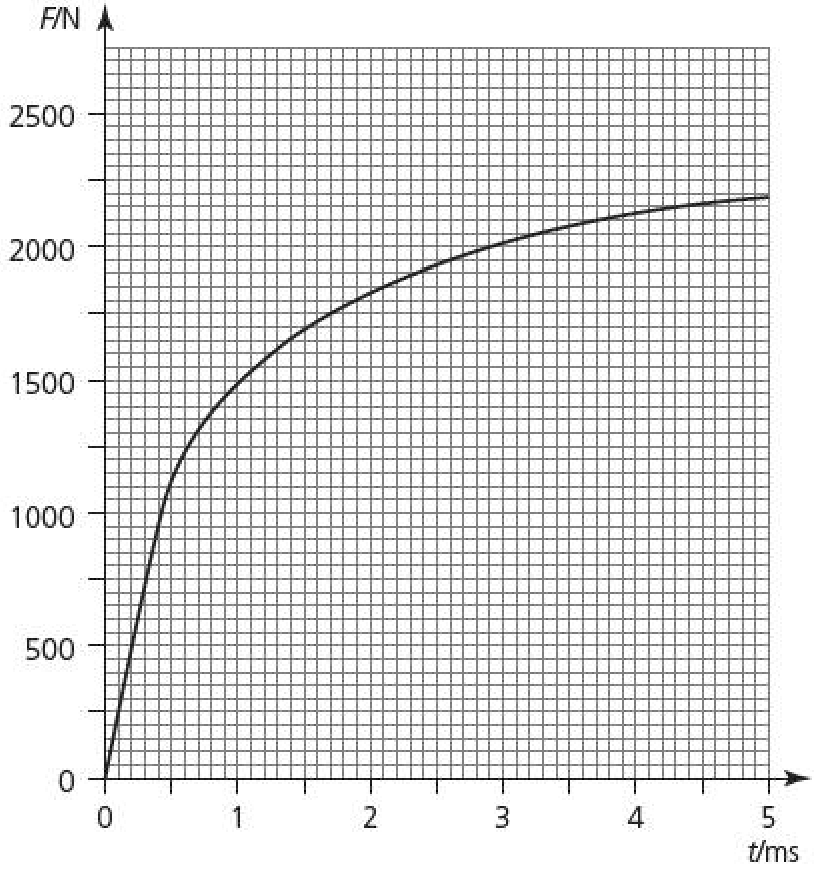
Standard Level Paper 1, Nov 09, Q7

Q15 A weight W is suspended from the ceiling on the end of a length of string. According to Newton's third law of motion there must be another force equal and opposite to the weight. This second force is the

A the downwards force of the string on the ceiling.
B the upwards force of the string on the weight.
C the upwards force exerted by the weight on the Earth.
D the tension in the string.

Paper 2 IB questions and IB style questions

Q1 A bullet of mass 32 g is fired from a gun. The graph shows the variation of the force F on the bullet with time t as it travels along the barrel of the gun.



The bullet is fired at time $t = 0$ and the length of the barrel is 0.70 m.

a State and explain why it is inappropriate to use the equation $s = ut + \frac{1}{2}at^2$ to calculate the acceleration of the bullet. [2]

b Use the graph to

i determine the average acceleration of the bullet during the final 2.0 ms of the graph. [2]

ii show that the change in momentum of the bullet, as the bullet travels along the length of the barrel, is approximately 9 N s. [3]

c Use the answer in **b ii** to calculate the

i speed of the bullet as it leaves the barrel [2]

ii average power delivered to the bullet. [3]

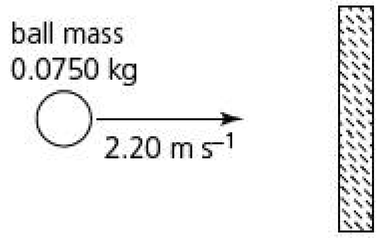
d Use Newton's third law to explain why a gun will recoil when a bullet is fired. [3]

Standard Level Paper 2, May 09 TZ2, QB1 (Part 1)

Q2 This question is about impulse.

a A net force of magnitude F acts on a body. Define the *impulse* I of the force. [1]

b A ball of mass 0.0750 kg is travelling horizontally with a speed of 2.20 m s^{-1}. It strikes a vertical wall and rebounds horizontally.

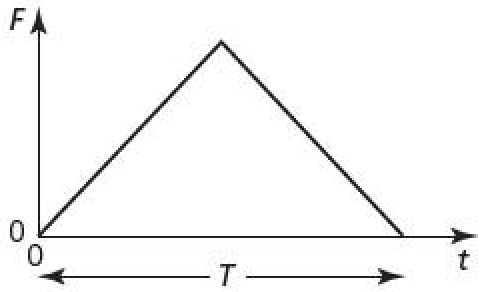


Due to a collision with the wall, 20% of the ball's initial kinetic energy is dissipated.

i Show that the ball rebounds from the wall with a speed of 1.97 m s^{-1}. [2]

ii Show that the impulse given to the ball by the wall is 0.313 N s. [2]

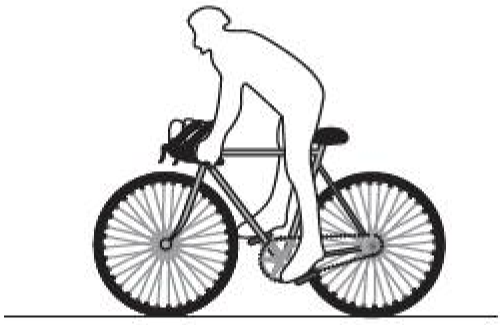
c The ball strikes the wall at time $t = 0$ and leaves the wall at time $t = T$.
The sketch graph shows how the force F that the wall exerts on the ball is assumed to vary with time t.



The time T is measured electronically to equal 0.0894 s.
Use the impulse given in **b ii** to estimate the average value of F. [4]

Standard Level Paper 2, May 09 TZ1, QA2

Q3 a A system consists of a bicycle and cyclist travelling at a constant velocity along a horizontal road.



i State the value of the net force acting on the cyclist. [1]

ii On a copy of the diagram, draw labelled arrows to represent the vertical forces acting on the bicycle. [2]

iii With reference to the horizontal forces acting on the system, explain why the system is travelling at constant velocity. [2]

b The total resistive force acting on the system is 40 N and its speed is $8.0\,m\,s^{-1}$. Calculate the useful power output of the cyclist. [1]

c The cyclist stops pedalling and the system comes to a rest. The total mass of the system is 70 kg.

i Calculate the magnitude of the initial acceleration of the system. [2]

ii Estimate the distance taken by the system to come to rest from the time the cyclist stops pedalling. [2]

iii State and explain **one** reason why your answer to **c ii** is only an estimate. [2]

Standard Level Paper 2, Nov 09, QB1 (Part 2)

3 Thermal physics

STARTING POINTS

- When trying to understand the everyday world around us, scientists often look for explanations by considering what is happening to the tiny particles that are found in everything. In this way, small-scale (microscopic) explanations are developed into the theories that help explain our large-scale (macroscopic) world.
- Most of the macroscopic physical properties of substances can be explained through an understanding of the forces between molecules, how molecules are arranged and how molecules move.
- Because there are electrical forces between the molecules in solids and liquids, the substances contain electrical potential energy.
- Temperature is often measured on the Celsius scale (°C). The Celsius scale has two fixed points: the melting point and the boiling point of pure water, which are chosen to be 0°C and 100°C.
- All moving masses (including molecules) have kinetic energy. The faster they move, the greater their kinetic energy: $KE = \frac{1}{2}mv^2$.
- Work is the energy transferred when a force is used to move something.
- The total energy of any closed system is constant.
- Power is the rate of transfer of energy: $\text{power} = \frac{\text{energy transferred}}{\text{time taken}}$
- Density is mass per unit volume: $\text{density} = \frac{\text{mass}}{\text{volume}}$

3.1 Thermal concepts

Thermal energy and temperature differences

3.1.1 State that temperature determines the direction of thermal energy transfer between two objects.

The transfer of energy from hotter to colder as a result of a temperature difference is called **thermal energy (heat)**.

Objects continuously exchange energy with their surroundings. An object which is hotter than its surroundings will give out (emit) more energy than it takes in, and a colder object will take in (absorb) more energy than it gives out.

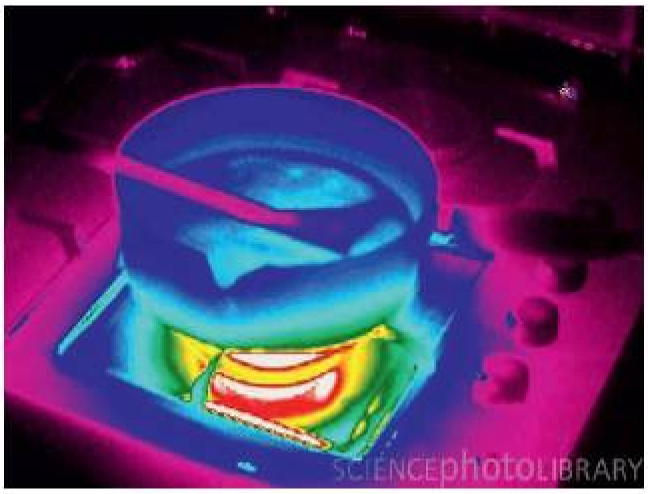


Figure 3.1 This thermogram, taken using infrared radiation, uses colour to show different temperatures in a saucepan on a cooker. The scale runs from white (hottest), through red, yellow, green and blue to pink (coldest)

Figure 3.1 shows the temperature differences in a saucepan and cooker which will result in a flow of thermal energy. Thermal energy is *always* transferred from higher temperature to lower temperature.

Temperature determines the direction of thermal energy transfer between two objects.

Consider the simple example of two objects (or substances) at different temperatures able to transfer thermal energy between themselves, but isolated from everything around them (their surroundings). The hotter object will transfer energy to the colder object and cool down, while at the same time the colder object warms up. As the temperature difference between the two

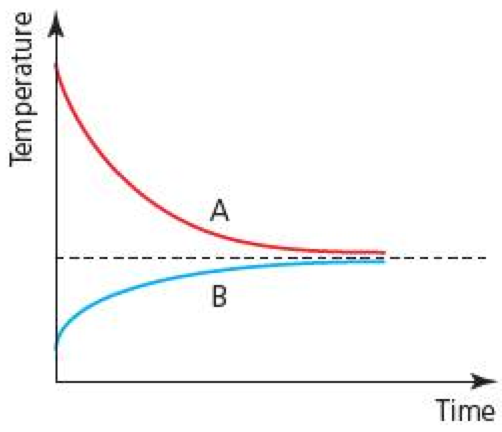


Figure 3.2 Two objects at different temperatures, insulated from their surroundings but not from each other, will reach thermal equilibrium

objects gets smaller, so too does the *rate* of thermal energy transfer. This is represented in Figure 3.2, which shows how the temperature of two objects (A and B) may change when they are placed in thermal contact with each other. (Being in 'thermal contact' means that thermal energy can be transferred between them, by *any* means.) Eventually they will reach the same temperature.

When the temperatures have stopped changing and both objects are at the same temperature, the objects are said to be in **thermal equilibrium** and there will be no net flow of thermal energy between them.

In any realistic situation, it is not possible to completely isolate/**insulate** two objects from their surroundings, so the concept of thermal equilibrium may seem to be idealized. The concept of hotter objects getting colder, and colder objects getting hotter suggests that eventually *everything* will end up at the same temperature.

Temperature scales

3.1.2 State the relationship between the Kelvin and Celsius scales of temperature.

The temperature scales that we use today were designed for simplicity and easy reproduction. On the **Celsius scale (°C)**, sometimes called the centigrade scale, 0 °C is the temperature at which pure water forms ice (at normal atmospheric pressure), and 100 °C is the temperature at which pure water boils (at normal atmospheric pressure). It is important to realize that this temperature scale was devised for convenience, that is, these values were *chosen* – they were not discovered. In particular, 0 °C is definitely *not* a zero of temperature, nor a zero of energy. It has no other significance other than being the melting point of ice. (For example, 10 °C cannot be considered to be 'twice as hot' as 5 °C.)

After it was predicted that almost all molecular motion stops (in all substances) at –273 °C, it made sense to make this the true zero of temperature. This temperature is commonly called **absolute zero**.

The **Kelvin (absolute) temperature scale** is an adaptation of the Celsius scale with its zero at –273 °C. (A more precise value is –273.15 °C.) On this scale the units are **kelvins, K** (not °K). *Changes* in temperature of 1 °C and 1 K are identical, which makes conversion from one scale to the other very straightforward:

$$T/\mathrm{K} = t/^\circ\mathrm{C} + 273$$

Table 3.1 A comparison of temperatures in degrees Celsius and in kelvin

Temperature	°C	K
absolute zero	–273	0
melting point of water	0	273
body temperature	37	310
boiling point of water	100	373

In the equation above, note that use of the symbol T for temperature implies the Kelvin scale and the symbol t implies the Celsius scale. θ is also widely used for temperature in degrees Celsius. Table 3.1 compares some important temperatures on the two scales.

When making calculations involving temperature *changes*, either degrees Celsius or kelvins may be used, but it is important to remember that when dealing with calculations involving just *one* temperature, kelvins must be used. There is more about temperature scales in Chapter 10.

Worked example

1 The normal freezing point of mercury is –39 °C. What is this temperature in kelvin?

$T/\mathrm{K} = t/^\circ\mathrm{C} + 273$
$T = -39 + 273$
$T = 234\,\mathrm{K}$

1 a The world's highest and lowest recorded temperatures are reported to be 58 °C (in Libya) and –89 °C (in Antarctica). What are these temperatures on the Kelvin scale?
b The temperature of some water is raised by 38 °C. What is the temperature rise in kelvins?

2 a The volume of a gas is 37 cm^3 when its temperature is 23 °C. If the volume is proportional to the absolute temperature, at what temperature (°C) will the volume be 50 cm^3?
b What will the volume be when the gas is at –15 °C?

Internal energy

3.1.3 State that the internal energy of a substance is the total potential energy and random kinetic energy of the molecules of the substance.

All substances contain moving particles. In the context of this chapter, the word 'particle' is a general term that might apply to a molecule, an atom or an ion. Although different substances may contain any or all of these, most substances are *molecular* and in the rest of this chapter the term *molecule* will be used to describe the particles in any substance.

Moving molecules have **kinetic energy**. The molecules may be moving in different ways, which gives rise to three different forms of random molecular kinetic energy:

- Molecules may be vibrating about fixed positions (as in a solid) – this gives the molecules **vibrational kinetic energy**.
- Molecules may be moving from place to place (translational motion) – this gives the molecules **translational kinetic energy**.
- Molecules may also be spinning (rotating) – this gives the molecules **rotational kinetic energy**.

Molecules may have **potential energy** as well as kinetic energy. In solids and liquids, it is the **electrical forces** (between charged particles) that keep the molecules from moving apart or moving closer together. Wherever there are electrical forces there will be electrical potential energy in a system, in much the same way as gravitational potential energy is associated with gravitational force. (There is more about electrical forces and electrical potential energy in Chapters 6 and 9.)

In gases, however, the forces between molecules are usually negligible because of the greater separation between molecules. This is why gas molecules can move freely and randomly. The molecules in a gas, therefore, usually have negligible electrical potential energy. All the energy is in the form of kinetic energy.

So, to describe the total energy of the molecules in a substance, we need to take account of both the kinetic energies and the potential energies. This is called the **internal energy** of the substance and is defined as follows:

The internal energy of a substance is the total random kinetic energy and potential energy of all the molecules inside it.

It is important not to call this energy 'heat'. That is, we should not refer to the thermal energy (or heat) *in* anything.

In the definition above, the word 'random' means that the molecular movements are **disordered** and unpredictable. That is, they are not linked to each other, or **ordered** – as their motions would be if they were all moving together, such as the molecules in a macroscopic motion of a car, for example. The molecules in a moving car have both the ordered kinetic energy of macroscopic movement and the random kinetic energy of internal energy.

Summarizing the differences between temperature, internal energy and thermal energy

3.1.4 Explain and distinguish between the macroscopic concepts of temperature, internal energy and thermal energy (heat).

Temperature, internal energy and thermal energy (often called heat) are widely used and very important concepts throughout all of science, but they are commonly misunderstood and misused terms. To stress their importance, the meaning of these concepts is summarized below.

- *Internal energy* is the total energy (random kinetic and potential) of the molecules inside a substance.

- If energy is transferred to a substance its molecules move faster. We say that it has become *hotter* and this is measured as an *increased temperature*. A more precise meaning of temperature is given later in this chapter.
- *Thermal energy* (heat) is energy flowing from a higher temperature to a lower temperature. In any particular example, the object or substance we are considering is often called the *system* and thermal energy flows between the system and the *surroundings*. You may know that thermal energy is transferred by **conduction**, **convection** and **radiation**, but detailed knowledge of these processes is not needed in this chapter.

In Figure 3.3 thermal energy is being conducted into one hand and out of the other.

Figure 3.3 One hand is receiving thermal energy and the other hand is losing it

3 Equal masses of two different gases at 20 °C are both heated to 40 °C.
  a Discuss whether they have the same amount of internal energy to begin with.
  b Discuss whether the same amount of thermal energy was transferred to both of them.

4 Figure 3.4 sketches how the resultant force between two molecules varies with separation. x_0 represents the average equilibrium separation between molecules in a solid. The molecules in a gas are typically ten times further apart than in a solid.

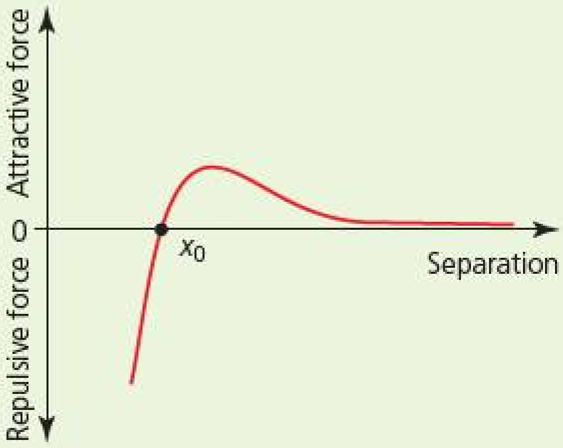


Figure 3.4 How intermolecular force varies with molecular separation

  a Describe how the resultant force between molecules changes if they move
    i slightly further apart
    ii slightly closer together.
  b What can you conclude from this graph about how the size of forces between molecules in a gas is different from forces in a solid?
  c Explain why you might expect the density of gases to be about 1000 times smaller than the density of solids.

5 The 'sparkles' emitted from the sparkler (Figure 3.5) are very hot. Explain why they do not usually cause any harm when they land on people or their clothing. Use the terms 'temperature', 'internal energy' and 'thermal energy' in your explanation. (The hot sparkler itself will cause burns if touched.)

Figure 3.5 Sparklers are usually not as dangerous as they may look!

Heating and working

Apart from supplying thermal energy to a system (heating), there is another, fundamentally different, way to make something hotter: we can do *mechanical work* on it. A simple example would be the force of friction causing a temperature rise when surfaces rub together. Heating is a *non-mechanical* transfer of energy.

Figures 3.6 and 3.7 provide examples of these different ways of raising the temperature of an object. In Figure 3.6, the internal energy of the screw will rise as it gains thermal energy from the hand holding it. In Figure 3.7, its internal energy will rise because a force is twisting the screw into a piece of wood against resistive forces opposing it, doing mechanical work.

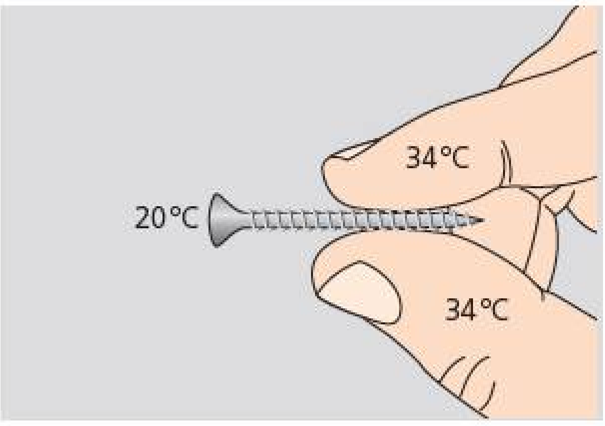


Figure 3.6 Getting hotter by being in contact with something at a higher temperature

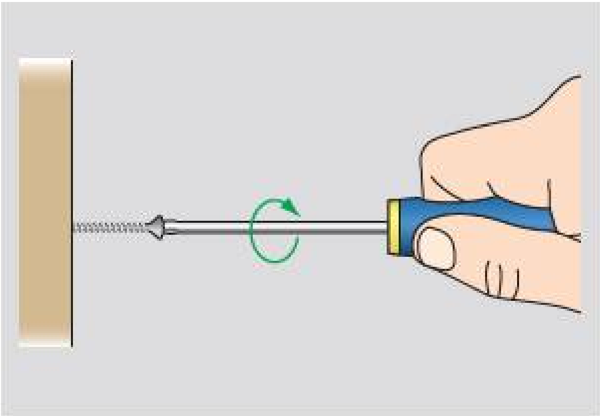


Figure 3.7 Getting hotter because work is being done

Additional Perspectives

James Prescott Joule

The SI unit of energy is named after the Englishman, James Prescott Joule (Figure 3.8), who was a 19th century physicist and manager of a brewery. His experiences as a brewer may have contributed to his renowned skills in accurate measurement, especially those involving small temperature changes. In the middle of the 19th century, 'heat' was believed to be an undetectable 'calorific fluid' that flowed from hotter to colder things, but Joule tried repeatedly to show that 'heat' was just another form of energy that could be transformed from an equal amount of other forms, such as kinetic energy or gravitational potential energy.

Figure 3.8 James Prescott Joule

In particular, he is remembered for his 'mechanical equivalent of heat' experiments in which mechanical energy was used to raise the temperature of water. It is even claimed that he spent part of his honeymoon trying to measure a very small temperature difference between the top and bottom of waterfalls, which is not an easy thing to do!

His work united the separate topics of energy and 'heat', and was important in the later development of the law of conservation of energy and the first law of thermodynamics. Joule also worked with Lord Kelvin on thermometry and temperature scales.

Question

1 The use of the word 'heat' can be confusing, especially if it is used to represent both energy inside objects as well as energy being transferred. Discuss the use of the word 'heat' in this Additional Perspective.

Counting molecules

3.1.5 **Define** the mole and molar mass.
3.1.6 **Define** the Avogadro constant.

If we need to make calculations about mass, speed and kinetic energy of molecules, we need to understand the link between the macroscopic measurement of mass and the microscopic numbers of molecules.

The *amount* of a substance is a measure of the number of particles it contains and it is measured in moles.

One **mole** (mol) of a substance is the amount that contains as many (of its defining) particles as there are atoms in exactly 12 g of carbon-12.

The number of particles in a mole is called the **Avogadro constant**, N_A.

$N_A = 6.02 \times 10^{23}\,\text{mol}^{-1}$ (to three significant figures)

The Avogadro constant is given in the IB *Physics data booklet*.

Therefore, by definition, one mole of carbon-12 has a mass of exactly 12 g and contains 6.02×10^{23} atoms. The same number of atoms of hydrogen would have a mass of only 1.0 g because each hydrogen atom is only $\frac{1}{12}$ the mass of a carbon atom. Similarly, 63.5 g of copper would contain the same number of atoms because, on average, each copper atom is 63.5 times more massive than a hydrogen atom.

Figure 3.9 One mole of water, sugar, copper and aluminium

Many substances are molecular, and each molecule may consist of two or more atoms. A mole of a molecular substance contains Avogadro's number of *molecules*. For example, oxygen is a molecule with two atoms, O_2, so a mole of oxygen molecules (6.02×10^{23} molecules) has a mass of 32 g. The **molar mass** of oxygen is $32\,\text{g}\,\text{mol}^{-1}$.

The molar mass of a substance is defined as the mass that contains one mole of (its defining) particles (unit: $\text{g}\,\text{mol}^{-1}$) .

Table 3.2 lists the molar masses of some common substances, as well as the number of particles that this involves. Figure 3.9 shows what one mole of a number of different substances looks like.

Table 3.2 Molar masses

Substance	Molar mass/$\text{g}\,\text{mol}^{-1}$	Particles
aluminium	27.0	6.02×10^{23} atoms of aluminium
copper	63.5	6.02×10^{23} atoms of copper
gold	197	6.02×10^{23} atoms of gold
hydrogen	2.02	12.04×10^{23} atoms of hydrogen combined to make 6.02×10^{23} molecules
oxygen	32.0	12.04×10^{23} atoms of oxygen combined to make 6.02×10^{23} molecules
water	18.0	6.02×10^{23} atoms of oxygen combined with 12.04×10^{23} atoms of hydrogen to make 6.02×10^{23} molecules of water
air (at normal temperature and pressure)	~ 29	6.02×10^{23} molecules from a mixture of gases
sugar (sucrose)	342	6.02×10^{23} molecules

Worked example

2 The molar mass of water is $18\,g\,mol^{-1}$. How many molecules are there in 1 kg of water?

Number of moles (amount of water) in $1\,kg = \frac{1000}{18} = 55.6\,mol$

Number of molecules in $55.6\,mol = 55.6 \times (6.02 \times 10^{23}) = 3.34 \times 10^{25}$ molecules

You will need to use data from Table 3.2 to answer these questions.

6 **a** What mass of aluminium will contain exactly 4 mol of atoms?
b What mass of sucrose will contain 1.0×10^{22} molecules?
c How many moles are there in 2.00 kg of carbon?
d How many molecules are there in 1.0 kg of carbon dioxide?

7 A gold ring had a mass of 12.3 g.
a How many moles of gold atoms did it contain?
b When it was weighed 50 years later its mass had decreased to 12.2 g. How many atoms were 'lost' from the ring on average every second during this time?

8 **a** If the density of air in a classroom is $1.3\,kg\,m^{-3}$, what is the total mass of gas in a room with dimensions $2.5\,m \times 6.0\,m \times 10.0\,m$?
b Approximately how many moles of air are in the room?
c Approximately how many molecules are in the room?

9 The density of aluminium is $2.7\,g\,cm^{-3}$.
a What is the volume of one mole of aluminium?
b What is the average volume occupied by one atom of aluminium?
c Approximately how far apart are atoms in aluminium?

10 What mass of oxygen contains 2.70×10^{24} molecules?

3.2 Thermal properties of matter

Transferring energy to something to make it hotter has been an activity carried out by humans for many thousands of years. We can also make something colder by removing thermal energy from it. Such energy transfers may also result in a change of **phase** (for example a liquid becoming a gas, or a liquid becoming a solid).

The temperature rise produced when thermal energy is transferred to an object depends on the amount of energy transferred, what material(s) the object is made from and its mass. Clearly, large amounts of energy supplied to small masses will tend to produce big temperature rises, but some substances heat up much more quickly than others.

Heating and cooling graphs

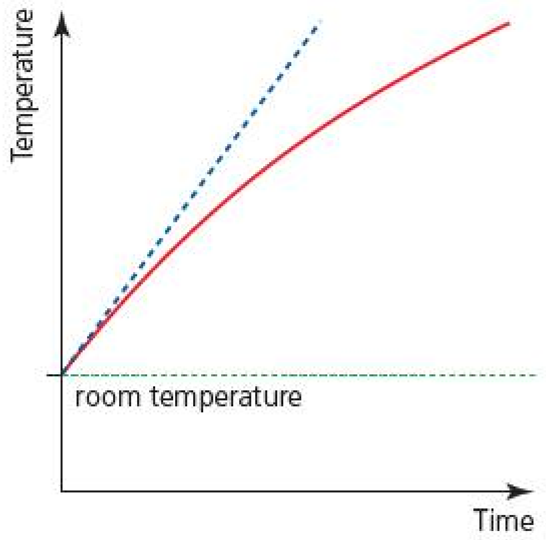


Figure 3.10 A typical graph of temperature against time for heating at a constant rate

The dotted blue straight line in Figure 3.10 shows how the temperature of an object heated at a constant rate changes with time under the idealized circumstances of no heat losses. The temperature rises by equal amounts in equal times. However, heat losses to the surroundings are unavoidable, so the curved red line represents a more realistic situation. The curve shows that the *rate* of temperature rise decreases as the object gets hotter. This is because thermal energy losses (from the object to the surroundings) are greater with larger temperature differences. If energy continues to be supplied, the object will eventually reach a constant temperature when the input power and rate of thermal energy loss to the surroundings are equal to each other.

When something is left to cool naturally, the rate at which thermal energy is transferred away decreases with time because it depends on the temperature difference between the object and its surroundings. This can be seen in Figure 3.11, in which the rate of cooling (as shown by the gradients at different times) gets less and less.

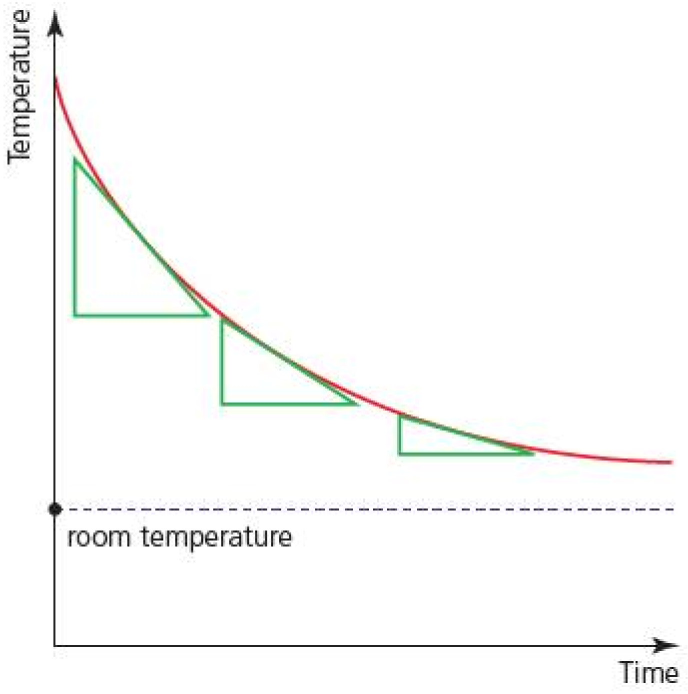


Figure 3.11 A typical graph of temperature against time for an object cooling down to room temperature. Note how the gradient decreases with time

Specific heat capacity and thermal capacity

3.2.1 Define *specific heat capacity* and *thermal capacity*.

Specific heat capacity

To compare how different substances respond to heating, we need to know how much thermal energy will increase the temperature of the same mass (usually one kilogram) of each substance by the same amount (1 K or 1 °C). This is called the substance's **specific heat capacity**, *c*. (The word 'specific' is used here simply to mean that the heat capacity is related to a certain amount of the material, namely 1 kg.)

The specific heat capacity of a substance is the amount of energy needed to raise the temperature of 1 kg of the substance by 1 K. (Unit: $\text{J kg}^{-1}\text{K}^{-1}$, but sometimes $°\text{C}^{-1}$ is used instead of K^{-1}.)

The values of specific heat capacity for some common materials are given in Table 3.3.

Table 3.3 Some specific heat capacities of different materials

Material	Specific heat capacity/$\text{J kg}^{-1}\text{K}^{-1}$
copper	390
aluminium	910
water	4180
air	1000
dry earth	1250
glass (typical)	800
concrete (typical)	800

In simple terms, substances with high specific heat capacities heat up slowly compared with equal masses of substances with lower specific heat capacities (given the same power input). Similarly, substances with low specific heat capacities will cool down more quickly. It should be noted that water has an unusually large specific heat capacity. This is why it takes the transfer of a large amount of energy to change the temperature of water and the reason why water is used widely to transfer energy in heating and cooling systems.

If a quantity of thermal energy, Q, was supplied to a mass, m, and produced a temperature rise of ΔT, we could calculate the specific heat capacity from the equation:

$$c = \frac{Q}{m\Delta T}$$

(Remember that the **delta sign**, Δ, is used widely in science and mathematics to represent a small change of something.)

This equation is more usually written as follows:

$$Q = mc\Delta T$$

This equation is included in the IB *Physics data booklet*.

When a substance cools, the thermal energy transferred away can be calculated using the same equation.

The simplest way of determining the specific heat capacity of a substance experimentally is to supply a known amount of energy from an electrical heater placed inside (immersed in) the substance. Such a heater is called an **immersion heater**. This is easy enough with liquids, but for solids it is often necessary to drill a hole in the substance to allow the heater to be placed inside, thus ensuring good thermal contact.

In the two experiments shown in Figures 3.12 and 3.13, a **joulemeter** is used to measure directly the energy transferred. (More commonly, the energy can be calculated using knowledge of electrical circuits, which is covered in Chapter 5, from the equation $E = VIt$.)

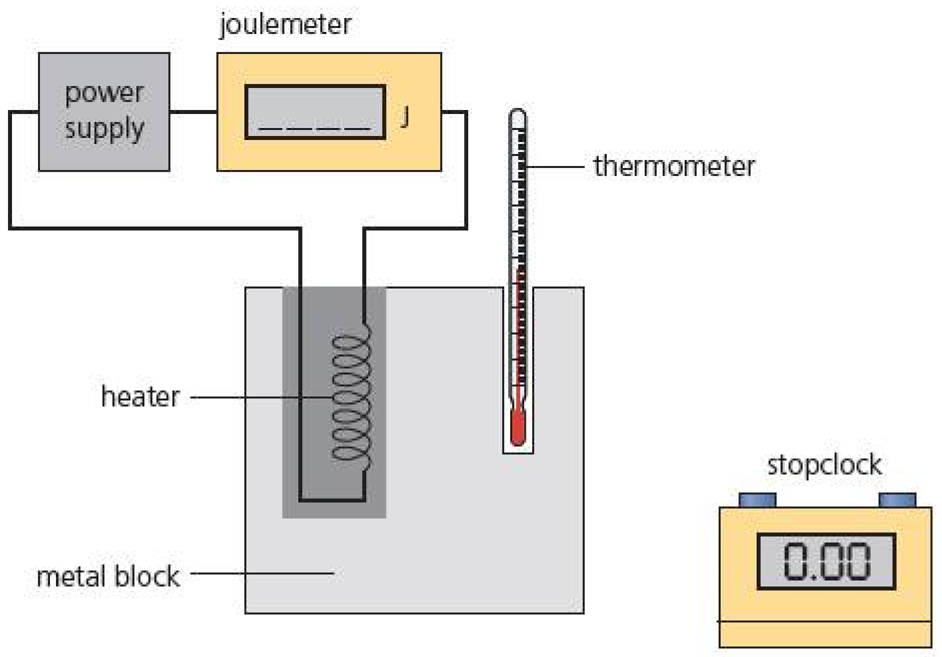


Figure 3.12 Determining the specific heat capacity of a metal

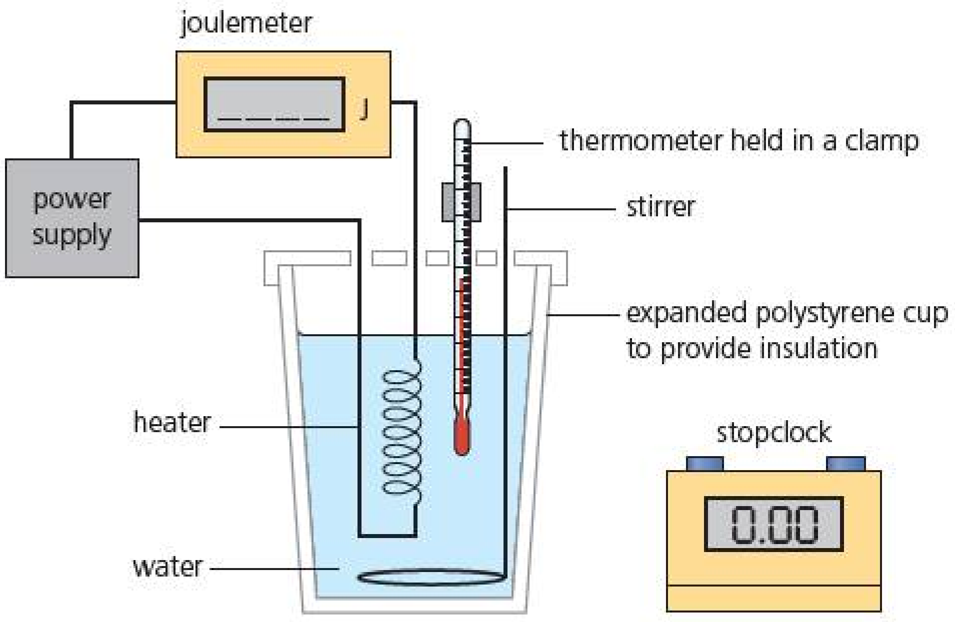


Figure 3.13 Determining the specific heat capacity of water

Worked example

3 Suppose that the metal block shown in Figure 3.12 had a mass of 1500 g and was heated for 5 minutes with an 18 W heater. If the temperature of the block rose from 18.0 °C to 27.5 °C, calculate its specific heat capacity.

$$c = \frac{Q}{m\Delta T}$$

and $Q = Pt$

so, $$c = \frac{18 \times (5 \times 60)}{1.5 \times (27.5 - 18.0)}$$

$$= 380\,\text{J kg}^{-1}\,°\text{C}^{-1}$$

In doing such calculations, it has to be assumed that all of the substance was at the same temperature and that the thermometer accurately recorded that temperature at the relevant times. In practice, both of these assumptions may lead to significant inaccuracies in the calculated value.

Furthermore, in any experiment involving thermal energy transfers and changes in temperature, there will be unavoidable losses (or gains) from the surroundings. If accurate results are required, it will be necessary to use **insulation** to limit these energy transfers, which will otherwise usually lead to an overestimate of the substance's specific heat capacity (because some of the energy input went to the surroundings rather than into the substance). The process of insulating something usually involves surrounding it with a material which traps air (a poor conductor) and is often called **lagging.**

3.2.2 Solve problems involving specific heat capacities and thermal capacities.

In the following questions, assume that there is no energy transferred to or from the surroundings.

11 How much energy is needed to raise the temperature of a block of metal of mass 3.87 kg by 54 °C if the metal has a specific heat capacity of 456 J kg^{-1} K^{-1}?

12 What is the specific heat capacity of a liquid that requires 3840 J to raise the temperature of 156 g by 18.0 K?

13 Air has a density of 1.3 kg m^{-3} and specific heat capacity of 1000 J kg^{-1} °C^{-1}. If 500 kJ was transferred to a room of volume 80 m^3, what was the temperature rise?

14 If 1.0 MJ of energy is transferred to 15.0 kg of water at 18.0 °C, what will the final temperature be?

15 A drink of mass 500 g has been placed in a glass of mass 250 g glass (of specific heat capacity 850 J kg^{-1} °C^{-1}) in a refrigerator. How much energy must be removed to cool the drink and the glass from 25 °C to 4 °C? (Assume the drink has the same specific heat capacity as water.)

16 A 20 W immersion heater is placed in a 2.0 kg iron block at 24 °C for 12 minutes. What is the final temperature? (Specific heat capacity of iron = 444 J kg^{-1} °C^{-1}.)

17 How long will it take a 2.20 kW kettle to raise the temperature of 800 g of water from 16.0 °C to its boiling point?

18 An air conditioner has a cooling power of 1200 W and is placed in a room containing 100 kg of air (specific heat capacity 1000 J kg^{-1} °C^{-1}) at 30 °C. What will the temperature be after the air conditioner has been running for 10 minutes?

19 A water heater for a shower is rated at 9.0 kW. If water at 15 °C flows through it at a rate of 15 kg every 3 minutes, what will be the temperature of the water in the shower?

20 A burner on a gas cooker raises the temperature of 500 g of water from 24 °C to 80 °C in exactly 2 minutes. What is the effective average power of the burner?

Additional Perspectives

Geothermal power

At a number of locations around the world, naturally occurring hot water has been used for thousands of years for bathing and for heating rooms. In recent years, attention has turned to extracting more of this thermal energy, with the aim of reducing the adverse effects of using fossil fuels (discussed in detail in Chapter 8).

Since its formation, the core of the Earth has been extremely hot. It has remained hot due to energy released in the decay of radioactive materials. Some of this internal energy is continuously transferred as thermal energy to rocks close to the surface of the Earth. A volcano is a dramatic example of this, and the use of geothermal energy is mostly (but not exclusively) to be found in countries known for their volcanic activity.

Figure 3.14 Iceland gets about 30% of its energy needs from geothermal power

Currently, geothermal power provides much less than 1% of the total world's needs, but three countries each generate more than 25% of their electrical power using geothermal energy: Iceland (Figure 3.14), the Philippines and El Salvador.

In some locations the energy is directly and conveniently available in the form of hot water, but in other places water has to be pumped down to hot rocks a long way beneath the surface. The most common and efficient use of geothermal energy is achieved with the direct use of hot water for heating, but where temperatures are high enough electricity can be generated.

Questions

1 Where are the nearest 'hot springs'/spas to where you live? How hot is the water and what is it used for?

2 Use the Internet to find out as much information as you can about any one particular geothermal power station from around the world. Prepare a presentation for the rest of your group. (Answer questions such as: Why is it at that location? Did engineers have to dig into the Earth's crust or was the energy accessible close to the surface? Is the internal energy mostly in rocks or water? How is the energy transferred for use? Is there any pollution? What is the useful power output? What is the efficiency?). Make sure that you are using the terms 'thermal energy' and 'internal energy' correctly. Try not to use the word 'heat'.

Exchanges of thermal energy

Figure 3.2 showed the temperature–time graphs of two substances at different temperatures placed in good thermal contact so that thermal energy can be transferred relatively quickly, assuming that the system is insulated from its surroundings. Under these circumstances *the heat given out by one substance is equal to the heat absorbed by the other substance*. Exchanges of thermal energy can be used as an alternative means of determining a specific heat, as illustrated by question **21**, or in the determination of the energy that can be transferred from a food or a fuel. (This may be done in a **calorimeter**, which is a piece of apparatus designed to be used to measure the amount of thermal energy transferred in heating processes, chemical reactions, changes of phase, etc.)

3.2.2 Solve problems involving specific heat capacities and thermal capacities.

21 A large metal bolt of mass 53.6g was heated for a long time in an oven at 245°C. The bolt was transferred as quickly as possible from the oven into a beaker containing 257.9g of water initially at 23.1°C (Figure 3.15). The water was stirred continuously and the temperature rose to a maximum of 26.5°C.

a Calculate the energy transferred to the water.
b Why was the bolt kept in the oven for a long time?
c Why was the transfer made quickly?
d Calculate the specific heat capacity of the material of which the bolt is made.
e Why was it necessary to stir the water?
f Is the value for the specific heat capacity of the bolt likely to be an underestimate or an overestimate of its true value? Explain your answer.

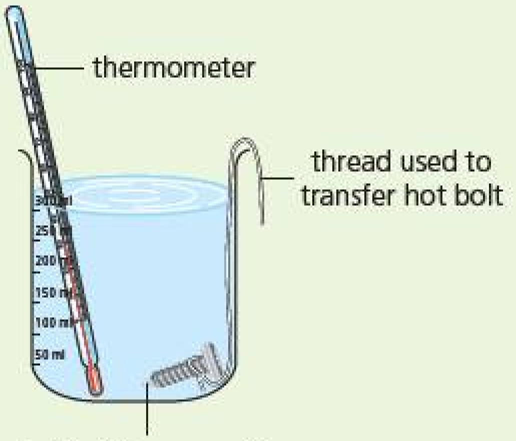


Figure 3.15 A hot metal bolt placed in cold water

22 When 14.5g of a certain fuel was burned, thermal energy was transferred to 63.9g of water. The water temperature rose from 18.7°C to 42.4°C. Assuming that no thermal energy was transferred into the surroundings, calculate the maximum amount of energy that can be transferred from 1kg of this fuel.

23 When running water into a bath tub, 84kg of water at 54°C was added to 62kg of water at 17°C.
a What was the final temperature?
b What assumption did you make?

24 Sand initially at 27.2°C is added to an equal mass of water of at 15.3°C. If the specific heat capacity of sand is 822 $\text{J kg}^{-1}\,°\text{C}^{-1}$ and the specific heat capacity of water is 4180 $\text{J kg}^{-1}\,°\text{C}^{-1}$, what will be the final temperature of the mixture? (Assume no energy is transferred to the surroundings.)

Thermal capacity

Everyday objects are not often made of only one substance, and referring to a specific amount (a kilogram) of such objects is not sensible. In such cases we refer to the **thermal capacity** of the whole object, for example we might want to know the thermal capacity for a room and its contents when choosing a suitable heater or air conditioner.

The thermal capacity of an object is the amount of energy needed to raise its temperature by 1K. (Unit: J K^{-1} or $\text{J}\,°\text{C}^{-1}$.)

Thermal capacity, $C = \dfrac{Q}{\Delta T}$

Worked example

4 How much thermal energy is needed to increase the temperature of a kettle and the water inside it from 23°C to 77°C if its thermal capacity is 6500 J K^{-1}?

$Q = C\Delta T$
$Q = 6500 \times (77 - 23)$
$Q = 3.51 \times 10^5\,\text{J}$

3.2.2 Solve problems involving specific heat capacities and thermal capacities.

For the following questions, assume that no energy is transferred to or from the surroundings.

25 The thermal capacity of a saucepan and its contents is 7000 J K^{-1}. How long will it take a 1.5kW electric cooker to raise the temperature of the saucepan and its contents from 22°C to 90°C?

26 If it took 32 minutes for a 2.5kW heater to raise the temperature of a room from 8°C to 22°C, what was the room's thermal capacity?

27 If the contents of a refrigerator with a cooling power of 405W have a thermal capacity of 23.9 kJ K^{-1}, how long will it take for the average temperature to be reduced from 19.5°C to 5.20°C?

Additional Perspectives

Buildings with high thermal capacity

In countries with hot, dry climates, keeping cool in the daytime can be a major problem, especially if is not possible or desirable to use air conditioners. People in these countries have known for centuries about the advantages of building homes with large mass and, therefore, large thermal capacity (Figure 3.16). In such climates, it is also common for the air temperature to fall significantly at night because of the low humidity and the lack of clouds, which allows thermal energy to be radiated away.

Figure 3.16 A thick-walled traditional African house

During the day, the radiated thermal energy from the Sun warms the building, but using large amounts of materials like earth and stone (which have relatively high specific heat capacities) to make a home with a large thermal capacity ensures that the temperature rise is not too quick. At night, for the same reason, the temperature of the building will not fall suddenly and the people inside can keep warm. This effect produces a pleasant thermal 'lag', with the temperature inside the building 'cooler' in the morning and 'hotter' in the late afternoon or evening.

In essence, buildings with high thermal capacity 'average' out the temperature extremes that would otherwise be caused by the weather conditions and any significant changes in temperature between day and night. This is a useful property of nearly all buildings, whether they are located in hot or cold climates, and most buildings are best designed to have large thermal capacities. However, the economic costs of using large quantities of otherwise unnecessary building materials will usually limit the mass of buildings.

Question

1 a Make estimates of the various quantities needed, and then calculate the approximate thermal capacity of a house like that shown in Figure 3.16.

b Assuming that on a hot day the radiated thermal energy from the Sun arrives perpendicularly down on the building's roof at a rate of $850\,\mathrm{W\,m^{-2}}$, calculate the maximum temperature rise produced during one hour in the middle of the day. Assume that 5% of the thermal energy is absorbed in, and spread evenly throughout, the building. (It is likely that every effort will be made to ensure that the building does *not* absorb radiant energy falling on it – by using light colours to reflect back the energy.)

c The temperature rise you calculated in **b** was probably much higher than that which actually happens. Suggest why.

d Sketch a graph to show how you think that the air temperature in a hot, dry country may vary over a period of two cloudless days and nights. Then add to your graph a sketch showing how the inside temperature of a building such as that shown in Figure 3.16 varies over the same time.

e Why do you think that air temperatures are always measured in the shade?

The transfer of mechanical energy to internal energy

We know that, without a force pushing them forwards, all moving objects will tend to decelerate and eventually stop because of the forces of *friction*. Sometimes we need to provide extra forces to stop a moving object, such as a car. This, too, is usually done with the help of friction. When friction acts to slow motion, the macroscopic ordered kinetic energy of the moving molecules is transferred to the random, disordered kinetic energies of molecules in both surfaces. There will be a rise in internal energy and temperature. Thermal energy will also be transferred to the surroundings.

The transfer of ordered energy to disordered energy cannot be stopped or reversed. If a moving object has been stopped by friction, it is simply not possible to transfer the increased internal energy (involving the random kinetic energies of molecules) back into the macroscopic overall, ordered kinetic energy of the moving object.

Consider the example of a car stopping under the action of its brakes. Friction between the tyres and the road, as well as air resistance on the surface of the vehicle, will contribute to the resistive forces, but for simplicity we might assume that all the kinetic energy of the car is transferred to raising the temperature of the brakes (Figure 3.17). An example of this kind of calculation can be found in question 28 below.

Falling objects transfer gravitational potential energy to kinetic energy. If they do not bounce when they hit the ground, we can assume in the short-term (for ease of calculation), that all of the energy is transferred into the internal energy of the object. In other words, no thermal energy is transferred to the ground or surroundings. An interesting example is the expected small rise in temperature of water after it has fallen over a waterfall (Figure 3.18).

Figure 3.17 The disc brakes on a car can get very hot

Figure 3.18 The water temperature at the bottom of the waterfall will be a little greater than at the top

28 A car of total mass 1200 kg, travelling at 17.3 m s^{-1} has four disc brakes, each of mass 424 g and specific heat capacity 1580 J kg^{-1} °C^{-1}.
- **a** Calculate the kinetic energy of the car.
- **b** Estimate the maximum temperature rise of the brakes when they are used to decelerate the car to rest.
- **c** Why would a temperature rise be greater if the deceleration was greater?

29 A 12 g bullet travelling at 670 m s^{-1} (twice the speed of sound) is fired into a wooden block of mass 650 g.
- **a** Estimate the temperature rise if the combined thermal capacity of the bullet and wood was 1880 J K^{-1}.
- **b** What assumption have you made?
- **c** Without using a detailed calculation, what temperature rise would be produced by a similar bullet travelling at half the speed?

30 Using a steady force of 160 N, a paddle wheel was turned inside a tank containing 8.00 kg of water at 18.6 °C (Figure 3.19).
- **a** If the rotating force was moved a total distance of 250 m, how much work was done?
- **b** Estimate the final temperature of the water.
- **c** Why is it difficult to demonstrate that mechanical energy can be directly transferred to an equivalent amount of internal energy?

31 A 1800 g piece of lead fell 14.8 m off the roof of a building.
- **a** If the specific heat capacity of lead is 130 J kg^{-1} °C^{-1}, estimate the temperature rise of the lead after it hit the ground.
- **b** What assumptions did you have to make in order to be able to do this calculation?
- **c** Explain why a piece of lead of twice the mass would have approximately the same temperature rise if it fell the same distance.

32 The Horseshoe Falls at Niagara has a vertical height of 53 m. What is the maximum possible temperature rise of water falling from the top to the bottom of this waterfall?

Figure 3.19 Joule's apparatus for turning mechanical energy into internal energy

Phase changes

A **phase** is a region of space in which all the physical and chemical properties of a substance are the same. A particular substance may exist in a solid phase, a liquid phase or a gaseous phase. These are sometimes called the three *states* of matter. For example, water can exist in three phases (states): liquid, ice (solid) and steam (gas) (Figure 3.20).

Solids, liquids and gases

3.2.3 Explain the physical differences between the solid, liquid and gaseous phases in terms of molecular structure and particle motion.

- In **solids** the molecules (or atoms, or ions) are held close together by strong forces (sometimes called 'bonds'), usually in regular patterns. The molecules **vibrate** about their mean positions.
- In **liquids** the molecules still vibrate, but the forces between some molecules are overcome, allowing them to move around a little. The molecules are still almost as close together as in solids, but there is little or no regularity in their arrangement, which is constantly changing.
- In **gases** the molecules are much further apart than in solids and liquids, and the forces between them are very, very small and usually negligible (except when they collide). This results in all molecules moving independently in random directions with a range of different (usually fast) speeds.

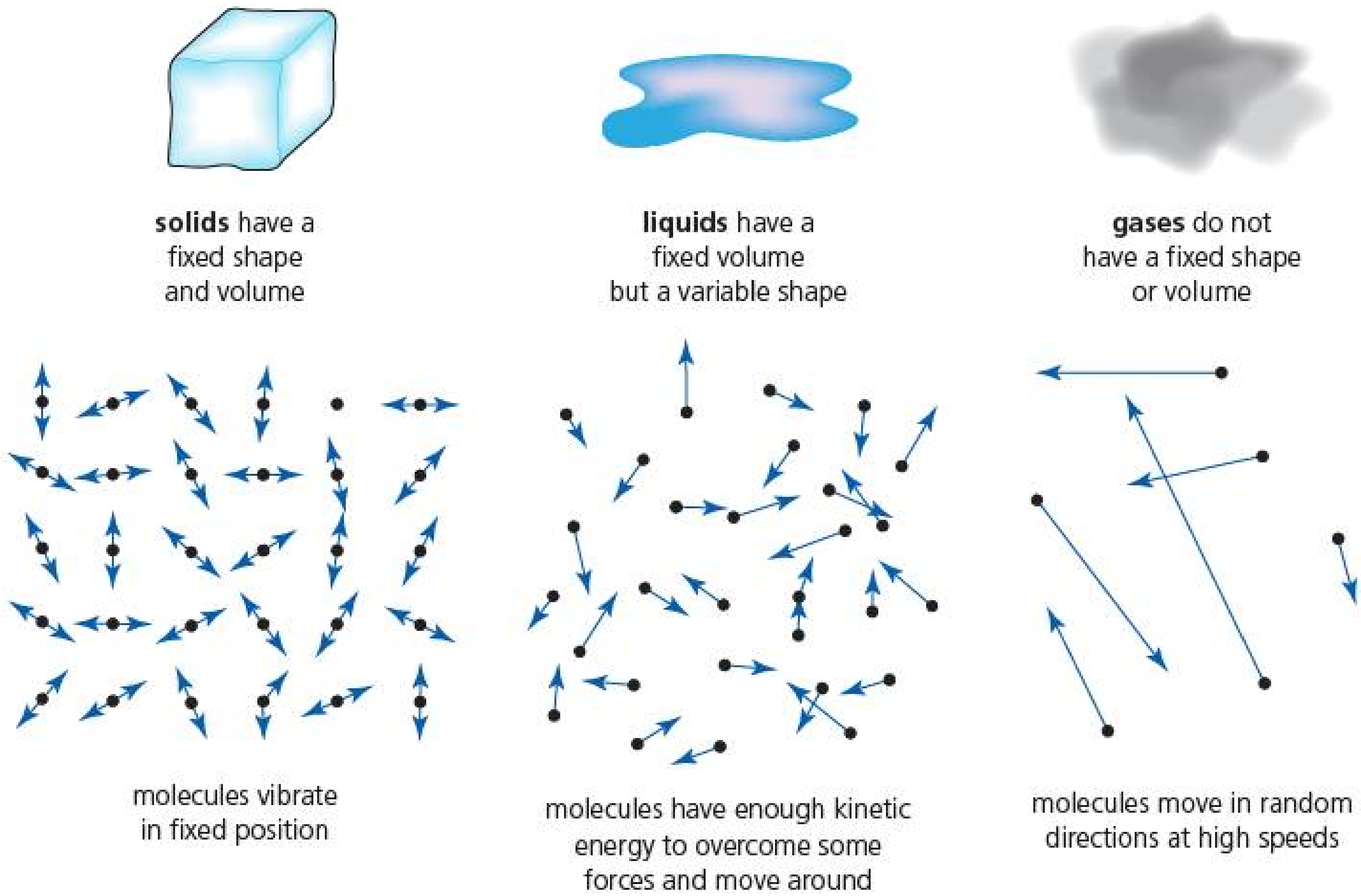


Figure 3.20 Differences between solids, liquids and gases. The arrows represent the velocities of the molecules

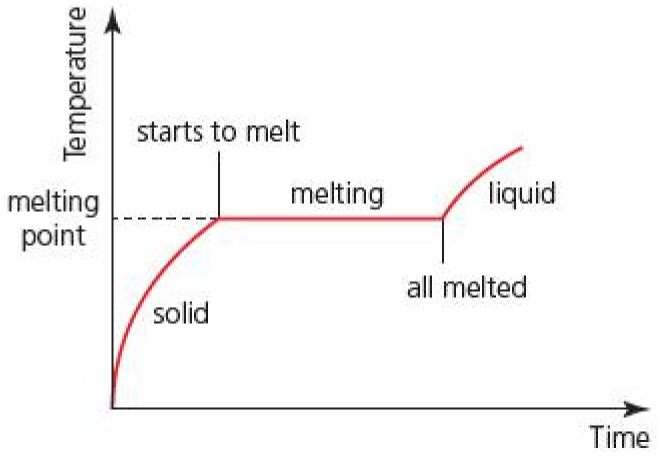


Figure 3.21 Temperature changes while a solid is heated and melted

Molecular behaviour during phase changes

3.2.4 Describe and **explain** the process of phase changes in terms of molecular behaviour.
3.2.5 Explain in terms of molecular behaviour why temperature does not change during a phase change.

When thermal energy is transferred to a solid it will usually get hotter. For many solid substances, once they reach a certain temperature they will begin to **melt** (change from a solid to a liquid), and while they are melting the temperature does not change (Figure 3.21). This temperature is called the **melting point** of the substance, and it has a fixed value at a particular air pressure (see Table 3.4). Melting is an example of a **phase change**. Another word for melting is **fusion**.

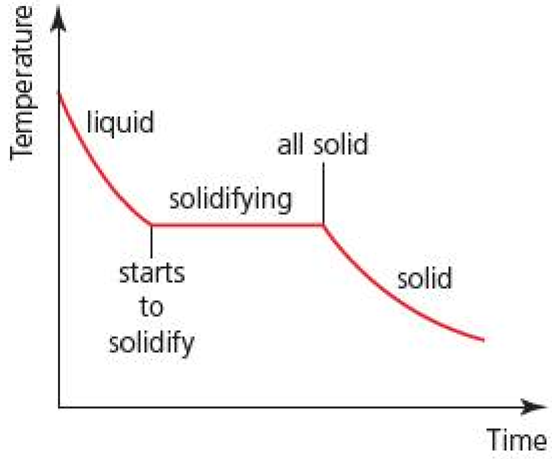


Figure 3.22 Temperature changes when a liquid cools and freezes (solidifies)

Similarly, when a liquid cools, its temperature will be constant at its melting point while it changes phase from a liquid to a solid (Figure 3.22). This process is known as *solidifying* or **freezing**. Be careful – the word 'freezing' suggests that this happens at a low temperature, but this is not necessarily true (unless we are referring to turning water into ice, for example). The phase changes of water are such common events in everyday life that we all tend to think of them as the obvious examples but, of course, many other substances can melt and freeze. For example, molten iron 'freezes' at 1538 °C.

A change of phase also occurs when a liquid becomes a gas, or when a gas becomes a liquid. The change of phase from a gas to a liquid may be by boiling or evaporation, but generally this process can be called **vaporization**. Changing from a gas to a liquid is called **condensation**. The temperature at which boiling occurs is called the **boiling point** of the substance, and it has a fixed value for a particular air pressure (see Table 3.4). Boiling points may vary considerably with different surrounding air pressures.

A graph showing the temperature change of a liquid being heated to boiling will look very similar to that for a solid melting (Figure 3.21), while the graph for a gas being cooled will look very similar to that for a liquid freezing (Figure 3.22).

To melt a solid or boil a liquid, it is necessary to heat them – thermal energy is transferred. However, as we have seen, melting and boiling occur at constant temperatures, which means that during melting or boiling there is no increase in the average random kinetic energies of the molecules. Instead, the energy supplied is used to overcome intermolecular forces and to increase separations and, therefore, electric potential energies. In the case of melting, some forces are overcome, but in the case of boiling all the remaining forces are overcome.

When a liquid freezes (solidifies) the same amount of energy per kilogram is emitted as was needed to melt it (without a change in temperature). Similarly, boiling and condensing involve equal energy transfers.

Figure 3.23 represents the four principle phase changes. Table 3.4 lists melting and boiling points of some common substances.

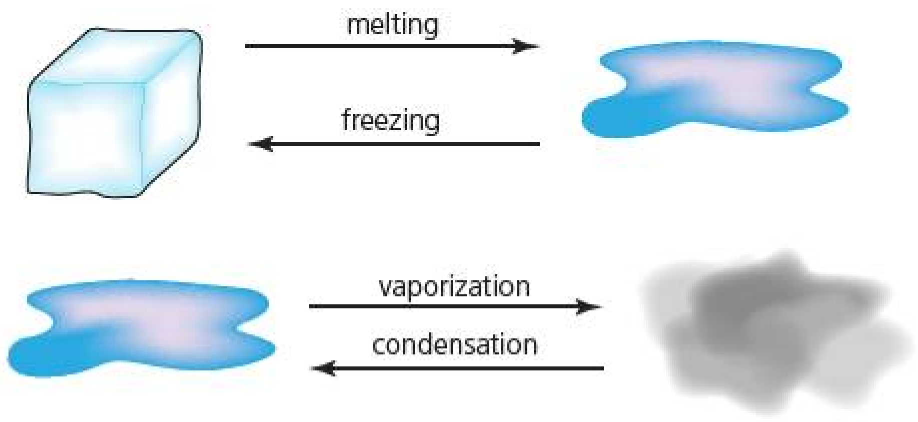


Figure 3.23 Changes of phase

Table 3.4 Melting points and boiling points of some substances (at normal atmospheric pressure)

Substance	Melting point		Boiling point	
	/°C	/K	/°C	/K
water	0	273	100	373
mercury	–39	234	357	630
alcohol (ethanol)	–117	156	78	351
oxygen	–219	54	–183	90
copper	1083	1356	2580	2853
iron	1538	1811	2750	3023

Evaporation and boiling

3.2.6 Distinguish between evaporating and boiling.

Molecules in a liquid have a range of different energies and energy is continuously transferred between them. This means there will always be some molecules near the surface that have enough energy to overcome the attractive forces that hold the molecules together in the liquid. Such molecules can escape from the surface and the macroscopic effect is called **evaporation**. The loss of the most energetic molecules means that the average kinetic energy of the molecules remaining in the liquid must decrease (until thermal energy flows in from the surroundings). This microscopic effect explains the macroscopic fall in temperature (cooling) that always accompanies evaporation from a liquid (Figure 3.24).

Evaporation occurs only from the surface of a liquid and may occur at any temperature, although the rate of evaporation increases significantly with rising temperature (between the melting and boiling points). **Boiling** occurs at a precise temperature – the temperature at which the molecules have enough kinetic energy to form bubbles inside the liquid.

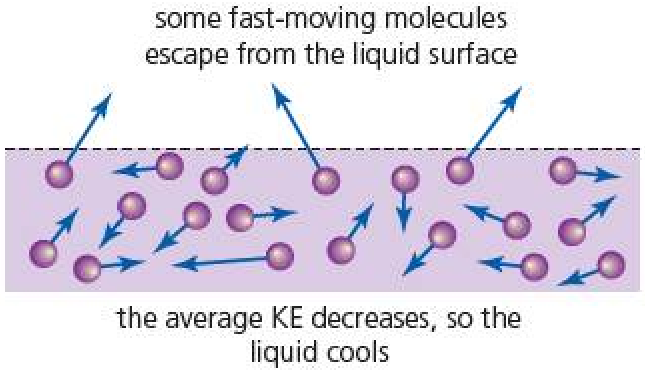


Figure 3.24 Molecules leaving a surface during evaporation

Additional Perspectives

Cooling by evaporation

The cooling effect produced by water evaporating has been used for thousands of years to keep people and buildings cool. For example, in central Asia, open towers in buildings encouraged air flow over open pools of water, increasing the rate of evaporation and the transfer of thermal energy up the tower by **convection currents**. The flow of air past people in buildings also encourages the human body's natural process of cooling by sweating.

Modern refrigerators and air conditioners also rely on the cooling produced when a liquid evaporates. The liquid or gas used is called the **refrigerant**. Ideally, it should take a large amount of thermal energy to turn the refrigerant from a liquid into a dense gas at a little below the desired temperature.

In a refrigerator, for example, after the refrigerant has removed thermal energy from the food compartment, it will be a gas and be hotter. In order to re-use it and turn it back into a cooler liquid again, it must be compressed and its temperature reduced. To help achieve this, thermal energy is transferred from the hot, gaseous refrigerant to the outside of the refrigerator (Figure 3.25).

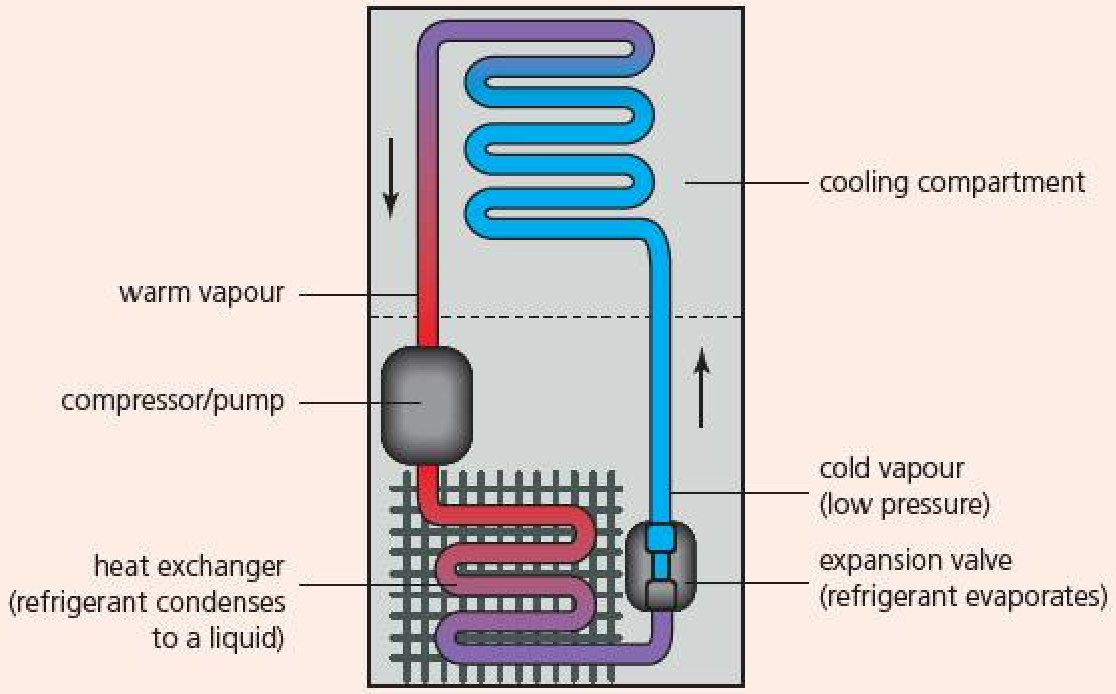


Figure 3.25 Schematic diagram of a refrigerator

As well as having suitable thermal properties, a refrigerant must also be non-poisonous and must not be harmful to the environment. This is because, over a period of time, some of the refrigerant may leak out of the system into the environment. In the 1970s and 1980s it was realized that the refrigerants being widely used (CFCs) were damaging the environment (the ozone layer in particular) and their use was discouraged and banned in many countries. More recently concern has centred on the effect of refrigerants on global warming. A lot of research has gone into producing a variety of synthetic refrigerants with the right combinations of physical and chemical properties.

Questions

1 Draw a simple sketch of an old central Asian building showing the tower, pool and convection currents.
2 Water is cheap, non-poisonous to us and to our environment, and needs a large amount of energy to make it evaporate. So why isn't it used widely as a refrigerant in air conditioners?
3 Discuss whether a fountain in the living room of your home could help to keep you cool on a hot day.
4 As explained above, thermal energy has to be removed from the refrigerant during its **cycle**. Suggest how this can be done.

Specific latent heat

3.2.7 Define *specific latent heat.*

We now want to consider the energy transfers involved with changes of phase. As an example, Figure 3.26 shows the changes in temperature that might occur when a quantity of crushed ice is heated *continuously*, first to become water, and then to become steam, as the water boils. There are two flat sections on the graph. The first flat section, at 0 °C, the melting point of water, shows where the additional energy input from the heater is used to break some molecular bonds in the ice. The temperature does not begin to rise until all the water is in the liquid phase. The second flat section occurs at 100 °C, the boiling point of water. Here the energy being supplied is used to free all the molecules from intermolecular forces.

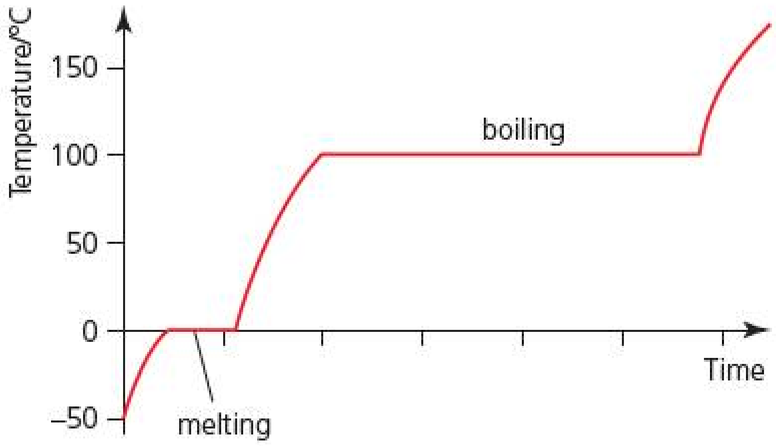


Figure 3.26 A graph of temperature against time for the heating of crushed ice

The thermal energy involved with changing potential energies during any phase change is called **latent heat** ('latent' means hidden). During melting or boiling latent heat must be transferred *to* the substance. During condensing or freezing latent heat is transferred *from* the substance. The latent heat associated with melting or freezing is called **latent heat of fusion, L_f**. The latent heat associated with boiling or condensing is known as **latent heat of vaporization, L_v**.

The **specific latent heat** of a substance, L, is the amount of energy transferred when one kilogram of the substance changes phase *at a constant temperature*. (The units are J kg^{-1}.)

That is, specific latent heat, L = thermal energy transferred/mass = Q/m, which is more commonly written as:

$$Q = mL$$

This equation can be found in the IB *Physics data booklet.*

For example, the specific latent heat of fusion of lead is $2.45 \times 10^4\,J\,kg^{-1}$ and its melting point is 327 °C. This means that $2.45 \times 10^4\,J$ is needed to melt 1 kg of lead at a constant temperature of 327 °C.

Experiments to determine specific latent heats (water is often used as a convenient example) have many similarities with specific heat capacity experiments. Usually either an electrical heater of known power is used to melt or boil a substance, or a warmer liquid is used to melt a cooler solid. Question 37 below describes such an experiment.

Worked example

5 The latent heats of vaporization of water and ethanol are $2.27 \times 10^6\,J\,kg^{-1}$ and $8.55 \times 10^5\,J\,kg^{-1}$.

a Which one is 'easier' to turn into a gas/boil (at the same pressure)?

b How much thermal energy is needed to turn 50 g of ethanol into a gas at its boiling point of 78.3 °C?

a It is 'easier' to boil ethanol because much less energy is needed to turn each kilogram into gas.

b $Q = mL$

$Q = 0.050 \times (8.55 \times 10^5)$

$Q = 4.3 \times 10^4\,J$

3.2.8 Solve problems involving specific latent heats.

33 If the latent heat of fusion of a certain kind of chocolate is 160 000 $J\,kg^{-1}$, how much thermal energy is removed from you when a 10 g bar of chocolate melts in your mouth?

34 Water is heated in a 2250 W kettle. When it reaches 100 °C it boils and in the next 180 s the mass of water reduces from 987 g to 829 g. Use these figures to estimate the latent heat of vaporization of water.

35 Why should you expect that the latent heats of vaporization of substances are usually greater than their latent heats of fusion?

36 0.53 g of steam at 100 °C condensed on an object and then the water rapidly cooled to 35 °C.

a How much thermal energy was transferred from the steam:

i when it condensed?

ii when the water cooled down?

b Suggest why a burn received from steam is much worse than from water at the same temperature (100 °C). (The latent heat of vaporization of water is $2.27 \times 10^6\,J\,kg^{-1}$ and the specific heat capacity of water is $4180\,J\,kg^{-1}\,°C^{-1}$.)

37 In the experiment shown in Figure 3.27, two identical 50 W immersion heaters were placed in some ice in two separate funnels. The heater above beaker A was switched on, but the heater above B was left off. After five minutes it was noted that the mass of melted ice in beaker A was 54.7 g, while the mass in beaker B was 16.8 g.

a What was the reason for having ice in *two* funnels?

b Use these figures to estimate the latent heat of fusion of ice.

c Suggest a reason why this experiment does not provide an accurate result.

d Describe one change to the experiment that would improve its accuracy.

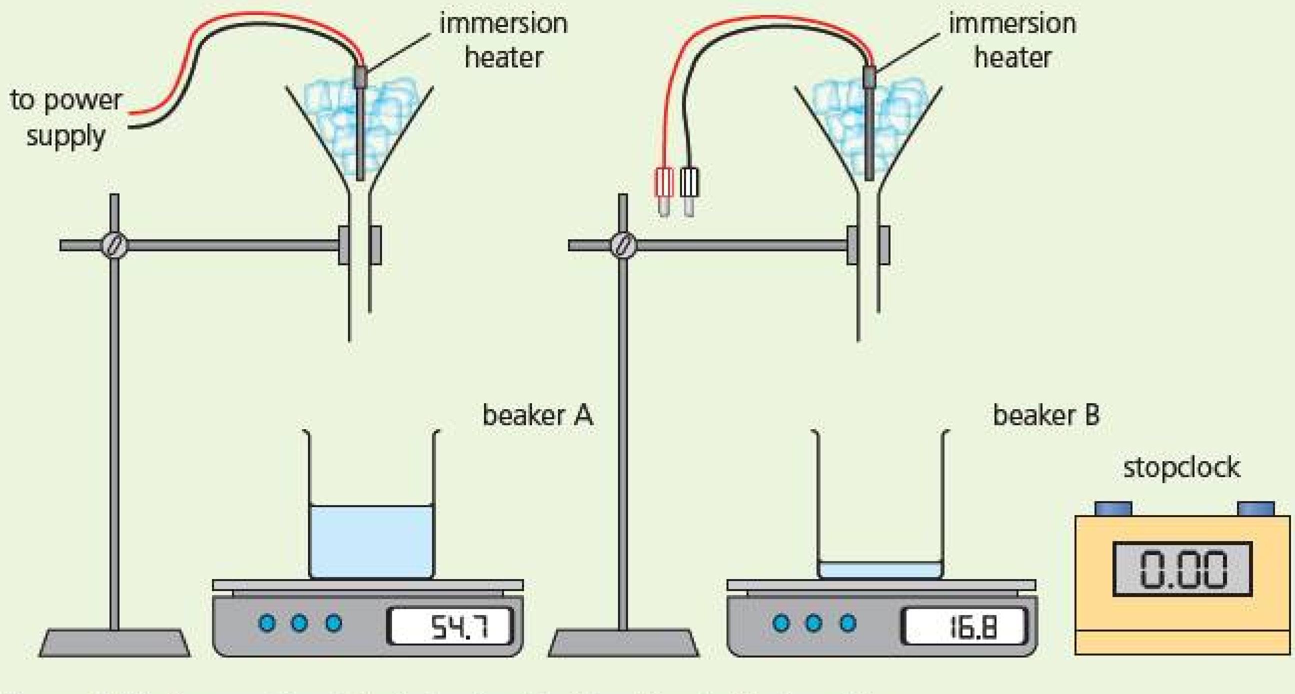


Figure 3.27 An experiment to determine the latent heat of fusion of ice

38 120 g of water at 23.5 °C was placed in a plastic tray for making ice cubes. If the tray was already at 0 °C, calculate the thermal energy that has to be removed from the water to turn it to ice at 0 °C. (The latent heat of fusion of water is $3.35 \times 10^5\,\mathrm{J\,kg^{-1}}$.)

39 Clouds are condensed droplets of water and sometimes they freeze to become ice particles. Suppose a cloud had a mass of 240 000 kg, how much thermal energy would be released if it all turned to ice at 0 °C?

40 Some water in a glass container are both at a temperature of 23 °C and have a thermal capacity of $1500\,\mathrm{J\,K^{-1}}$. If a 48 g lump of ice at −8.5 °C is placed in the water and the mixture is stirred until all of the ice has melted, what is the final temperature? (Specific heat capacity of ice is $2100\,\mathrm{J\,kg^{-1}\,K^{-1}}$. The latent heat of fusion of water is $3.35 \times 10^5\,\mathrm{J\,kg^{-1}}$.)

Additional Perspectives

Freezing hot water

To turn water into ice, thermal energy needs to be removed from it by placing it in a freezer. The time that this takes will depend on the amount of water and the cooling power of the freezer. Energy needs to be removed to bring the water down from its starting temperature to zero, and then the latent heat of fusion must be extracted. Strangely, it is sometimes said that 'hot water freezes quicker than cold water'.

Question

1 Can this ever be true? Use the Internet to investigate whether there might be any truth in this surprising claim. (Or, is it just a myth?) How would you carry out experiments to test the claim for yourself?

Kinetic model of an ideal gas

A good understanding of internal energy, thermal energy (heat) and temperature is fundamental to so much of science. Such an understanding develops from knowledge of molecular behaviour. Of the three phases of matter, gases are the easiest to understand *because* we can usually assume that the motions of the molecules are random and independent. The random microscopic behaviour of countless billions of individual molecules results in a totally predictable macroscopic behaviour of gases that is much simpler to understand than the more complex intermolecular interactions in solids and liquids.

We can begin by making some simplifying assumptions about the behaviour of gas molecules. This theory is known as the **kinetic model of an ideal gas**. It is called the 'kinetic' theory because it involves moving molecules. The model can then be used to explain pressure, temperature and the macroscopic behaviour of gases.

3.2.10 State the assumptions of the kinetic model of an ideal gas.

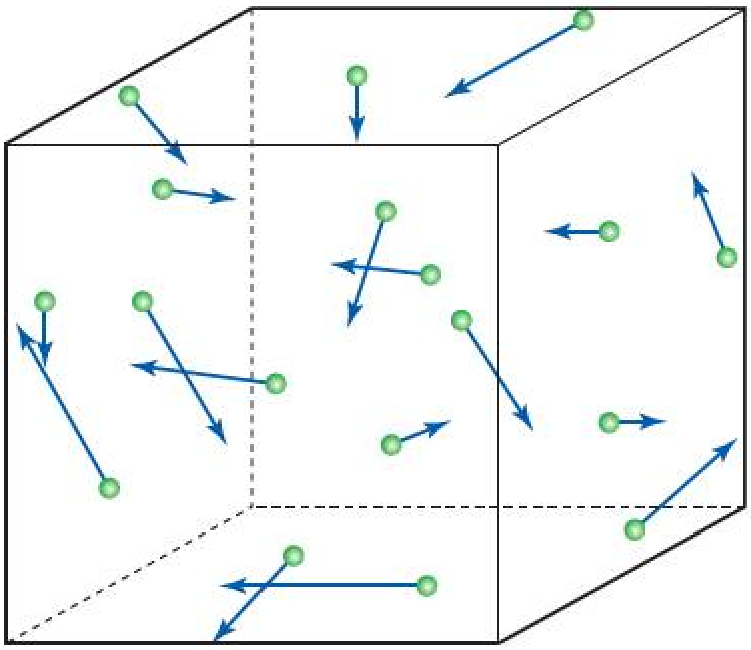


Figure 3.28 Gas molecules moving around at random in a container

Assumptions of the kinetic model of an ideal gas

The microscopic model of a gas consists of a large number of molecules moving randomly. We can imagine these as the 'molecules in a box', as shown in Figure 3.28. Arrows of different lengths represent the random velocities. There are many computer simulations of gas behaviour to represent how the molecules move over a period of time and it is recommended to view some of these.

To simplify the theory, the following assumptions are made and then the gas can be described as 'ideal'.

- An ideal gas contains a very large number of identical molecules.
- The volume of the molecules is negligible compared to the total volume occupied by the gas.
- The molecules are in completely random motion.
- There are negligible forces between the molecules, except when they collide. This means that any changes of internal energy of an ideal gas are assumed to be only in the form of changes of random kinetic energy. The molecules have no electrical potential energy.
- All collisions are elastic, that is, the total kinetic energy of the molecules remains constant at the same temperature. This means that there is no energy transferred from a gas to its surroundings and the average random speed of its molecules will not decrease. If this were not true, all gases would cool down and their molecules would fall to the bottom of their containers and condense to liquids!

Using the kinetic theory of an ideal gas to explain gas pressure

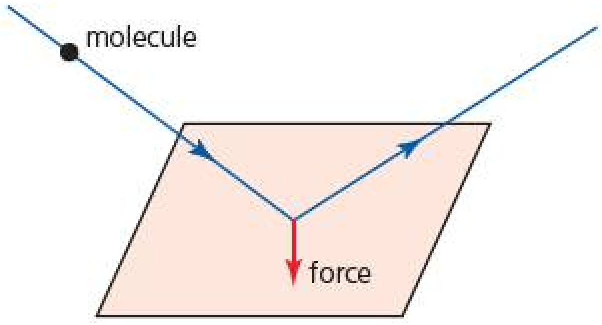


Figure 3.29 Every molecular collision with a wall creates a tiny force on the wall

When a molecule hits (collides with) a wall of a container, it exerts a force on the wall (Figure 3.29).

The size of the force could be calculated knowing the mass and change in momentum of the molecule, together with the duration of the impact (see Chapter 2). Each and every collision may result in a different sized force, and it is not realistic to know the size of the forces that individual molecules exert on the walls. However, because the molecular motions are random and because of the incredibly large number of them, the total force caused by many molecular collisions on any given area of container wall will be completely predictable and will (usually) be the same at all places in the container. This is called the **pressure** of the gas.

Collisions between molecules (intermolecular collisions) are happening all the time, but they simply result in random changes to molecular velocities and have no overall effect on pressure or other macroscopic properties of the gas.

Pressure

3.2.9 Define pressure.

The effect of a force often depends upon the area on which it acts. For example, when the weight of a solid pushes down on a surface, the consequences usually depend on the area underneath it, as well as the magnitude of the weight.

Pressure is defined as:

pressure = force per unit area

$$P = \frac{F}{A}$$

This formula is in the IB *Physics data booklet*. The SI unit of pressure is the pascal, Pa. 1 Pa = 1 N m^{-2}.

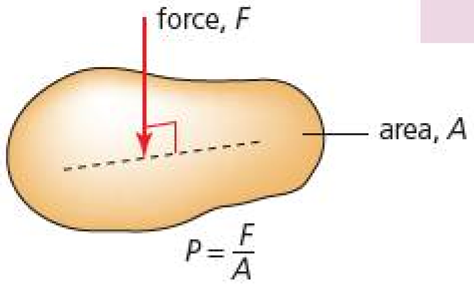


Figure 3.30 Defining pressure

Note that the force used to calculate pressure is perpendicular to the surface, as shown in Figure 3.30. It is a **normal** force.

The pressure caused in a gas by molecular collisions acts equally in all directions and has a typical value in the order of 10 N on every square centimetre. In SI units, the usual pressure of the air around us (**atmospheric pressure**) is 1.0×10^5 Pa at sea level. Atmospheric pressure acts upwards and sideways as well as downwards.

41 Normal air pressure is about 10^5 Pa.

a What total force acts on an area of 1 mm^2?

b If a student had a body with a total surface area of 1.5 m^3, what would the total force on their skin be?

c Why doesn't such a large force have any noticeable effect?

42 Why does the pressure on underwater divers increase when they go deeper?

43 **a** The pressure under our feet acts down on the ground, but the pressure in a gas acts in all directions. Explain this.
b In what direction(s) does pressure in a liquid act?

Using the kinetic theory of an ideal gas to explain gas temperature

3.2.11 **State** that temperature is a measure of the average random kinetic energy of the molecules of an ideal gas.

Earlier in this chapter, temperature was explained as a way of determining the direction of thermal energy transfer, now we can interpret temperature in terms of molecular energies. When a gas is heated, the average speed of its molecules increases. The thermal energy transferred into the gas increases the kinetic energy of the molecules.

Temperature is a measure of the average random kinetic energy of the molecules.

Although we are discussing *ideal* gases, this interpretation of temperature can be applied to all substances. For example, in a comfortable air temperature for humans, the molecules are moving with an average kinetic energy of about 6×10^{-21} J.

It is important to realize that individual molecules will be moving with a wide range of speeds and kinetic energies. The value given above is just an average value. Temperature is a concept that can only be applied to a collection of a very large number of molecules – not to individual molecules. That is, temperature is a macroscopic concept, although statistically it has a microscopic interpretation.

Additional Perspectives

The distribution of molecular speeds

The important link between temperature and average molecular kinetic energies can be quantified if we use the Kelvin scale (rather than the Celsius scale). For an 'ideal' gas, we can write:

$$\text{average molecular kinetic energy} = \tfrac{3}{2}kT$$

The term k is a very important constant linking macroscopic temperature to microscopic energies. It is called the **Boltzmann constant** ($k = 1.38 \times 10^{-23}\,\text{J K}^{-1}$). With this equation we can easily calculate an approximate average kinetic energy of, say, molecules in the air that you are breathing (6×10^{-21} J).

Oxygen molecules (O_2) have a mass of 5.3×10^{-26} kg. Using $KE = \tfrac{1}{2}mv^2$, we can calculate that a typical speed for oxygen molecules is about $480\,\text{m s}^{-1}$.

Figure 3.31 shows the range and distribution of molecular speeds in a typical gas, and how it changes as the temperature increases. This is known as the Maxwell–Boltzmann distribution.

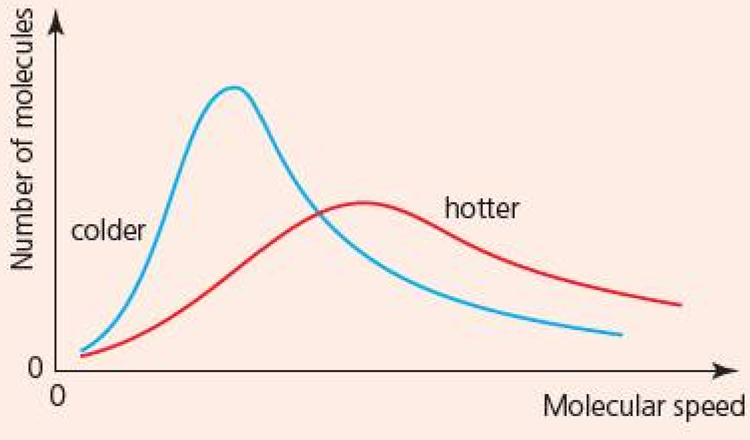


Figure 3.31 Typical distributions of molecular speeds in a gas

Note that there are no molecules with zero speed and very few with very high speeds. Molecular speeds and directions (that is, molecular velocities) are continually changing as the result of intermolecular collisions. As we have seen, higher temperature means higher kinetic energies and therefore higher molecular speeds, but the range of speeds also broadens (and so the peak becomes lower, keeping the area under the graph, which represents the number of molecules, constant).

Molecules of different gases have the same average kinetic energy at the same temperature. This means that more massive molecules travel more slowly (because the average value of $\frac{1}{2}mv^2$ is unchanged). This is illustrated by the distributions in Figure 3.32.

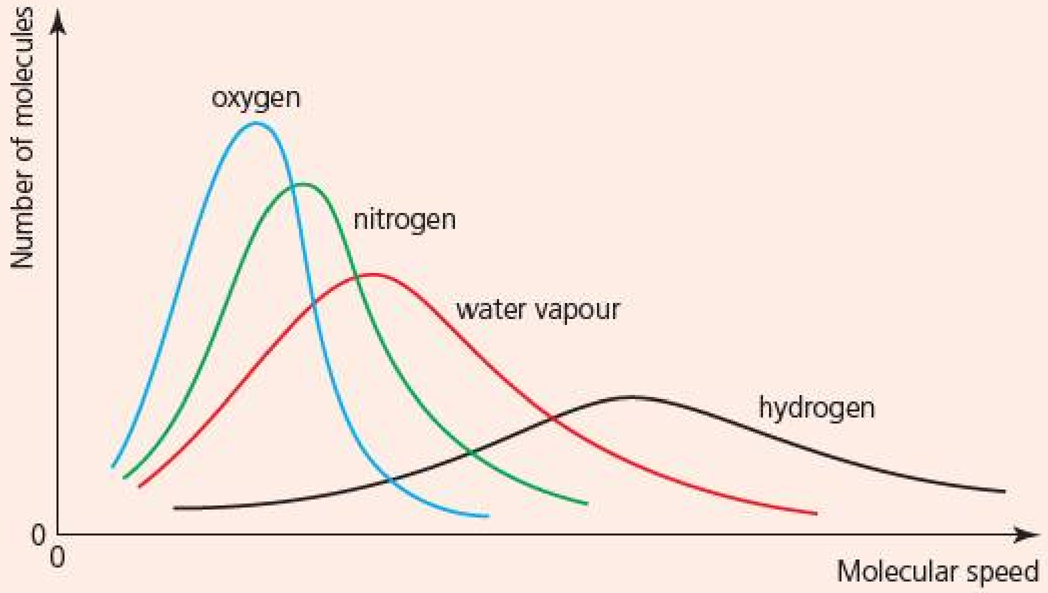


Figure 3.32 Distribution of molecular speeds in different gases at the same temperature

Question

1 a What is the average kinetic energy of hydrogen molecules at room temperature?
  b Calculate a typical speed for hydrogen molecules at 20 °C. (Mass of a hydrogen molecule = 3.3×10^{-27} kg.)
  c Explain why hydrogen gas diffuses more quickly than other gases at the same temperature.
  d If a cylinder of hydrogen gas at 20 °C was placed in a plane that later travelled at $500\,\mathrm{m\,s^{-1}}$, what would happen to the average speed of the molecules and the temperature?
  e What is the average kinetic energy of ideal gas molecules: **i** at 0 °C; **ii** on the surface of the Sun (5500 °C)?
  f At what temperature (°C) would ideal gas molecules have an average kinetic energy of 5×10^{-21} J?

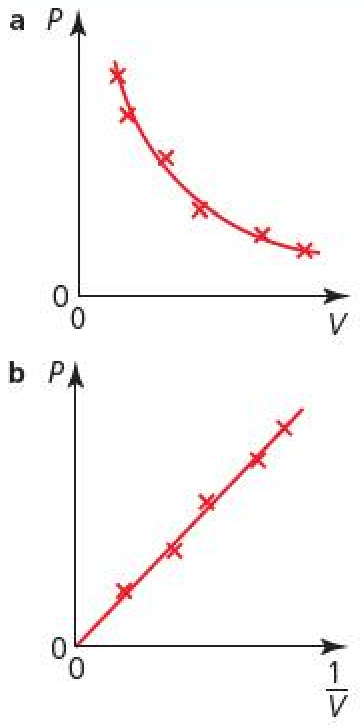


Figure 3.33 Two graphs showing that gas pressure is inversely proportional to volume

Macroscopic behaviour of real gases

Given a closed container with a gas inside it, there are four macroscopic, physical properties of the gas we can measure experimentally: mass, volume, temperature and pressure. If you have a fixed mass of gas, the other three properties will be interlinked – change one and you may change the other two. Three classic physics experiments showed that *all* gases, under most circumstances, follow the same simple patterns of behaviour (these are often called the 'gas laws').

1 For a fixed mass of gas at constant temperature, pressure is inversely proportional to volume (*Boyle's law*).

$$P \propto \frac{1}{V}$$

This relationship is represented in Figure 3.33a. The line on this graph is called an **isotherm** ('iso' means 'the same' – all the data points on an isotherm are for the gas at the same temperature). Figure 3.33b represents the same data re-drawn to produce a linear graph.

Figure 3.34 shows two sets of apparatus that can be used to demonstrate Boyle's law.

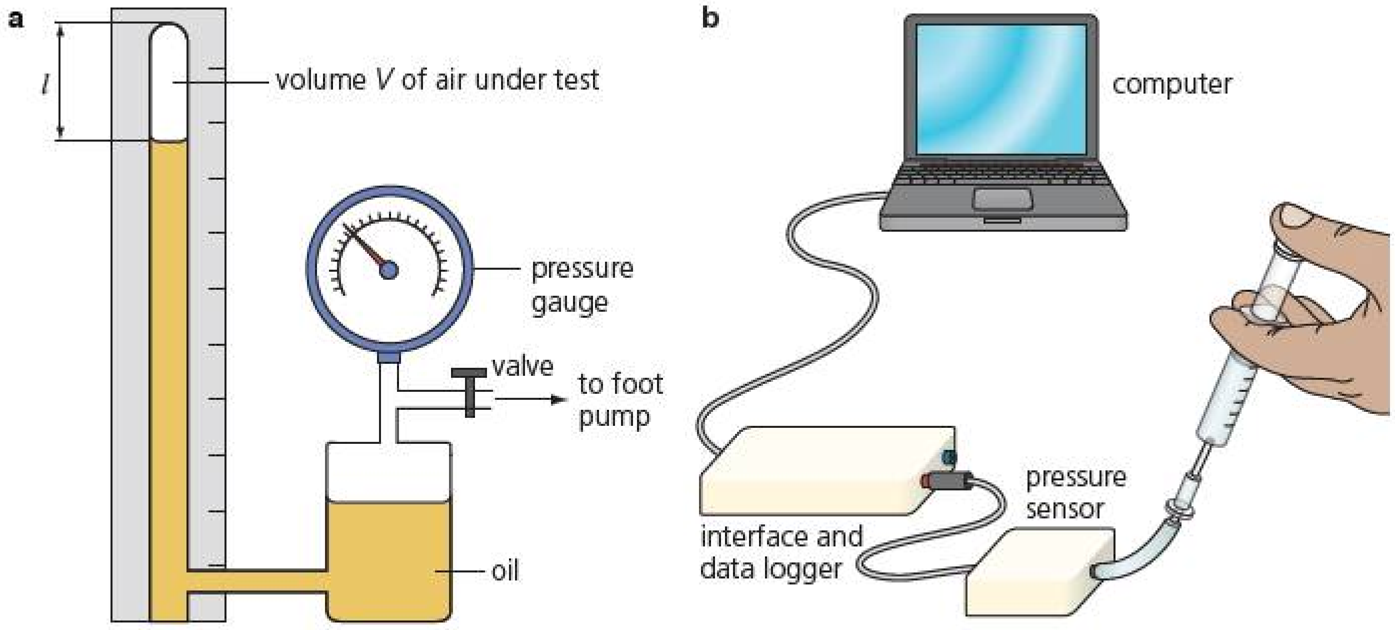


Figure 3.34 Two sets of apparatus that can be used to investigate how the pressure of a fixed mass of gas at constant temperature depends on its volume

2 For a fixed mass of gas at constant pressure, the volume is proportional to the temperature (K) (*Charles' law*).

$V \propto T$

This relationship is shown in Figure 3.35. Experimental results are usually taken within the range 0 °C to 100 °C (273 K to 373 K) and then the graph is extrapolated back for lower values.

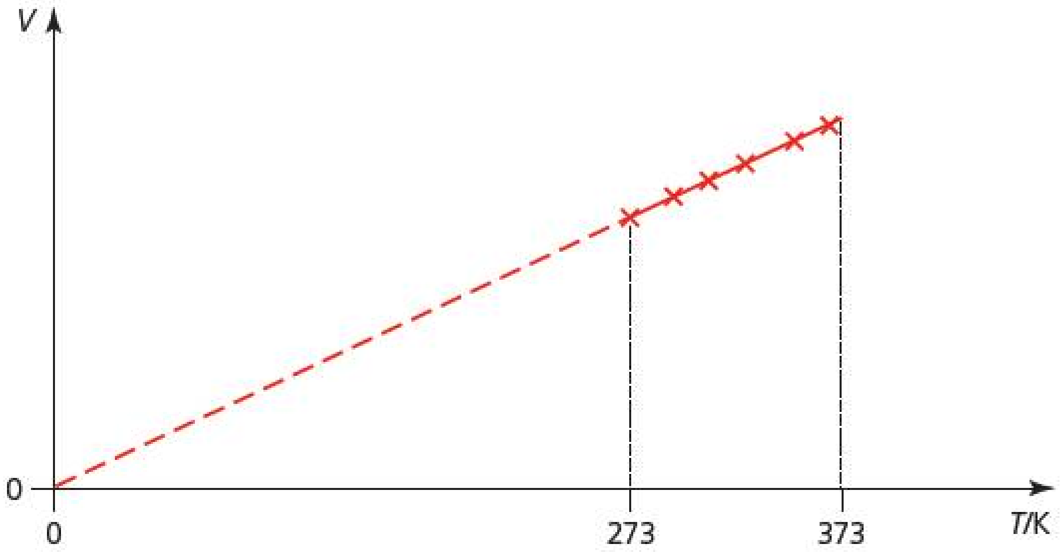


Figure 3.35 Gas volume is proportional to temperature (K)

3 For fixed mass of gas at constant volume, the pressure is proportional to the temperature (K) (the *pressure law*).

$P \propto T$

This relationship is shown in Figure 3.36. Experimental results are usually taken within the range 0 °C to 100 °C (273 K to 373 K) and then the graph is extrapolated back for lower values.

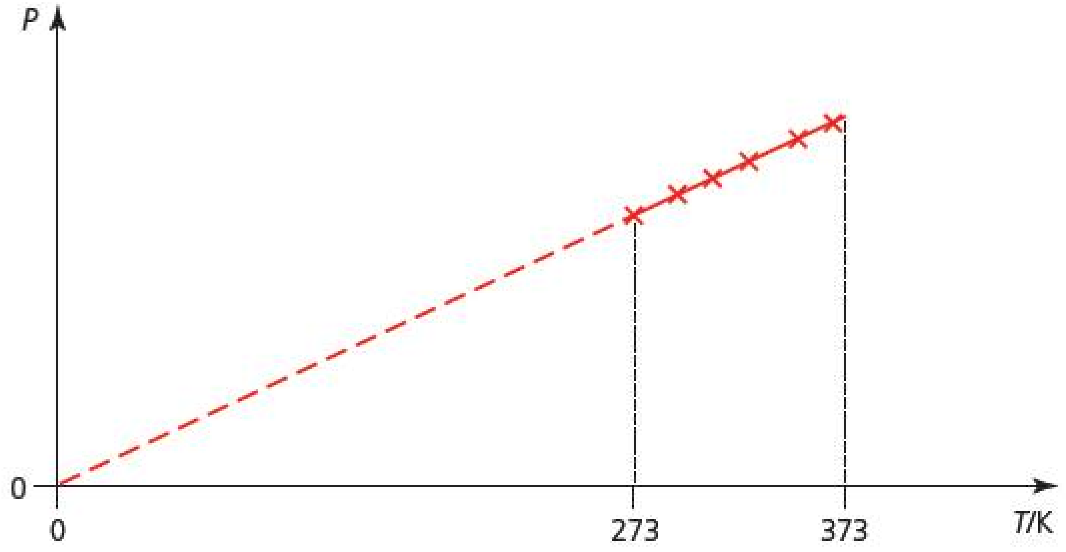


Figure 3.36 Gas pressure is proportional to temperature (K)

Figure 3.37 shows two different sets of apparatus that can be used to demonstrate the pressure law.

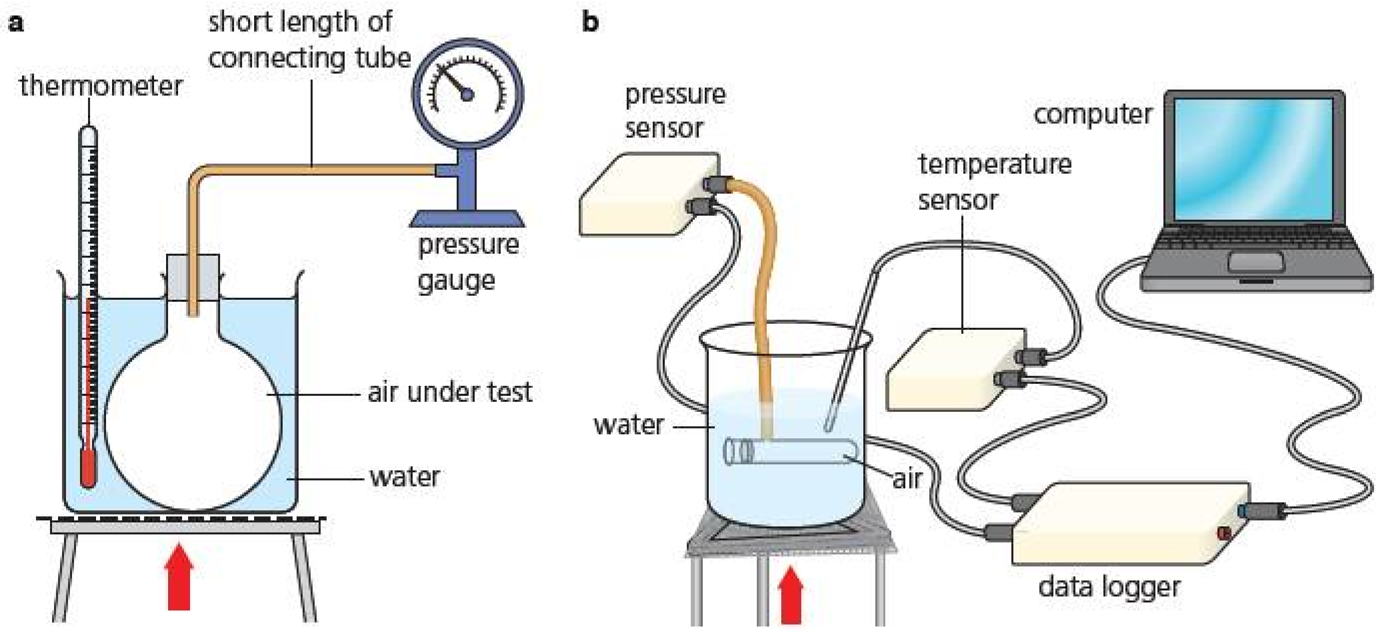


Figure 3.37 Two sets of apparatus that can be used to investigate how the pressure of a fixed mass of gas at constant volume depends on its temperature

Using kinetic theory to explain gas behaviour

3.2.12 Explain the macroscopic behaviour of an ideal gas in terms of a molecular model.

The kinetic theory of an ideal gas was developed to explain the macroscopic properties of real gases as listed above, and it was very successful – the simplified kinetic model of an *ideal* gas is very good at describing *real* gases. The following explanations are qualitative, but a detailed mathematical treatment of the kinetic theory of an ideal gas can also produce accurate quantitative predictions.

1 *Boyle's law* – As the volume of a gas is decreased, the molecules will hit the walls more frequently because they have less distance to travel between collisions. So, the smaller the volume, the higher the pressure.
2 *Charles' law* – When the temperature of a gas increases, the average molecular speed increases, so that the molecules collide with the walls more frequently and with more force. This results in a net outward force on the walls of the container, but, if any of the walls are moveable, the gas will expand, keeping the pressure on the inside the same as the pressure on the outside of that wall. So, the higher the temperature, the larger the volume (if the gas is able to expand freely).
3 *Pressure law* – When the temperature of a gas increases, the average molecular speed increases. The molecules collide with the walls more frequently and with more force. (There are more collisions every second with the walls.) So, the higher the temperature, the higher the pressure.

An ideal gas would follow the three 'gas laws' perfectly. The behaviour of real gases will be a close approximation to the gas laws, except under extreme circumstances, for example, at very high pressures or very low temperatures.

44 **a** Explain why the forces between molecules in a solid or liquid are much greater than the intermolecular forces in a gas.
b Can intermolecular forces between molecules ever be truly zero? (Look back at Figure 3.4.) Explain your answer.

45 **a** Explain in (microscopic) detail what happens to the molecules in a warm gas placed inside a colder container.
b What happens to the internal energy and temperature of the walls and the gas?

46 **a** Explain why the pressure in an ideal gas will become zero if it is cooled to −273 °C.
b If they are cooled enough, real gases will condense to liquids long before they get to −273 °C. Explain this observation with reference to molecular speeds and intermolecular forces.

47 Make a copy of Figures 3.33a and 3.33b and then add another isothermal line to each to represent the results that would be obtained with the same gas in the same apparatus, but at a higher temperature.

TOK Link: Models in physics

In order to explain the world around us, we need to reduce its complexities and represent our observations in simplified ways. By doing this we hope to get a better understanding, to learn from the past and to be able to make predictions about the future. Much of science is about making such 'models' and the kinetic model of an ideal gas (sometimes called the kinetic 'theory' of gases) is an excellent example of how modelling is used in science. (Everyday life is also complicated and we all cope with its many complexities by making simplifying assumptions about people and situations – models of our own reality.)

Models can have many different forms. A drawing can be a simple model, for example a scale map representing the location of a geographical feature, or a more stylized map representing a city's metro lines (which is not drawn to scale).

Ask people to give an example of a model, and they are most likely to describe scale models of cars, boats, etc. Physics, too, uses models in this way, for example an engineer may make a scale model of a bridge before its actual construction, in order to predict its possible behaviour when exposed to strong winds, earthquakes or other forces.

Models are also used to represent things that are much too large or too small to visualize, like the models of the solar system or atomic structure that are discussed elsewhere in this course. However, in such examples, there are important differences from more simple models, because in these cases, it is probable that there will be aspects of the 'system' being modelled that have not been seen or are not known or not understood. For example, great scientific insight (and imagination) was needed to develop a model for the structure of atoms that nobody had ever seen. In such cases we may well describe the model as a 'theory'. (The differences between facts, models, theories and 'laws' is another interesting discussion.)

Computer models (simulations) of physical processes have become very popular and useful, especially to model how complex systems change over short or long periods of time (such as climate models, discussed in Chapter 8).

One of the most common and useful kinds of modelling in physics is by the use of mathematics. Graphs and equations used to represent the behaviour of various systems are found in all branches of physics and for good reason. Verbal and written descriptions can be vague, difficult to understand or inconvenient to use. But mathematical models are precise and unambiguous, as well as being ideal for making predictions. The equations of motion (covered in Chapter 2) are good examples of mathematical models in physics.

Models are intentionally less complicated than the reality they are meant to represent. Simplifications and/or assumptions have to be made and these may be just a matter of convenience (too much detail is simply not needed), or may reflect an ignorance of important facts about the system being modelled. Or maybe the system is just too complicated to deal with without reducing its complexity in ways that, it is hoped, will not affect the usefulness of the model being developed. The kinetic model of a gas is a good example of this feature.

The usefulness of a model in being able to answer 'what if ….' type questions (that is, to make predictions) may be considered to be as important as whether the model is an accurate representation of reality. For example, the Earth-centred model of the universe developed in some ancient civilizations is now known to be wrong, but at the time it was useful in predicting seasons, eclipses, etc.

Question

1 Make a list of five different *models* that you have used in your physics lessons.

Additional Perspectives

Randomness

Throughout this topic there have been frequent uses of the word 'random' with respect to energy or motion. But what exactly does 'random' mean? The word has various uses throughout science and more generally, often with slight differences in meaning. For example, we might say that the result of throwing a six-sided die is random because we cannot predict what will happen, although we probably appreciate that there is a one-in-six chance of any particular number ending up on top. In this case, all outcomes should be equally likely. Another similar example could be if we were asked to 'pick a card at random' from a pack of 52. Sometimes we use the word random to suggest that something is unplanned, for example a tourist might walk randomly around the streets of a town.

Unpredictability is a key feature of random events and that certainly is a large part of what we mean when we say a gas molecule moves randomly. All possible directions of motion may be equally likely, but the same cannot be said for speeds. Some speeds are definitely more likely than others. For example, at room temperature, a molecular speed of $500\ \mathrm{m\,s^{-1}}$ is much more likely than one of $50\ \mathrm{m\,s^{-1}}$. Similarly, when we refer to random kinetic energies of molecules (in any of the three phases of matter) we mean that we cannot know or predict the energy of individual molecules, although some values are more likely than others. But there is a further meaning: we are suggesting that individual molecules behave independently and that their energies are disordered.

Perhaps surprisingly, in the kinetic theory the random behaviour of a very large number of individual molecules on the macroscopic scale leads to complete predictability in our everyday macroscopic world. Similar ideas occur in other areas of physics, notably in radioactive decay (Chapter 7), where the behaviour of an individual atom is unknowable, but the total activity of a radioactive source is predictable. Of course, insurance companies and casinos can make good profits by understanding the statistics of probability.

Questions

1 If ten coins were tossed and three came down as heads and seven came down as tails, it would not be too surprising. But if 100 were tossed and only 30 were heads it would be amazing, and if a thousand tosses produced 300 heads and 700 tails it would be almost completely unbelievable.
Explain these comments.

2 Explain why individual gas molecules colliding with the walls of their container produce a predictably constant pressure.

3 Choose five numbers between 1 and 10 at random. How did you make your choice? How could anyone tell from looking at the numbers that you have chosen, that they were really picked at random?

SUMMARY OF KNOWLEDGE

3.1 Thermal concepts

- All substances contain particles (molecules) that are vibrating and/or moving around. The molecules have random kinetic energy. The molecules in solids and liquids also have potential energy because there are electric forces between them. The total of all of these energies inside a substance is called its internal energy. (It should not be called thermal energy or heat.)
- When objects (or substances) are at different temperatures, energy will always flow from hotter to colder. This flow of energy is called thermal energy (sometimes called heat). If two or more objects are at the same temperature there will be no net flow of thermal energy between them and they are described as being in thermal equilibrium.
- The Kelvin (absolute) temperature scale has a zero at the point where (almost) all molecular motion has stopped. Calculations involving single temperatures should use the Kelvin scale, but calculations involving *changes* of temperature can also use the Celsius scale. $T/\mathrm{K} = t/^{\circ}\mathrm{C} + 273$.

- The mass of a substance is measured in kilograms, but the amount of a substance is a measure of how many particles it contains. Amount of substance is measured in moles. One mole is defined as the amount of a substance that contains the same number of particles as in exactly 12 g of carbon-12. (This number is Avogadro's constant, N_A, and equals $6.02 \times 10^{23}\,\text{mol}^{-1}$.) Molar mass is defined as the mass of a substance that contains one mole.

3.2 Thermal properties of matter

- The molecules in solids are held close together by strong forces and they are usually in regular patterns. The molecules vibrate about their mean positions.
- In liquids the molecules still vibrate, but sometimes the forces between them are overcome, allowing molecules to move around a little. The molecules are still almost as close together as in solids, but there is little or no regularity in their arrangement, which is constantly changing.
- In gases the molecules are much further apart than in solids and liquids, and the forces between them are very, very small (except when they collide). This results in all molecules moving independently in random directions with a range of different (usually fast) speeds.
- The thermal capacity of an object is defined as the amount of energy needed to raise its temperature by one kelvin, $C = Q/\Delta T$.
- The specific heat capacity of a certain substance is defined as the amount of energy needed to raise the temperature of one kilogram by one kelvin. $c = Q/m\Delta T$.
- It is common for gravitational potential energy or kinetic energy to be transferred into internal energy. The temperature rise can be calculated by equating the mechanical energy transferred to the increase of internal energy, $mc\Delta T$.
- When objects or substances at different temperatures are placed together in good thermal contact, the decrease in internal energy of one can be equated to the increase in internal energy of the other if they are isolated from their surroundings.
- Calculations involving thermal energy transfers usually have to assume that there are no transfers of energy to or from the surroundings.
- The change of phase from a solid to a liquid is called melting or fusion. The reverse is called freezing. Changing from a liquid to a gas is called boiling or evaporation. The reverse is called condensation.
- Evaporation can occur at any temperature (at which the substance is liquid) and it occurs only on the surface. Boiling occurs at a precise temperature throughout the liquid.
- Energy must be transferred when there is a change of phase because there is a difference in potential energy of the molecules. This energy is called latent heat of fusion, or vaporization. Whenever there is a change of phase there is no change in temperature because the average kinetic energy of the molecules does not change.
- The specific latent heat of a substance is defined as the amount of energy needed to change the phase of one kilogram of the substance at constant temperature. $L = Q/m$.
- The molecular description of a gas can be developed further into the 'kinetic model of an ideal gas'. The theory can then be used to explain the macroscopic properties of real gases. To do this, some important simplifying assumptions are made:
  - An 'ideal' gas is assumed to consist of a very large number of molecules that are identical and have negligible size.
  - The molecules move around randomly because there are negligible forces between them.
  - Collisions between molecules are elastic.
- When a gas is placed in a container we can measure the macroscopic physical properties of volume, temperature and pressure (as well as the amount or mass of the gas). Classic physics experiments investigated how these properties are interconnected.
- Pressure is defined as force/area. Gas pressure is caused by molecules colliding with the surfaces of the container. Each collision results in a tiny force on the surface. Collisions are so frequent that the average force on a given area (pressure) usually remains constant.
- Temperature is a measure of the average kinetic energy of the molecules in a substance.
- The kinetic model of an ideal gas uses ideas about molecular motion to correctly predict how changes in temperature and/or volume affect pressure. The model can also explain why the volume of a gas at constant pressure must increase if the temperature rises.

Examination questions – a selection

Paper 1 IB questions and IB style questions

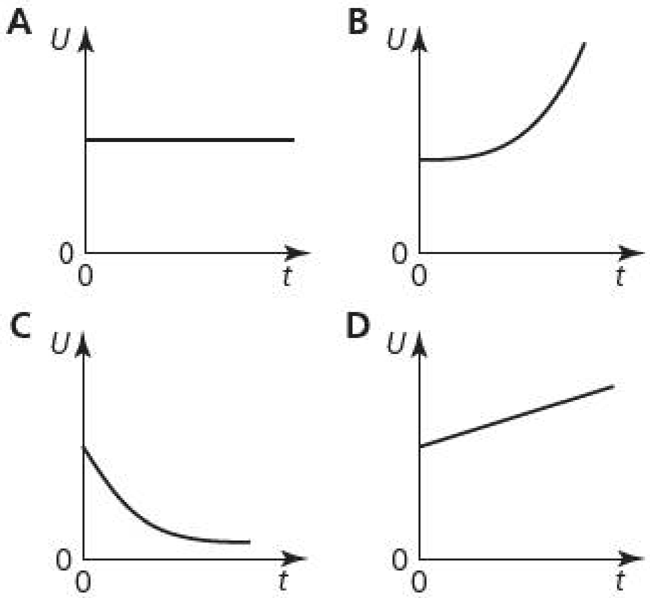
Q1 The temperature of an ideal gas is a measure of the
- **A** average momentum of the molecules.
- **B** average speed of the molecules.
- **C** average kinetic energy of the molecules.
- **D** average potential energy of the molecules.

Q2 If the volume of a fixed mass of an ideal gas is decreased at a constant temperature, the pressure of the gas increases. This is because
- **A** the molecules collide more frequently with each other.
- **B** the molecules collide more frequently with the walls of the container.
- **C** the molecules are moving at a higher average speed.
- **D** the molecules exert greater average forces on the walls during collisions.

Q3 The specific heat capacity of a substance is defined as the amount of thermal energy needed to raise the temperature of
- **A** the mass of the substance by 1 K.
- **B** the volume of the substance by 1 K.
- **C** unit volume of the substance by 1 K.
- **D** unit mass of the substance by 1 K.

Q4 Which of the following is an important assumption of the kinetic theory of ideal gases?
- **A** The forces between molecules are zero.
- **B** All the molecules travel with the same speed.
- **C** The molecular potential energies are constant.
- **D** The molecules have zero momentum.

Q5 A system consists of an ice cube placed in a cup of water. The system is thermally insulated from its surroundings. The water is originally at 20 °C. Which graph best shows the variation of total internal energy U of the system with time t?

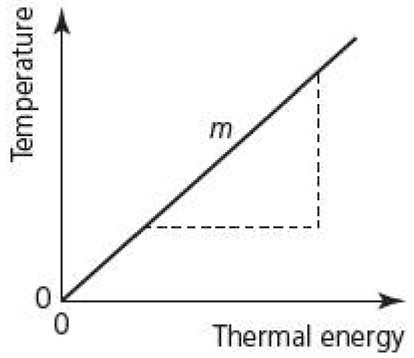


Standard Level Paper 1, Nov 10, Q9

Q6 Which of the following is the correct conversion of a temperature of 100 K to degrees Celsius?
- **A** −373 °C
- **B** −173 °C
- **C** 173 °C
- **D** 373 °C

Q7 Which of the following is a correct statement about the energy of the molecules in an ideal gas?
- **A** The molecules only have kinetic energy.
- **B** The molecules only have thermal energy.
- **C** The molecules only have potential energy.
- **D** The molecules have kinetic and potential energy.

Q8 A copper block of mass M was heated with an immersion heater. The graph shows how the temperature of the block was affected by the thermal energy supplied to it. The gradient of the line is m.



Which of the following expressions equals the specific heat capacity of the block?

A m **B** $\frac{1}{m}$ **C** mM **D** $\frac{1}{mM}$

Q9 Which of the following is *not* a factor which affects the rate of evaporation of a liquid?
- **A** specific heat capacity of the liquid
- **B** surface area of liquid
- **C** temperature of liquid
- **D** specific latent heat of vaporization of liquid

Q10 Two objects near each other are at the same temperature. Which of the following statements has to be true?
- **A** The objects have the same internal energy.
- **B** The objects have the same thermal capacity.
- **C** No thermal energy is exchanged between the objects.
- **D** The net thermal energy exchanged between the objects is zero.

Standard Level Paper 1, May 09 TZ2, Q9

Paper 2 IB questions and IB style questions

Q1 This question is about change of phase of a liquid and latent heat of vaporization.

a State the difference between evaporation and boiling with reference to

i temperature [1]

ii surface area of a liquid. [1]

b A liquid in a calorimeter is heated at its boiling point for a measured period of time. The following data are available.

Power rating of heater	= 15 W
Time for which liquid is heated at boiling point	$= 4.5 \times 10^2$ s
Mass of liquid boiled away	$= 1.8 \times 10^{-2}$ kg

Use the data to determine the specific latent heat of vaporization of the liquid. [1]

c State and explain **one** reason why the calculation in **b** will give a value of the specific latent heat of vaporization of the liquid that is greater than the true value. [2]

Standard Level Paper 2, May 10 TZ2, QA3

Q2 **a** The internal energy of a piece of copper is increased by heating.

i Explain what is meant, in this context, by internal energy and heating. [3]

ii The piece of copper has mass 0.25 kg. The increase in internal energy of the copper is 1.2×10^3 J and its increase in temperature is 20 K. Estimate the specific heat capacity of copper. [2]

b An ideal gas is kept in a cylinder by a piston that is free to move. The gas is heated such that its internal energy increases and the pressure remains constant. Use the molecular model of ideal gases to explain.

i the increase in internal energy. [1]

ii how the pressure remains constant. [3]

Standard Level Paper 2, May 09 TZ1, QB3 (Part 1)

4 Oscillations and waves

STARTING POINTS

- Objects can move in different ways. In translational motion an object starts in one place and moves somewhere else, for example, a student walking from one classroom to another, or a train moving from one station to the next (Chapter 2).
- An object can spin (rotate). For example, the Earth spins on its axis once every day and a wheel spins on a bicycle.
- Objects may move in circles. Circular motion was covered in Chapter 2.
- Objects can move backwards and forwards about the same place, with their displacement, velocity and acceleration continually changing in magnitude and direction. This is called an oscillation, although sometimes we may also refer to it as a vibration. This chapter is all about oscillations and oscillators (objects that oscillate).
- The relationships between displacement, velocity, acceleration and time are commonly represented using graphs (Chapter 2).
- Force = mass × acceleration or $F = ma$ (Newton's second law of motion), where F is the *resultant* force (Chapter 2). Forces and accelerations are vectors.

4.1 Kinematics of simple harmonic motion (SHM)

Examples of oscillations

4.1.1 Describe examples of oscillations.

There are repeating motions of many diverse kinds, both in nature and in the manufactured world in which we live. Examples of these oscillations include:

- tides on the ocean
- the movement of our legs when we walk
- a heart beating
- clock mechanisms
- atoms vibrating
- machinery and engines
- a guitar string playing a musical note
- our eardrums (when we hear a sound)
- electronic circuits that produce radio waves and microwaves.

And there are many, many more examples.

Oscillations can be very rapid and difficult to observe. In a school laboratory we usually begin the study of oscillations with experiments on very simple kinds of oscillator that are easy to observe and that oscillate at convenient rates, for example a pendulum and a mass on a spring (or between springs) (Figures 4.2 and 4.3).

Figure 4.1 Oscillations of a humming bird's wings

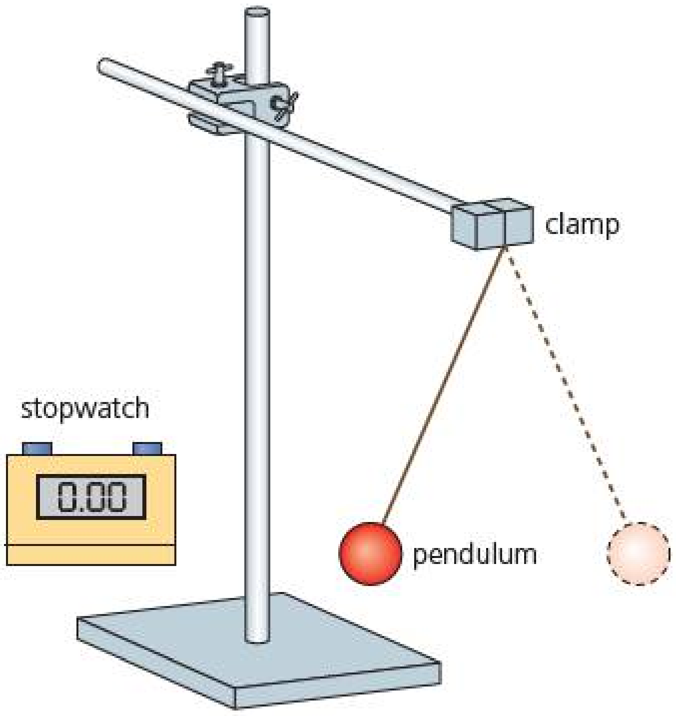


Figure 4.2 Investigating a pendulum

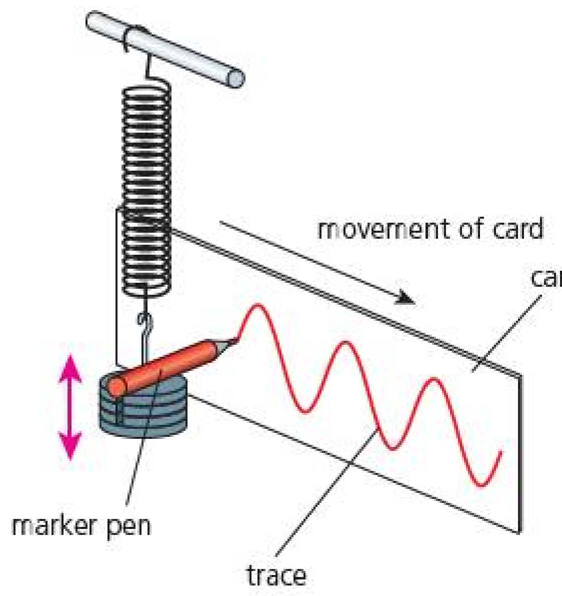


Figure 4.3 Investigating a mass oscillating on a spring

These oscillators behave in very similar ways to many other more complicated oscillators and we can use them as simple **models** that we can apply to other situations.

Describing oscillations quantitatively

4.1.2 Define the terms *displacement*, *amplitude*, *frequency*, *period* and *phase difference*.

Oscillations may occur whenever an object is displaced from its equilibrium position and it then experiences a '**restoring' force** pulling it back. The **equilibrium** (or mean) **position** is the position where there is no resultant force on the object. That is, the equilibrium position is the place where the object would stay if it had not been disturbed and the place to which it tends to return after the oscillations have stopped.

The **displacement**, x, of an oscillator is defined as the distance in a specified direction from its equilibrium position.

(This is similar to the more general definition of displacement given in Chapter 2.) The displacement varies continuously during an oscillation.

When describing an object which keeps repeating its motion, there are two very obvious questions to ask – how large are the oscillations and how quickly does it oscillate?

The **amplitude** of an oscillator is defined as its maximum displacement, that is, the distance from the equilibrium position to the furthest point of travel.

We use the symbol x_0 for amplitude. A greater amplitude means that there is more energy in the system.

The **(time) period**, T, of an oscillation is the time it takes for one complete oscillation.

(A complete oscillation is sometimes called a 'cycle'.)

Figure 4.4 shows these terms on a graph of the motion of an oscillator.

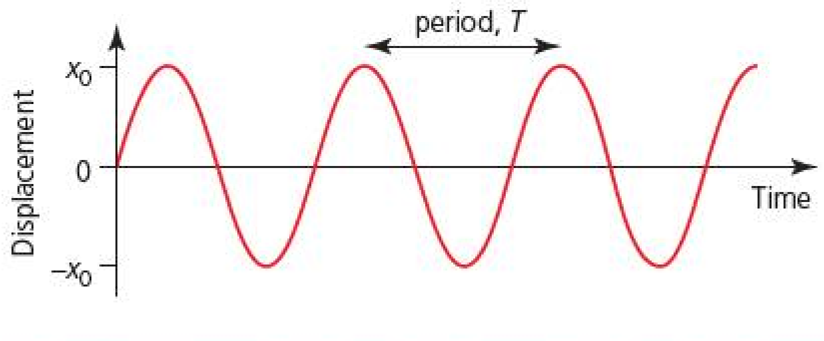


Figure 4.4 Graph of a simple oscillation

Most oscillations are quick, so that there are often a great many oscillations every second. Therefore, it is usually more convenient to refer to the number of oscillations every second (rather than the period).

The **frequency** of an oscillator is defined as the number of oscillations in unit time (usually every second). We use the symbol f for frequency.

When something is disturbed and then left to oscillate without further interference, it is said to oscillate at its **natural frequency**.

Of course, the frequency and time period of an oscillation are essentially the same piece of information expressed in two different ways. They are connected by the simple formula:

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

The concept of frequency is so useful that it has its own unit. A frequency of one oscillation per second is known as one **hertz, Hz**. The units kHz (10^3 Hz), MHz (10^6 Hz) and GHz (10^9 Hz) are also often used.

Worked example

1 A child on a swing went through exactly five complete oscillations in 10.4 s.
  a What was the period?
  b What was the frequency?

a $T = \frac{10.4}{5.00} = 2.08\,\text{s}$

b $f = \frac{1}{T} = \frac{1}{2.08} = 0.481\,\text{Hz}$

1 A pendulum was timed and found to have exactly 50 oscillations in 43.6 s.
  a What was its period?
  b What was its frequency?
  c How many oscillations will it undergo in exactly 5 minutes?
  d Explain why it is a good idea, in an experiment to find the period of a pendulum, to measure the total time for a large number of oscillations.

2 What is the period of a sound wave which has a frequency of 3.2 kHz?

3 Radio waves oscillate very quickly.
  a What is the frequency in Hz of a radio wave with a time period of 5×10^{-9} s?
  b Express the same frequency in MHz.

4 Look at the graph in Figure 4.5, which shows the motion of a mass oscillating on a spring. Determine:
  a the amplitude
  b the period
  c the displacement after 0.15 s
  d the displacement after 1.4 s.

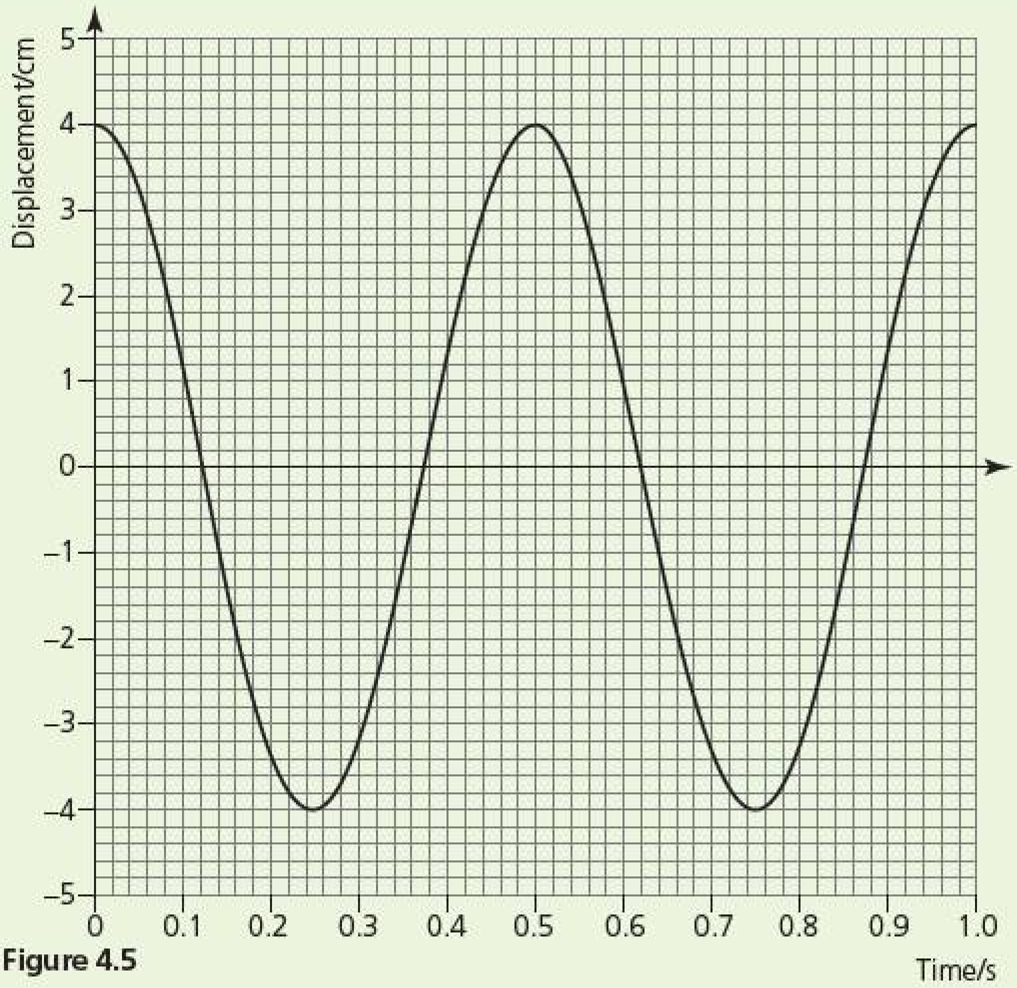


Figure 4.5

Additional Perspectives

Measurement of time

Our impression of the flow of time is an interesting philosophical and scientific issue based largely on repeating events, such as the daily spin of the Earth on its axis that gives us night and day, and the yearly orbit of the Earth around the Sun that gives us seasons. The measurement of time is also usually based on repeating events, such as the swing of a pendulum or the oscillations in a crystal.

Figure 4.6 A digital watch relies on electrical oscillations

Planet Earth's time system of years, days, hours, minutes and seconds would seem strange to aliens from another planet. In fact, the Earth spins once on its axis, not every 24 hours, but about every 23 hours and 56 minutes. Our 24-hour day is based on the time it takes for the Sun to return to its highest elevation as viewed from Earth and this is also affected, to a much lesser extent, by the Earth's daily movement in its orbit around the Sun.

The Earth's rotation is also decreasing at a rate of about 2 seconds every 100 000 years, so the length of the day is minutely but relentlessly increasing. This is mainly due to the gravitational interaction between the Earth and the Moon.

Questions

1 Before the invention of pendulum clocks, how did people measure short intervals of time?
2 Imagine you are watching a video of a pendulum swinging. Explain how you would know if the video was being played forwards or backwards. Discuss whether this helps us understand the meaning of time.

Oscillations and circular motion

There is a close connection between oscillations and circular motion. Indeed, viewed from the side, motion in a circle has the same pattern of movement as a simple oscillation.

Figure 4.7 shows a particle moving in a circle at constant speed. Point P is the projection of the particle's position onto the diameter of the circle.

As the particle moves in a circle, point P oscillates backwards and forwards along the diameter with the same frequency as the particle's circular motion, and with an amplitude equal to the radius, r, of the circle. One complete oscillation of point P can be considered as equivalent to the particle moving through an angle of 2π radians (or 360°).

If you were to plot a graph of displacement versus time for point P it would look exactly like the one in Figure 4.4, with the amplitude, x_0, being equal to r.

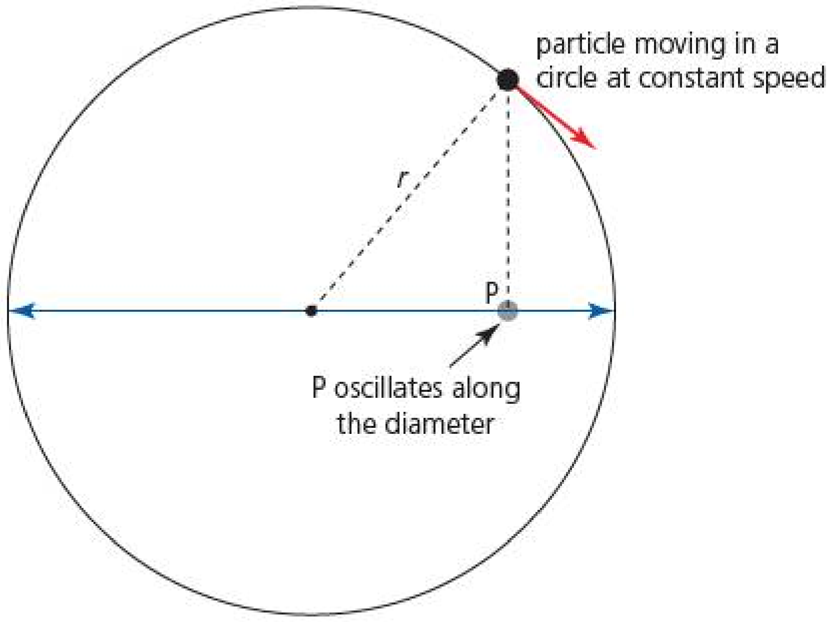


Figure 4.7 Comparing motion in a circle to an oscillation

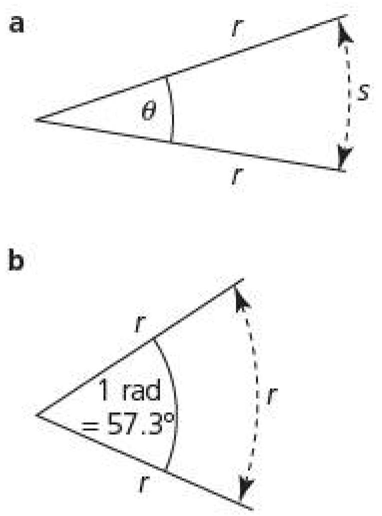


Figure 4.8 **a** Angle θ (in radians) is equal to s/r **b** one radian

Mathematical reminder: radians

- **Radians** are often a more convenient and useful way of measuring angles than degrees. In Figure 4.8a the angle θ (in radians) is equal to the distance s along the arc of the circle divided by the radius r.

$$\theta \text{ (in radians)} = \frac{s}{r}$$

- Note that angles expressed in radians are actually the ratio of two lengths, and therefore have no units. However we still usually refer to such angles using the term radians.
- One radian is the angle subtended at the centre of a circle by an arc of length equal to one radius ($\theta = r/r = 1$).
  As shown in Figure 4.8b, 1 radian = 57.3°. An angle of 360° = $2\pi r/r = 2\pi$ radians.
- For small angles (less than about 0.1 rad, or 6°) the shape shown in Figure 4.8a is similar to a right-angled triangle. This enables a useful approximation to be made:

$$\text{for small angles, } \theta \text{ (in radians)} \simeq \sin\theta \simeq \tan\theta$$

Angular frequency

A key feature in the description of translational motion is velocity, or 'change in displacement/change in time' (Δdisplacement/Δtime). Similarly, to describe circular motion and oscillations, we use 'change in angle/change in time' (Δangle/Δtime). This is called **angular frequency**, or sometimes 'angular speed'. It has the symbol ω and unit rad s^{-1}.

$$\text{angular frequency, } \omega = \frac{\Delta\text{angle}}{\Delta\text{time}}$$

$$\omega = \frac{\Delta\theta}{\Delta t}$$

Or, since in one oscillation, 2π radians are completed in period T,

$$\omega = \frac{2\pi}{T}$$

This equation is given in the IB *Physics data booklet*.

Also, since $T = \frac{1}{f}$,

$$\omega = 2\pi f$$

Time period, frequency and angular frequency are three concepts that represent the same information. For example, if an oscillator has a time period, T, of 0.2 s, we know that it has a frequency of 5.0 Hz (because $f = 1/T$) and an angular frequency of 31 rad s^{-1} (because $\omega = 2\pi/T$).

When information is presented to us, for example in a question, we are most likely to be given the period or the frequency of an oscillation, but for calculations the angular frequency, ω, is usually needed.

The angle, θ, passed through in time Δt, will be given by $\Delta\theta = \omega\Delta t$.

Worked example

2 If a pendulum undergoes exactly 20 oscillations in 34.6 s, calculate:
  a its frequency
  b its angular frequency.

a $f = \frac{1}{T} = \frac{1}{(34.6/20.0)} = 0.578\text{ Hz}$

b $\omega = 2\pi f = 2\pi(0.578) = 3.63\text{ rad s}^{-1}$

5 Using a slow-motion video replay of the movement of a bird's wings, it was calculated that they were beating at a frequency of 22 Hz. What was the time period of the oscillation and its angular frequency?

6 A car engine was recorded at 4300 rpm (revolutions per minute). What was its angular frequency?

7 The Earth spins once on its axis in 23 hours and 56 minutes.
a What is its angular frequency?
b Through what total angle will it rotate in 1 day (24 h), 3 hours and 24 minutes?

8 An object is oscillating with an angular frequency of 48.1 rad s^{-1}. What is its period?

9 Convert the following angles to radians:
a 180°
b 90.0°
c 45.0°
d 112°

Comparing oscillations

We often want to compare two (or more) similar oscillations that have the same frequency.

If oscillators are doing exactly the same thing at the same time, we say that they are **in phase**. They begin oscillations at the same time and reach their maximum displacements at the same times (Figure 4.9a).

If one oscillator is half an oscillation 'ahead' of (or 'behind') another, we say that they are **exactly out of phase** (Figure 4.9c).

We have seen that a complete oscillation can be represented by a time period, T, or movement through an angle of 2π radians (= 360°). So half an oscillation can be represented by a time, $T/2$, or movement through π radians (= 180°) and we say that oscillators that are exactly out of phase (half an oscillation, $T/2$) have a **phase difference** of π or 180°.

Oscillators that are exactly a quarter of an oscillation ($T/4$) out of phase have a phase difference of $\pi/2$ or 90° (Figure 4.9b).

A phase difference of 2π is the same a phase difference of zero, that is, the two oscillators are in phase.

Phase difference is defined as the angle (usually in radians) between two similar oscillations.

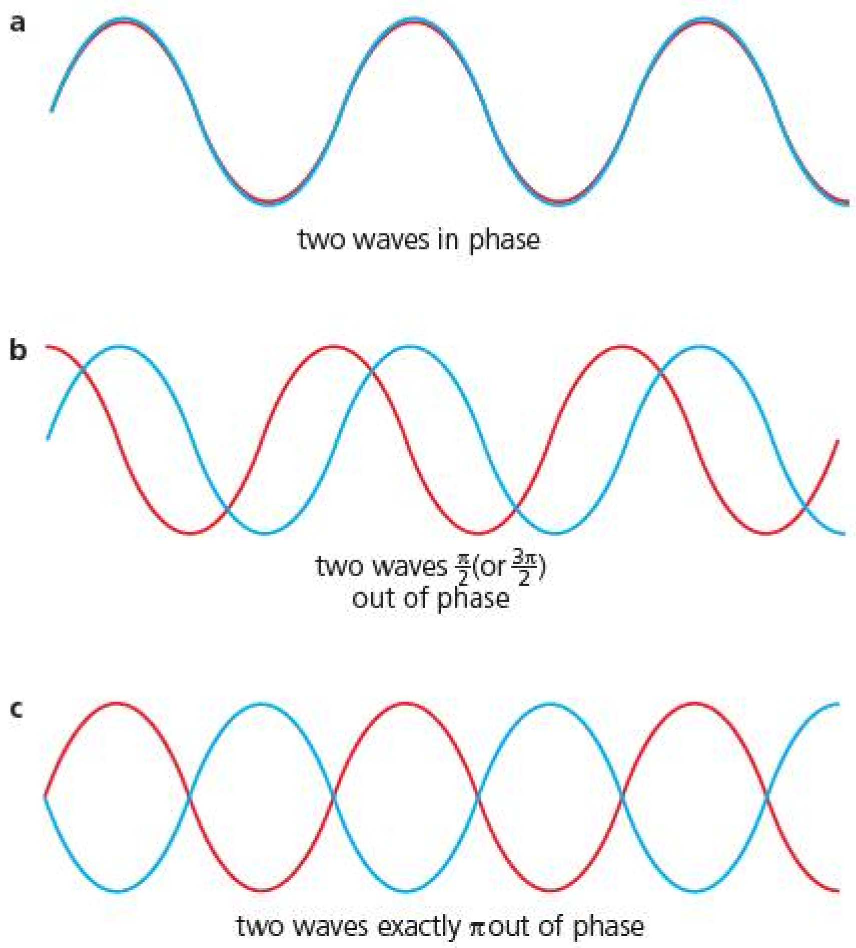


Figure 4.9 Phase differences between oscillations

Simple harmonic motion (SHM)

4.1.3 **Define** *simple harmonic motion (SHM)* and state the defining equation as $a = -\omega^2 x$.
4.1.4 **Solve** problems using the defining equation for SHM.

In advanced work on oscillators we may want to know, not only their amplitude and frequency, but also exactly where the object is at any time, and maybe also its velocity and/or its acceleration at that moment. However, the motion of real oscillators can be very complicated, so we need to start with a basic mathematical model in which we make some simplifying assumptions.

The simplest kind of oscillation will occur when an object, such as a mass attached to a spring (Figure 4.3), is displaced from its equilibrium position and experiences a restoring force, which causes it to vibrate. Provided the displacement is not too large, the size of the restoring force, F (and, therefore, the acceleration), is proportional to the displacement and in the opposite direction. Double the amplitude will result in twice the restoring force, so that the object will vibrate with the same *constant time period*. In general a constant time period may occur if:

$$F \propto -x$$

The negative sign in this equation indicates that the force is in the opposite direction to the displacement. In other words the force opposes the motion – it is a restoring force.

This kind of oscillation is called **simple harmonic motion (SHM)**. Although it is a theoretical model, many real-life oscillators approximate to this ideal model of SHM. SHM is defined in terms of accelerations, but since acceleration is proportional to force, the relationship has the same form.

Simple harmonic motion (SHM) is defined as an oscillation in which the acceleration, a, of a body is proportional to its displacement, x, from its equilibrium position and in the opposite direction:

$$a \propto -x$$

That is, the acceleration is always directed towards the equilibrium position, or, in other words, the object decelerates as it moves away from the equilibrium position.

We can rewrite the equations as:

$$a = -\text{constant} \times x$$

The constant must involve frequency because the magnitude of the acceleration is greater if the frequency is higher. The constant can be shown to be ω^2. So, the defining equation of SHM can be written as:

$$a = -\omega^2 x$$

This important equation is *not* in the IB *Physics data booklet*.

Remember that

$$\omega = \frac{2\pi}{T} = 2\pi f$$

This equation allows us to calculate the instantaneous acceleration (the acceleration at a given moment) of an oscillator of known frequency at any given displacement.

Worked example

3 A mass oscillates between two springs with a frequency of 1.4 Hz.
- **a** What is its angular frequency?
- **b** What is its acceleration when:
  - **i** its displacement is 1.0 cm
  - **ii** its displacement is 4.0 cm
  - **iii** it passes through its equilibrium position?

a $\omega = 2\pi f = 2\pi(1.4) = 8.8\,\text{rad s}^{-1}$

b i $a = -\omega^2 x = -(8.8)^2 \times 0.010 = -0.77\,\text{m s}^{-2}$

ii $a = -\omega^2 x = -(8.8)^2 \times 0.040 = -3.1\,\text{m s}^{-2}$

iii $a = -\omega^2 x = -(8.8)^2 \times 0.0 = 0.0\,\text{m s}^{-2}$

10 A pendulum has a period of 2.34 s. How far must it be displaced in order that its acceleration is $1.00\,\text{m s}^{-2}$?

11 During SHM a mass moves with an acceleration of $3.4\,\text{m s}^{-2}$ when its displacement is 4.0 cm. Calculate:
a its angular frequency
b its period.

12 A mass oscillating on a spring performs exactly 20 oscillations in 15.8 s.
a What is its acceleration when it is displaced 62.3 mm from its equilibrium position?
b What assumption did you make?

Solutions to the SHM equation ($a = -\omega^2 x$)

4.1.5 **Apply** the equations $v = v_0 \sin \omega t$, $v = v_0 \cos \omega t$, $v = \pm\, \omega \sqrt{(x_0^2 - x^2)}$, $x = x_0 \cos \omega t$ and $x = x_0 \sin \omega t$ as solutions to the defining equation for SHM.

4.1.6 **Solve** problems, both graphically and by calculation, for acceleration, velocity and displacement during SHM.

There are still many things about an oscillator which we cannot determine directly from the equation defining SHM, $a = -\omega^2 x$. For example, what are the values of displacement and velocity at any given time? To answer such questions we need either accurate graphs or mathematical '*solutions*' to the SHM equation. (Here, 'solutions' means either graphs or equations that will present the same information in more useful forms.)

Graphs of simple harmonic motion

Using data-gathering sensors connected to a computer, as shown in Figure 4.10, it is relatively easy to produce displacement–time graphs for a variety of oscillators.

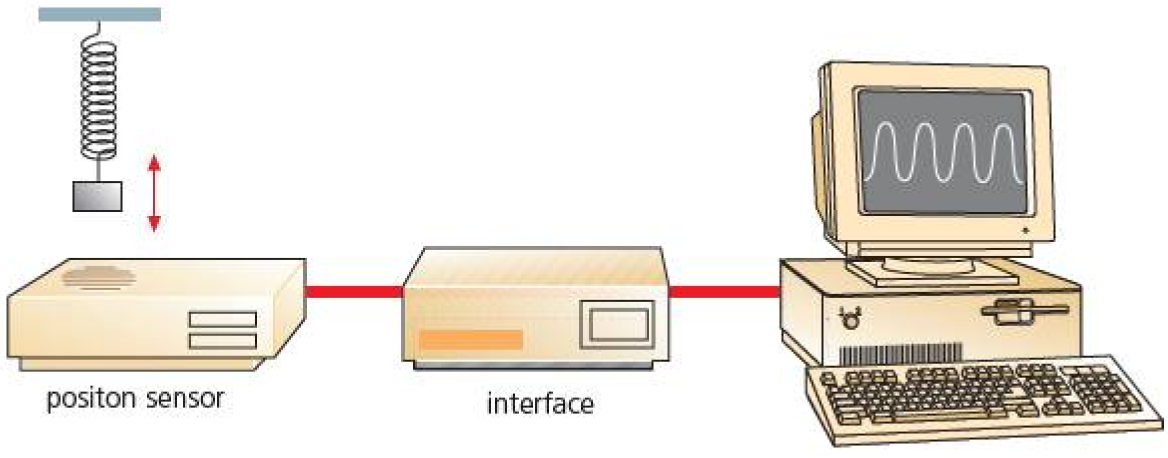


Figure 4.10 Collecting data on the oscillations of a mass on a spring using a sensor and data logger

The graph in Figure 4.11 shows the variation in displacement, x, with time for our idealized model of a particle moving with simple harmonic motion. Here the particle has zero displacement at the start of the timing, and has an amplitude of x_0. The graph is similar to the one in Figure 4.4.

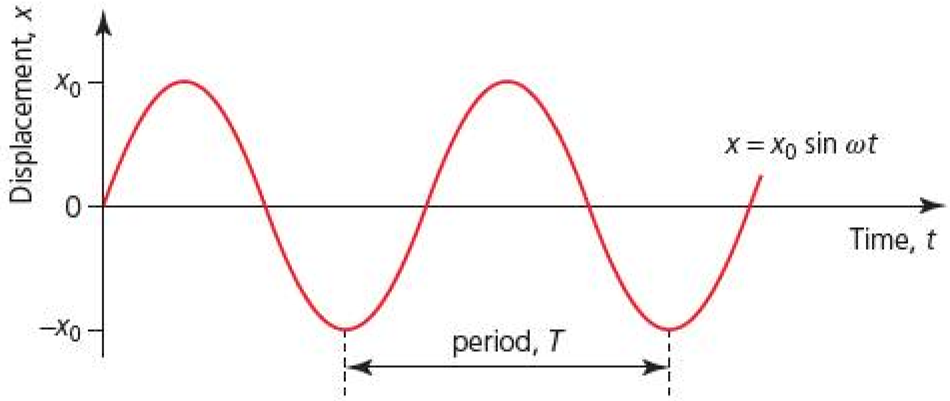


Figure 4.11 Displacement–time graph for simple harmonic motion (SHM), represented by a sine wave. Timing was started when the particle had zero displacement

The graph has a simple *sinusoidal* shape and it can be described mathematically by an equation using a sine:

$$x = x_0 \sin \theta$$

Since $\theta = \omega t$, the angle passed through in time t, this can be stated more usefully:

$$x = x_0 \sin \omega t$$

The velocity at a particular time or displacement can be found from the gradient of the displacement–time graph at that point (velocity, v = change in displacement/change in time).

Similarly, the acceleration at a given time or displacement is the gradient of the velocity graph (acceleration, $a = \Delta v/\Delta t$).

So the data from the displacement–time graph can be processed to calculate values for velocity and acceleration, and the corresponding graphs can be drawn, as shown in Figure 4.12.

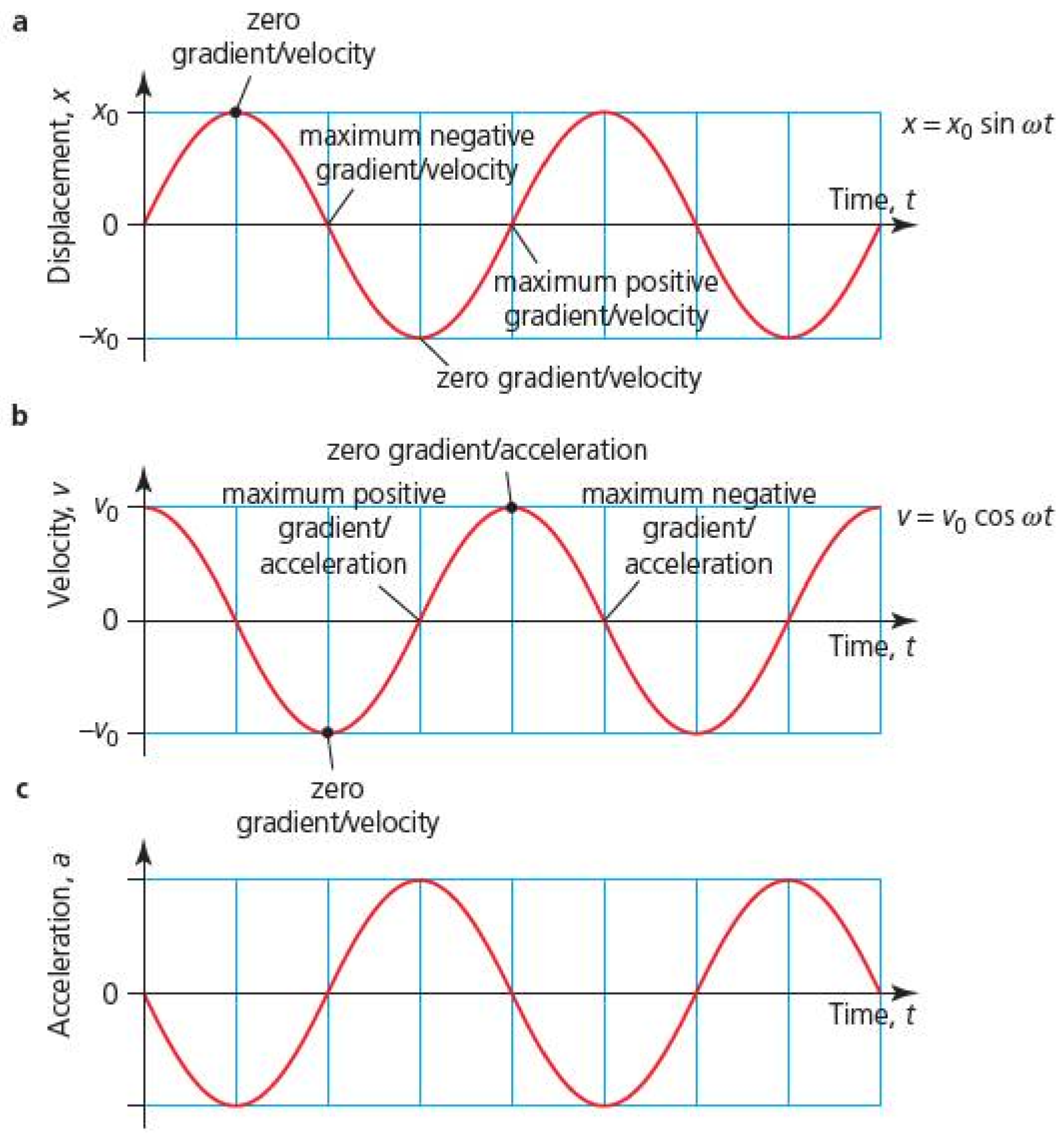


Figure 4.12 Displacement, velocity and acceleration graphs for simple harmonic motion (SHM), with timing starting at displacement $x = 0$: **a** displacement–time; **b** velocity–time; **c** acceleration–time

We see that the velocity has its maximum value, v_0, when the displacement x is zero, and the velocity is zero when the displacement is at its maximum, x_0. In other words, the velocity graph is $\pi/2$ out of phase with the displacement graph. This velocity graph can be represented by the following equation:

$$v = v_0 \cos \omega t$$

We can see from Figure 4.12 that the acceleration has its maximum value when the velocity is zero and the displacement is greatest. This is to be expected – when the displacement is greatest, the restoring force, acting in the opposite direction, is greatest. In terms of phase difference, the acceleration is $\pi/2$ out of phase with the velocity graph and π out of phase with the displacement graph.

Figure 4.13 shows all three graphs so that they can be easily compared. (Note that the amplitudes of the three graphs are arbitrary; they are not connected.)

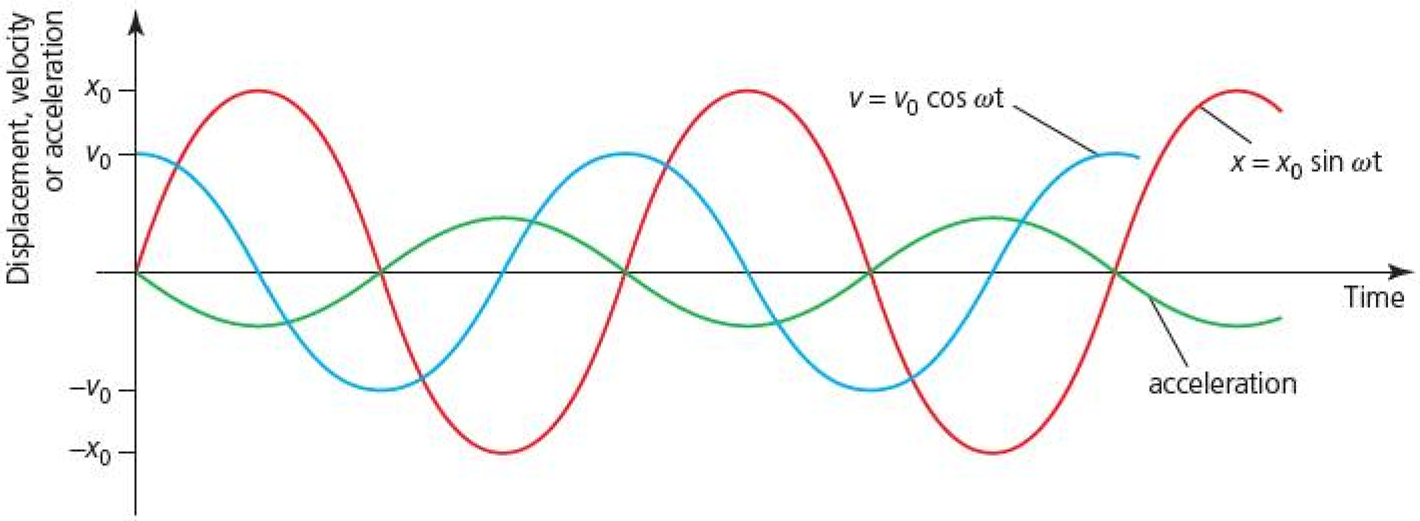


Figure 4.13 Comparing displacement, velocity and acceleration for SHM, with timing starting at displacement $x = 0$

Students are recommended to make use of a computer simulation of an object oscillating with SHM, combined with the associated graphical representations.

Solutions to the SHM equation

We have seen that the following equations show how displacement and velocity vary with time for SHM with zero displacement at the start of the timing ($x = 0$ when $t = 0$). We can say that they are *solutions to* the SHM equation ($a = -\omega^2 x$).

$$x = x_0 \sin \omega t$$
$$v = v_0 \cos \omega t$$

These equations are given in the IB *Physics data booklet*.

We could choose to start timing the SHM when the particle is at its maximum displacement from the equilibrium position during the oscillation, that is $x = x_0$ when $t = 0$. The graphs shown in Figure 4.13 would then appear as shown in Figure 4.14.

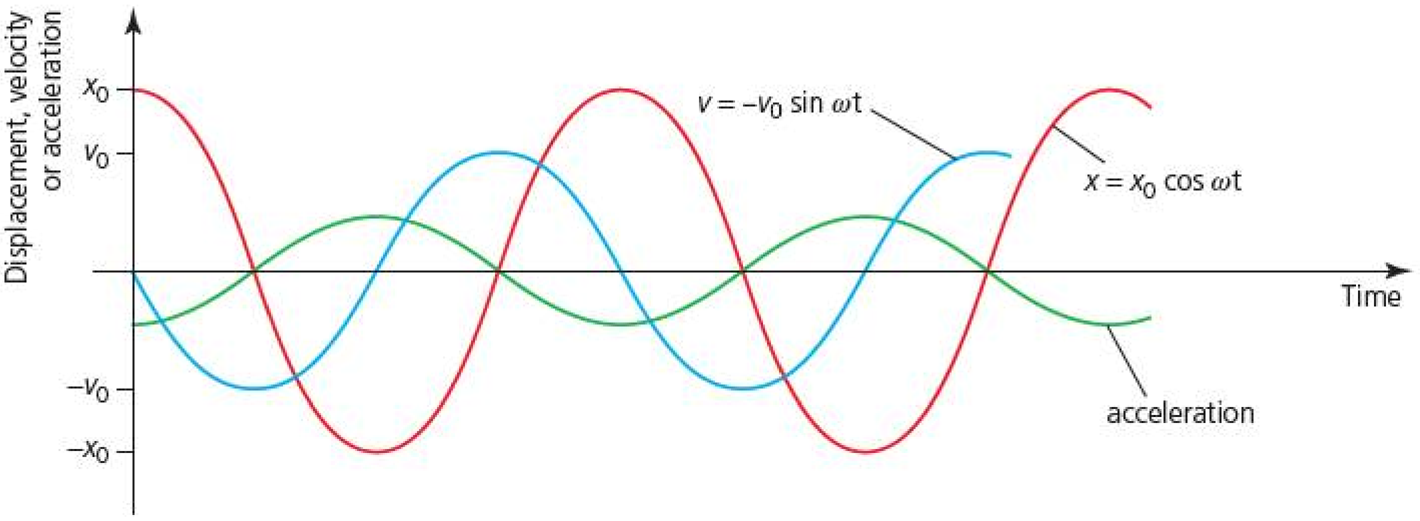


Figure 4.14 SHM with the particle at its maximum displacement at the start of timing, that is $x = x_0$ when $t = 0$

In this case, the solutions to the SHM equation are as follows:

$$x = x_0 \cos \omega t$$
$$v = -v_0 \sin \omega t$$

These equations are given in the IB *Physics data booklet*.

It can also be shown that the velocity at any known displacement can be calculated from the following equation:

$$v = \pm\omega\sqrt{(x_0^2 - x^2)}$$

This equation is also listed in the IB *Physics data booklet*.

This equation shows that the maximum velocity, v_0, will occur when the particle is moving through its equilibrium position (zero displacement, $x = 0$), so that

$$v_0 = \pm\omega x_0$$

Worked example

4 An oscillating mass is set in motion with SHM. It is at its maximum displacement of 12 cm when a stopwatch is started, and its period of oscillation is 2.4 s. Calculate:
a the displacement after 3.3 s
b its maximum speed
c its speed after 5.6 s
d its speed when its displacement is 8.8 cm.

a $x = x_0 \cos \omega t$ and $\omega = 2\pi/T$, with $x_0 = 0.12$ m, $T = 2.4$ s and $t = 3.3$ s

$$x = 0.12 \times \cos\left(\frac{2\pi}{2.4} \times 3.3\right) = 0.12 \times -0.704 = -0.084\,\text{m}$$

b $v_0 = \omega x_0 = \left(\frac{2\pi}{2.4}\right) \times 0.12 = 0.31\,\text{m s}^{-1}$

c $v = -v_0 \sin \omega t$ with $v_0 = 0.31\,\text{m s}^{-1}$ and $t = 5.6$ s

$$v = 0.31 \times \sin\left(\frac{2\pi}{2.4} \times 5.6\right) = 0.27\,\text{m s}^{-1}$$

d $v = \pm\omega\sqrt{(x_0^2 - x^2)}$, with $x = 0.088$ m

$$v = \left(\frac{2\pi}{2.4}\right) \times \sqrt{(0.12^2 - 0.088^2)} = 0.21\,\text{m s}^{-1}$$

13 A mass is oscillating between two springs with a frequency of 1.5 Hz and amplitude of 3.7 cm. It has a speed of 34 cm s^{-1} as it passes through its equilibrium position and a stopwatch is started. What are its displacement and velocity 1.8 s later?

14 An object of mass 45 g undergoes SHM with a frequency of 12 Hz and an amplitude of 3.1 mm.
a What is its maximum speed and kinetic energy?
b What is the object's displacement 120 ms after it is released from its maximum displacement?

15 A mass is oscillating with SHM with an amplitude of 3.84 cm. Its displacement is 2.76 cm at 0.0217 s after it is released from its maximum displacement. Calculate a possible value for its frequency.

16 A simple harmonic oscillator has a time period of 0.84 s and its speed is 0.53 m s^{-1} as it passes through its mean (equilibrium) position.
a What is its speed exactly 2.0 s later?
b If the amplitude of the oscillation is 8.9 cm, what was the displacement after 2.0 s?

17 The water level in a harbour rises and falls with the tides, with 12 h 32 min for a complete cycle. The high tide level is 8.20 m above the low tide level, which occurred at 4.10 am. If the tides rise and fall with SHM, what will be the level of the water at 6.00 am?

18 What does the area under a velocity–time graph of an oscillation represent?

4.2 Energy changes during simple harmonic motion

4.2.1 **Describe** the interchange between kinetic energy and potential energy during SHM.

4.2.2 **Apply** the expressions $E_K = \frac{1}{2}m\omega^2(x_0^2 - x^2)$ for the kinetic energy of a particle undergoing SHM, $E_T = \frac{1}{2}m\omega^2 X_0^2$ for the total energy and $E_P = \frac{1}{2}m\omega^2 x^2$ for the potential energy.

4.2.3 **Solve** problems, both graphically and by calculation, involving energy changes during SHM.

All mechanical oscillations involve a continual interchange between kinetic energy and some form of potential energy. For example, the potential energy involved with a pendulum swinging is gravitational, whereas for a mass oscillating horizontally between springs the potential energy is stored in the form of elastic strain energy in the spring.

We know that kinetic energy, $E_K = \frac{1}{2}mv^2$. Combining this with the equation for v given on page 129 leads to the following equation, which shows how KE varies with displacement during SHM.

$$E_K = \frac{1}{2}m\omega^2(x_0^2 - x^2)$$

This equation is given in the IB *Physics data booklet*.

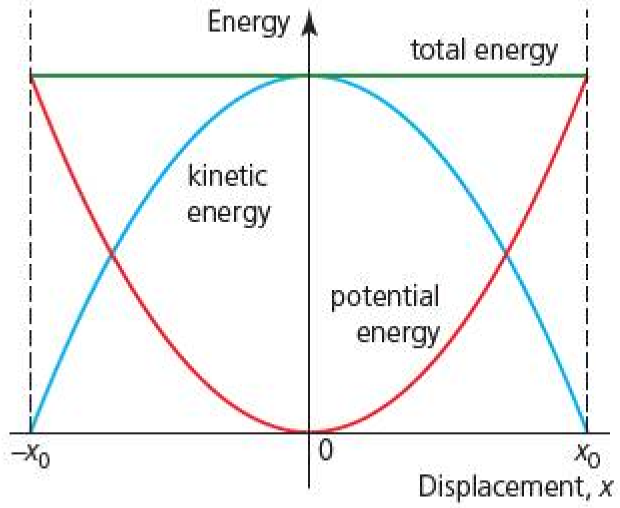


Figure 4.15 Variation of energies of a simple harmonic oscillator with displacement

This relationship is represented in Figure 4.15. Also included on the same graph are the lines for total energy and for potential energy.

The total energy is equal to the maximum value of KE – the value of KE with zero displacement and zero potential energy. That is,

$$E_T = E_{K(max)} = \frac{1}{2}m\omega^2 x_0^2$$

This equation is given in the IB *Physics data booklet*.

In a perfect simple harmonic oscillator the energy is constant and equals PE + KE. And, since $E_T = E_P + E_K$,

$$E_P = \frac{1}{2}m\omega^2 x^2$$

The equation shows an important point: the energy of an oscillation is proportional to its amplitude *squared*. For example, if pendulum A swings with three times the amplitude of an identical pendulum B, then pendulum A has nine (3^2) times the energy of pendulum B.

Worked example

5 A mass of 1250 g oscillates with a period of 0.56 s and an amplitude of 32 cm.

a What is its total energy?

b What is its potential energy when its displacement is 12 cm?

c What is its kinetic energy when displaced 12 cm?

d What is its kinetic energy when it has a displacement of 4.3 cm?

a $E_T = \frac{1}{2}m\omega^2 x_0^2$

$= 0.5 \times 1.25 \times \left(\frac{2\pi}{0.56}\right)^2 \times 0.32^2$

$= 8.1\,\text{J}$

b $E_P = \frac{1}{2}m\omega^2 x^2$

$= 0.5 \times 1.25 \times \left(\frac{2\pi}{0.56}\right)^2 \times 0.12^2$

$= 1.13\,\text{J}$

c $E_T = E_P + E_K$
$8.1 = 1.13 + E_K$
$E_K = 7.0\,\text{J}$

d $E_K = \frac{1}{2}m\omega^2(x_0^2 - x^2)$
$= 0.5 \times 1.25 \times \left(\frac{2\pi}{0.56}\right)^2 \times (0.32^2 - 0.043^2)$
$= 7.9\,\text{J}$

19 A large pendulum of mass 2.6 kg has a length, l, of 4.63 m.
- **a** Use the formula $T = 2\pi\sqrt{\frac{l}{g}}$ to determine its period, T.
- **b** Explain why a pendulum of twice the mass has the same period (if it has the same length).
- **c** The motion of a pendulum is a close approximation to SHM. Explain why the equation for the time period of an oscillator (like the pendulum) does not include the amplitude.
- **d** What is the angular frequency of this pendulum?
- **e** What is the oscillator's total energy if it has an amplitude of 1.2 m?
- **f** What is its speed as it passes through the equilibrium position?

20 A 435 g mass is oscillating horizontally between two springs with a frequency of 0.849 Hz. If its total energy is 4.28 J, what are its amplitude and its maximum speed?

21 An oscillator of mass 786 g has a total energy of 2.4 J.
- **a** Calculate its period if it is moving with an amplitude of 23 cm.
- **b** What is its speed when it has a displacement of 10 cm?

22 Real oscillators spread energy into the surroundings. Sketch a graph to show how potential and kinetic energy might change with time over several oscillations of a real oscillator.

4.3 Forced oscillations and resonance

Damping

4.3.1 **State** what is meant by damping.

The motions of all objects have frictional forces of one kind or another acting against them. Frictional forces always act in the opposite direction to the instantaneous motion of an oscillator, and result in a reduction of speed and the loss of kinetic energy (and consequently potential energy).

Therefore, as with all other mechanical systems, useful energy is spread from the oscillator into the surroundings in the form of thermal energy and sound. We say that the energy has been *dissipated*.

Consequently, the oscillator will travel at slower and slower speeds and its successive amplitudes will decrease in size. This effect is called **damping**.

Damping is the dissipation of energy of an oscillator due to resistive forces.

It is usual for the frequency (and time period) to remain approximately the same during damping because, although the displacements are less, the speeds and accelerations also decrease, as shown in Figure 4.16.

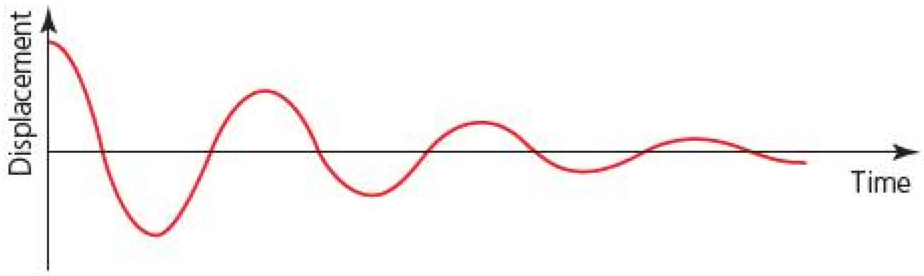


Figure 4.16 Decreasing amplitude (at constant frequency) of a damped oscillation

Oscillations are much more common than you may think at first. For example, most everyday objects make a noise if we hit them. This is because they oscillate and disturb the air next to them. The oscillations spread through the air to our ears – a sensation we call sound. The oscillations stop very quickly because of damping.

Examples of damped oscillations

4.3.2 **Describe** examples of damped oscillations.

The amount (*degree*) of damping in oscillating systems can be very different, as shown in Figure 4.17. Many oscillations are **heavily damped** because of considerable resistive frictional forces and the oscillations quickly reduce in amplitude. Conversely, occasionally damping can be very light and the oscillator may continue to oscillate, taking some time to dissipate its energy.

A pendulum and a mass oscillating on a spring are good examples of **lightly damped** systems. If the mass on the spring was placed in a beaker of oil (instead of air) it would then become heavily damped.

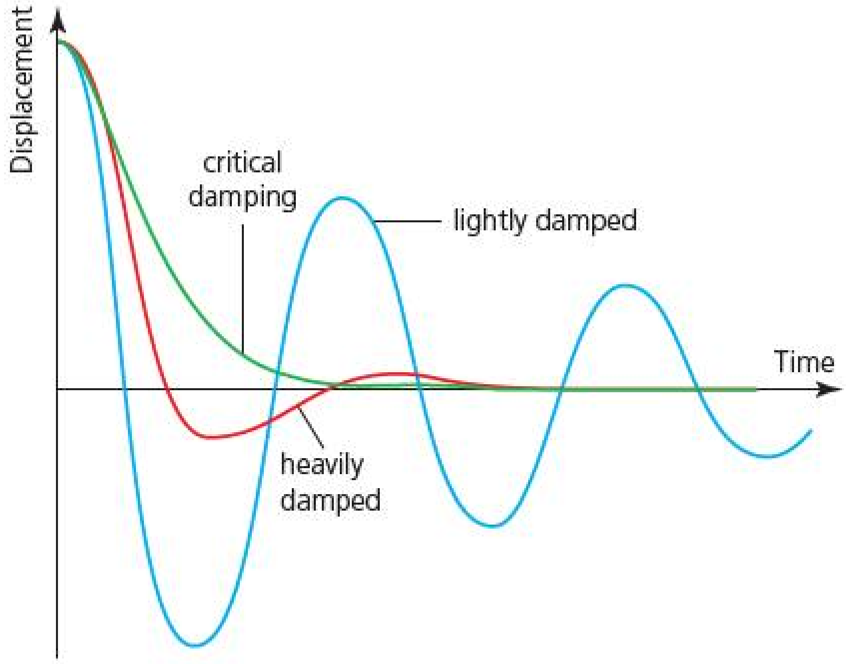


Figure 4.17 Heavy, light and critical damping

Figure 4.18 A shock absorber on a car is designed to dissipate energy

Oscillations are often unhelpful or destructive and we may wish to stop them as soon as possible. If an oscillation is stopped by resistive forces such that it settles relatively quickly back into its equilibrium position (without ever passing through it), the process is described as **critical damping**. A car's suspension (see Figure 4.18) is an example of this kind of damping, as are self-closing doors.

■ Additional Perspectives

Taipei 101

At the end of 2004 one of the world's tallest buildings, Taipei 101, was completed in Taiwan. Its height, measured to the top of its spire, is 509 m and, at the time of completion, it was the first building in the world to be more than half a kilometre tall.

Taiwan is in a region of the Earth which suffers from the effects of earthquakes and typhoons, so engineers had to be sure that their design could withstand the worst that could happen. A major design feature is a 730 tonne steel pendulum suspended inside the building from the 92nd to 88th floors. This is the major part of a system designed to dissipate energy in the event of strong oscillations produced by typhoons or earthquakes. It is the largest 'damper' in the world.

Without damping, wave energy from the earthquake could be transferred continually during the earthquake, leading to an increasing amplitude of vibration in the building, which could be destructive.

Figure 4.19 The Taipei 101 damper is designed to absorb energy in the event of an earthquake

Questions

1 Estimate how much energy could be safely transferred to this pendulum.

2 Very tall buildings usually give engineers many design problems and they are very expensive to build, so why do we build such tall structures? Suggest what kind of limitations there might be on the height of a building.

Forced oscillations

4.3.3 **State** what is meant by natural frequency of vibration and forced oscillations.

Figure 4.20 How can we increase the amplitude of a swing?

We are surrounded by a wide range of oscillations and it is important to consider how these oscillations affect other things. In other words, what happens when an oscillating force is continuously applied to another system? To understand this, it is helpful to first consider a very simple example – what happens when we keep pushing a child on a swing (see Figure 4.20)?

In this case it is fairly obvious, it depends on when we push the swing and in what direction. If we want bigger swings (increasing amplitudes) we should push once every oscillation in the direction the child is moving. In more scientific terms, we would say that we need to apply an external force which has the same frequency as, and is in phase with, the natural frequency of the swing.

When something is disturbed and then left to oscillate without further interference, it is said to oscillate at its **natural frequency** of vibration.

A **forced oscillation** occurs when an external oscillating force acts on another system tending to make it oscillate at a frequency which may be different from its natural frequency.

The most important examples of forced oscillation are those in which the frequency of the external force (often called the 'driving frequency' or the **forcing frequency**) is the same as the natural frequency. The child on the swing described on the previous page is an example of this and the effect is called **resonance**.

Resonance

4.3.4 **Describe** graphically the variation with forced frequency of the amplitude of vibration of an object close to its natural frequency of vibration.
4.3.5 **State** what is meant by resonance.

Resonance is the increase in amplitude and energy of an oscillation that occurs when an external oscillating force has the same frequency as the natural frequency of the system.

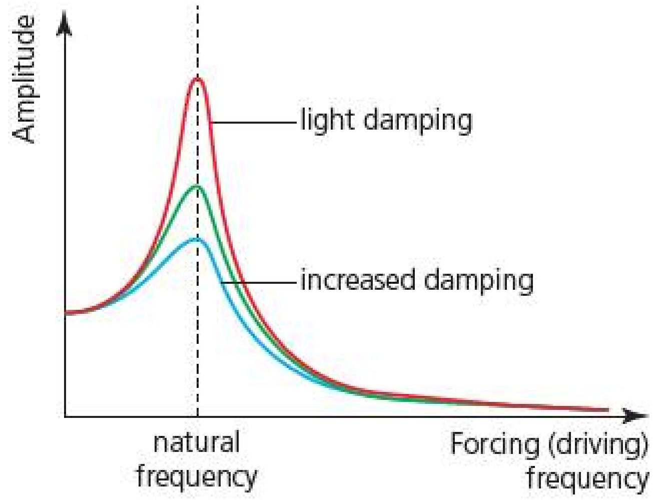


Figure 4.21 Typical frequency–response curves with different degrees of damping

Resonance causes oscillations to be amplified. Sometimes the increase in amplitude and the transfer of energy can be considerable.

Simple quantitative laboratory experiments into the effects of resonance can be difficult to perform, but they can produce interesting results which show how varying the driving/forcing frequency affects the amplitude. When the force is first applied, the oscillations may well be somewhat erratic, but the system will settle into a regular pattern of movement with a maximum amplitude.

A typical **frequency–response curve** (see Figure 4.21) rises to a maximum amplitude at the natural frequency, but the sharpness and height of the peak is also dependent upon the amount of damping in the system. The greater the damping, the greater the dissipation of energy, and therefore the smaller the amplitude.

There may be smaller resonance peaks at multiples of the natural frequency (not shown in diagram).

Examples of resonance

4.3.6 **Describe** examples of resonance where the effect is useful and where it should be avoided.

There are many important examples of resonance. Some are useful, but many are unwanted and we usually try to reduce their damaging effects.

Unwanted resonance

- Parts of almost all engines and machinery (and their surroundings) could vibrate destructively when their motors are operating at certain frequencies. For example, a washing machine may vibrate violently when the spinner is running at a certain frequency. Parts of a car can vibrate when the engine reaches a certain frequency.
- Earthquakes may well affect some buildings more than others. The buildings that are most damaged are often those which have natural frequencies close to the frequencies of the earthquake.
- Strong winds can cause dangerous resonance in structures like bridges.

To reduce the risk of damage from resonance two things can be done:

1 Change the design so that the natural frequencies are not the same as any possible forcing frequencies. This will mean changing the stiffness and mass of the relevant parts of the structure.
2 Ensure that there is enough damping in the structure and that it is not too rigid, so that energy can be dissipated.

Figure 4.22 Resonance may be one reason why some buildings collapse in an earthquake

Additional Perspectives

Resonance and bridges

If you have ever crossed a small suspension bridge for walkers (Figure 4.23), you will probably know how easy it is to set it vibrating with increasing amplitude by shaking it or stamping your feet at the 'right' frequency. This is because it would be too difficult or expensive to build such a bridge with a natural frequency which is very different from a frequency people can easily reproduce, or to design in greater damping features for a simple bridge.

Figure 4.23 Walker on a suspension bridge in Nepal

The resonance of bridges has been well understood for many years and the flexibility of suspension bridges makes them particularly vulnerable. The famous collapse of the newly built Tacoma Narrows Bridge in the USA in 1940 is widely given as a simple example of resonance caused by the wind, although this is only part of a much more complex explanation. More recently, the Millennium Bridge across the river Thames in London had to be closed soon after its opening in June 2000 because of excessive lateral (sideways) oscillations due to resonance (Figure 4.24).

Figure 4.24 The Millennium Bridge in London was affected by resonance

In this case *positive feedback* was important. The slow oscillations of the bridge made people sway with the same frequency and their motion simply increased the forces on the bridge which were causing resonance. The problem was solved by adding energy-dissipating dampers, but it was about 18 months before the bridge could reopen.

Question

1 Suggest what is meant by 'positive feedback' and give another example.

Useful resonance

- The molecules of certain gases (for example water and carbon dioxide) oscillate at the same frequency as infrared radiation emitted from the Earth. These gases absorb energy because of resonance, and this results in the planet being warmer than it would be without the gases in the atmosphere. This is known as the greenhouse effect (Chapter 8).
- Radios and TVs are 'tuned' by changing the frequency of an electronic circuit until it matches the driving frequency provided by the transmitted signal.
- The absorption of certain light frequencies from a white light spectrum can provide information about the energy levels within the absorbing material (Chapters 7 and 13).
- Your legs can be thought of as pendulums with their own natural frequency. If you walk with your legs moving at that frequency, energy will be transferred more efficiently and it will be less tiring (we tend to do this without thinking about it).
- Quartz crystals can be made to resonate using electronics and the resulting oscillations are useful in driving accurate timing in devices such as watches and computers.
- The sound from musical instruments can be amplified if the vibrations are passed onto a supporting structure that can resonate at the same frequency. An obvious example would be the strings on a guitar causing resonance in the box on which they are mounted. Because the box has a much greater surface area it produces a much louder sound than the string alone.
- Magnetic resonance imaging (MRI) is a widely used technique for obtaining images of features inside the human body. Electromagnetic waves of the right frequency are used to change the spin of protons (hydrogen nuclei) in water molecules.

23 Use the Internet to find out how scientists determine where the epicentre of an earthquake occurred.

24 A wing mirror on a car resonates at multiples of its natural frequency of 20 Hz.
  a Sketch a graph to show the frequency response of the mirror as the rpm (revolutions per minute) of the car engine increase from 1000 to 4000.
  b Suggest how the vibrations of the mirror could be reduced.
  c Add a second curve to your graph to show the new frequency response.

4.4 Wave characteristics

Progressive waves

4.4.1 **Describe** a wave pulse and a continuous progressive (travelling) wave.
4.4.2 **State** that progressive (travelling) waves transfer energy.

We are all familiar with water waves but there are many other types of wave, all of which have similar characteristics. These include many important waves which we cannot see. *Mechanical waves*, like sound, involve oscillating masses and need a material through which to travel, but waves like light (*electromagnetic waves*) can also travel across empty space.

Clearly, the study of waves is a very important topic in physics and we may summarize the importance of waves by saying that they can transfer energy from place to place without transferring matter. Furthermore, certain types of waves can be modified to transfer information – to communicate. Sound waves and radio waves are the most obvious examples.

The easiest way to develop an understanding of *all* kinds of waves is to begin by considering a simple one-dimensional system, for example, a wave sent along a rope by continuously shaking one end, as shown in Figure 4.25. Demonstrations with waves on ropes, strings or long springs are particularly useful because they are easy to produce and observe, and the waves travel at watchable speeds.

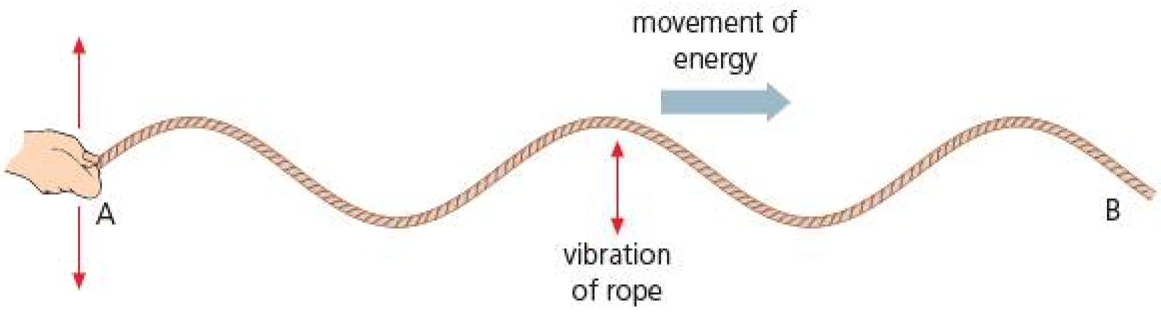


Figure 4.25 Creating a wave by shaking the end of a rope

Consider the idealized example of the end of a rope being shaken (oscillated) up and down, or from side to side, with SHM. The oscillations will be passed along the rope, from each part to the next, with a delay. Each part of the rope will, in turn, do exactly the same thing (if there is no energy dissipation). That is, each part of the rope will oscillate with SHM in the same direction as the original disturbance. It is easy to imagine how the characteristically shaped **sinusoidal wave** (as shown in Figure 4.25) is produced. 'Sinusoidal' means in the shape of a sine wave (or a cosine wave). The oscillations which transfer many different kinds of waves can be considered to be simple harmonic.

Scientists describe the motion of a wave away from its source as **propagation**. Any substance through which it is passing is called the **medium** of propagation. (The plural of medium is *media*.)

Waves which transfer energy away from a source are known as **continuous progressive waves** (or **travelling waves**).

Figure 4.26 Ocean waves transferring a lot of energy at Brighton, England, but there is no net movement of the water itself

It is important to realize that when we observe a wave, despite our impression of movement, the medium itself does *not* move continuously in the direction of apparent wave motion.

Progressive waves transfer energy without transferring matter. There is no *net* motion of the medium. For example, ocean waves may 'break' and 'crash' on to a shore or rocks (Figure 4.26), but there is no overall (net) movement of water from the sea because of the waves. A wooden log floating on water will just oscillate up and down as waves pass.

If the disturbance is not continuous but a single event (perhaps just one oscillation), then we describe the spreading disturbance as a **pulse**, rather than a wave.

Different kinds of waves

4.4.3 **Describe** and give examples of transverse and of longitudinal waves.
4.4.5 **Describe** the terms crest, trough, compression and rarefaction.

All waves are one of only two kinds: **transverse waves** or **longitudinal waves**. We first consider transverse waves.

Transverse waves

It helps our understanding of waves passing through a continuous medium, like a rope, to imagine a model in which the medium is represented by separate (discrete) particles, as shown in Figure 4.27. The black line represents a transverse wave moving to the right; the arrows show which way the particles are moving. The second wave, shown in red, represents the position of the same wave a short time later.

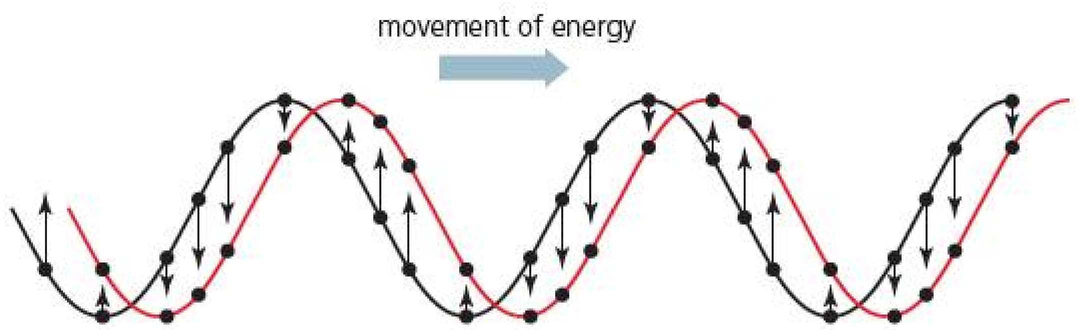


Figure 4.27 Movement of particles as a transverse wave moves to the right

Each part of the medium is doing the same thing (oscillating with the same frequency and same amplitude). But the different parts of the medium have different displacements at any particular time. That is, the parts of the medium are not all moving in *phase* with each other. The kind of wave shown in Figure 4.27 is described as transverse.

In a *transverse wave* each part of the medium oscillates *perpendicularly* to the direction in which the wave is transferring energy.

The 'tops' of transverse waves (especially water waves) are called **crests** and the 'bottoms' of the waves are called **troughs** (Figure 4.28).

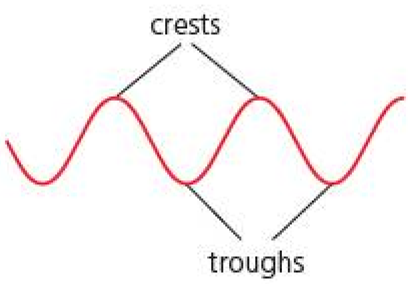


Figure 4.28 Crests and troughs of a transverse wave

Examples of transverse waves include light and all the other parts of the electromagnetic spectrum, water waves and waves on stretched strings and ropes. Transverse mechanical waves cannot pass through gases because of the random way in which particles move in gases.

Longitudinal waves

In a *longitudinal wave* each part of the medium oscillates *parallel* to the direction in which the wave is transferring energy.

We can demonstrate a longitudinal wave using a spring, but we first need to stretch it so that the coils are not touching (demonstrations often use 'slinky' springs). In order to make the longitudinal wave, the end of the spring is then oscillated 'backwards and forwards' (ideally with SHM) along the line of the spring.

A diagram of a longitudinal wave modelled by a slinky spring is shown in Figure 4.29. Characteristic **compressions** (where the spring is compressed) and **rarefactions** (where the spring is stretched) are marked. Longitudinal waves are sometimes called *compression waves*.

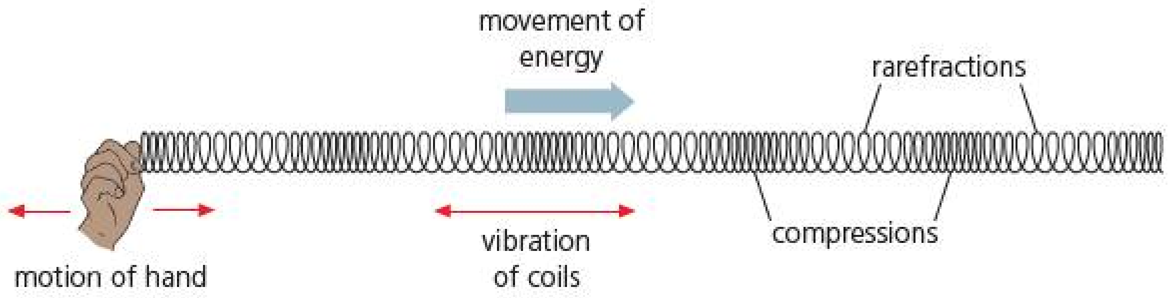


Figure 4.29 Oscillations of the medium transferring a longitudinal wave

As with transverse waves, when a longitudinal wave transfers energy through a medium, the medium itself does not undergo net translational motion away from the source.

Examples of longitudinal waves include sound and compression waves in solids. Earthquakes create both longitudinal and transverse waves in the Earth's rocks.

Waves in two dimensions

4.4.4 Describe waves in two dimensions, including the concept of wavefronts and of rays.

After discussing waves on a one-dimensional medium, like a rope, to further develop our understanding we will now consider progressive waves travelling in *two* dimensions. It will be helpful to think first of the waves on the surface of water, with which we are all familiar (Figure 4.30). Small waves (ripples) in tanks of shallow water are often used in laboratories to observe the behaviour of water waves (Figure 4.31 shows such a ripple tank).

Figure 4.30 Circular waves spreading out on a pond

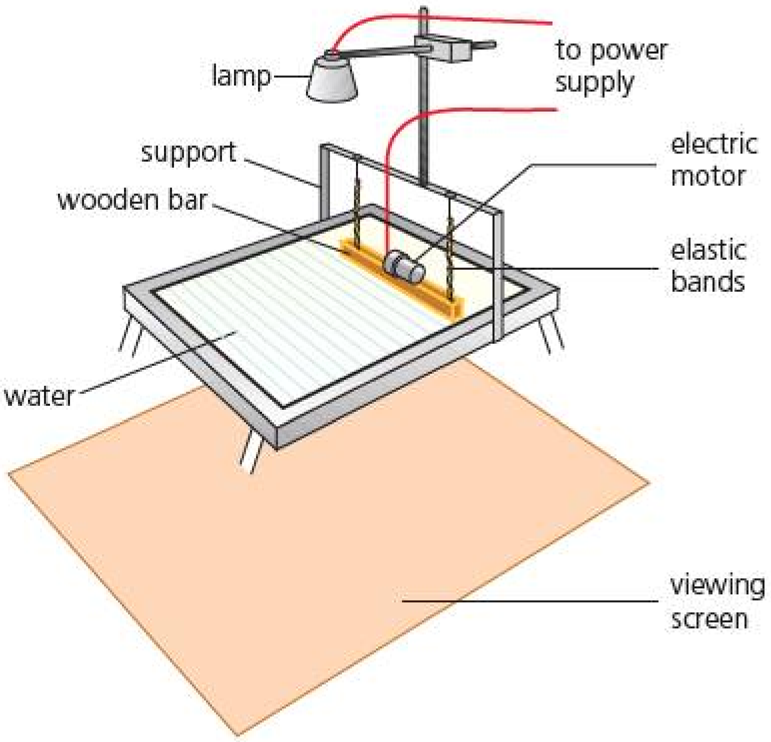


Figure 4.31 Ripple tank used to investigate wave behaviour

To make a wave on a flat water surface we need to disturb the water in some way, for example by repeatedly sticking a finger (or a stick) in and out of the water. Waves on the ocean are produced mainly by wind. Each part of the water affects the water around it and so the oscillation is passed on and away from the source, with a time delay. Waves on a ripple tank can also be made by a motor that vibrates a small dipper or beam suspended on the water surface.

Wavefronts and rays

When observing waves, we typically concentrate our attention on the crests of waves and, seen from above, waves spreading out on a smooth water surface with equal speeds in all directions will appear circular.

Although a wave is moving, we often want to represent the pattern of the waves (rather than their motion) on paper, or a computer screen. If we draw a momentary position of the moving waves on paper, then the lines we draw are called **wavefronts**. A wavefront is a line joining adjacent points moving in phase (for example, a line joining points where there are wave crests, or where there are troughs). The distance between adjacent wavefronts is called the **wavelength**.

Figure 4.32 shows circular wavefronts spreading from a point source. Note that the wave amplitude will decrease as the waves spread out because the same amount of energy is spread over a greater length.

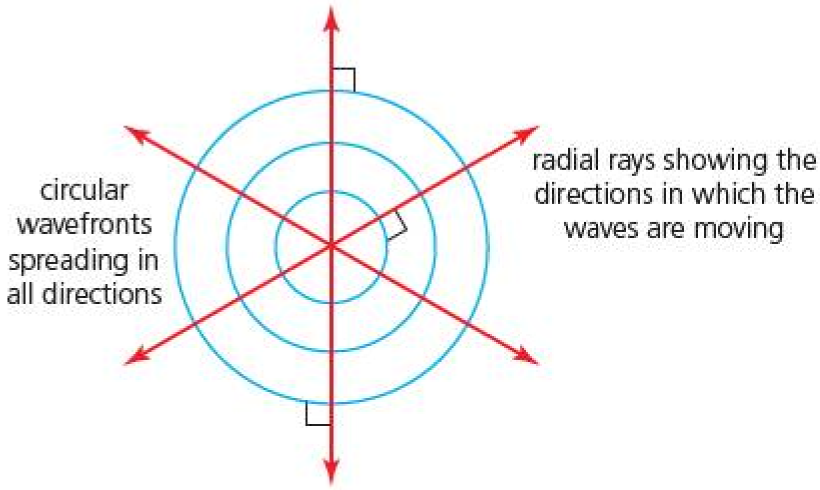


Figure 4.32 Circular wavefronts and radial rays spreading from a point source

Lines pointing in the directions in which the wave energy is being transferred are called **rays**. A ray is always perpendicular to the wavefronts that it is representing. The movement of circular wavefronts is represented by *radial* rays spreading from a point source, as shown in Figure 4.32.

Figure 4.33 shows waves which could have been made by a straight beam oscillated on a water surface. Wavefronts like this, which are parallel to each other, are called **plane wavefronts**. The movement of plane wavefronts is represented by *parallel* rays.

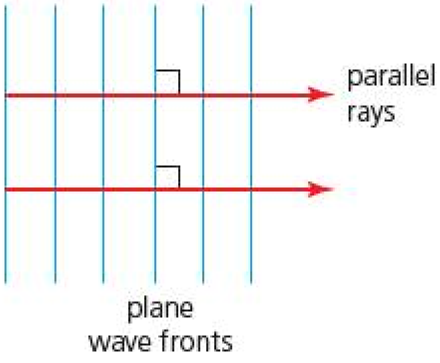


Figure 4.33 Plane wavefronts and parallel rays which are not spreading out

When circular wavefronts have travelled a long way from their source they become very nearly parallel and will approximate to plane wavefronts. An important example of this is light waves coming from a long way away, which we usually consider to be plane (for example, wavefronts from the Sun).

We have described wavefronts in terms of two-dimensional waves on water surfaces, but similar ideas and terminology can be used to describe *all* waves in two or three dimensions.

Representing waves graphically

4.4.7 **Draw** and **explain** displacement–time graphs and displacement–position graphs for transverse and for longitudinal waves.

Two similar graphs can be drawn to represent waves (transverse or longitudinal) and they are easily confused because they look similar. These are displacement–time graphs and displacement–position (distance) graphs. Compare Figures 4.34 and 4.35.

Figure 4.34 shows the position of *many* particles in the medium at an instant in time (a 'snapshot'). Figure 4.35 shows the movement of an *individual* particle as the wave passes through the medium. (This is similar to the SHM graphs shown earlier in this chapter.)

The shape of these graphs may suggest that they only represent transverse waves, but it is important to realize that similar graphs can also be used to represent the displacements in

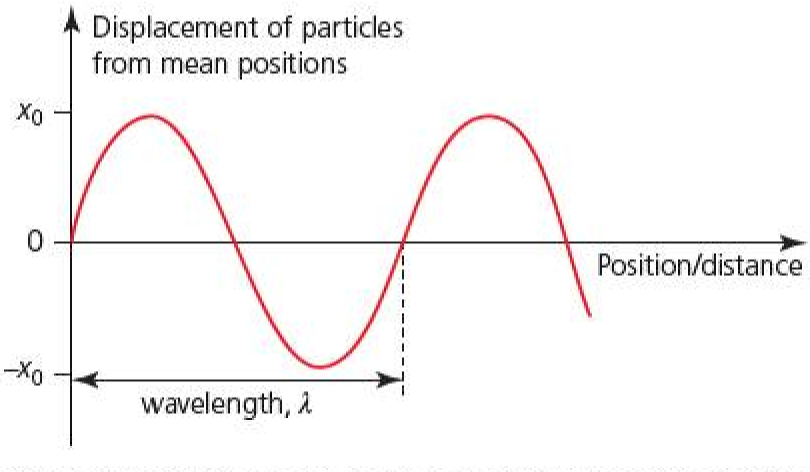


Figure 4.34 Displacement–position graph for a medium showing the meaning of wavelength, λ

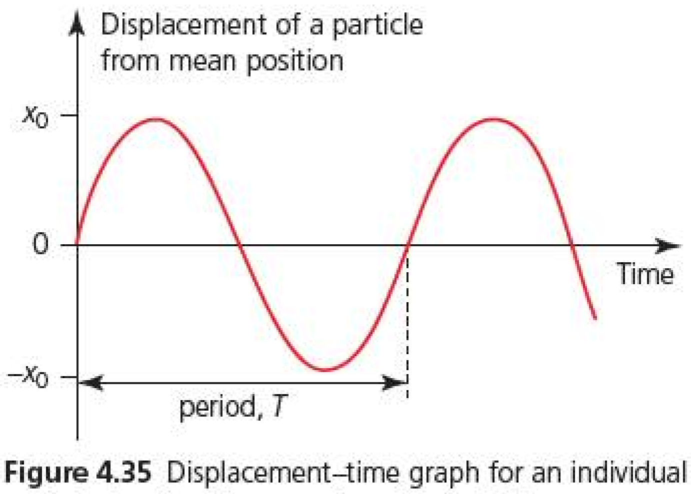


Figure 4.35 Displacement–time graph for an individual particle showing the meaning of period, T

longitudinal waves (or even the variations in pressure when sound waves travel through a gas). Figure 4.36 represents how the random arrangement of molecules in air might change as a longitudinal sound wave passes through. The compressions correspond to regions of higher than average air pressure and the rarefactions are regions of lower than average air pressure.

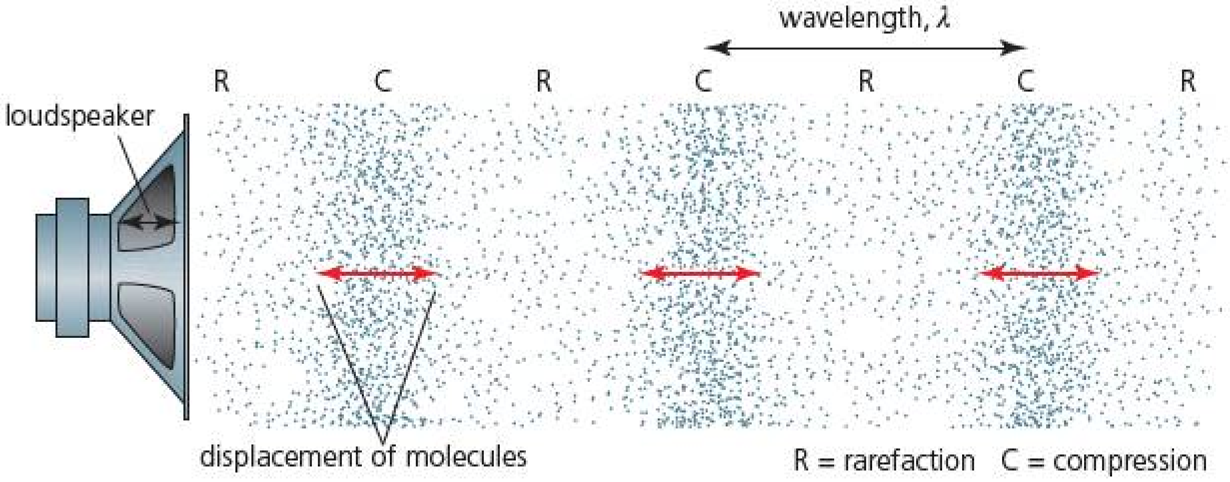


Figure 4.36 Arrangement of molecules in air as sound passes through

Basic terms used for describing waves

4.4.6 Define the terms *displacement, amplitude, frequency, period, wavelength, wave speed* and *intensity*.

Many of the terms we use to describe waves are similar to those introduced when discussing oscillations.

- **Displacement**, x, is defined as the distance a particle has moved from its equilibrium (mean) position in a specified direction.

- **Amplitude**, x_0, is defined as maximum displacement.

Waves with greater amplitude transfer more power. For example, a wave with twice the amplitude transfers four times as much power.

- **Intensity**, I, is defined as the wave power passing through unit area.

$$I = \frac{P}{A}$$

The unit of intensity is W m^{-2}. The intensity, of a wave is proportional to its amplitude squared:

$$I \propto x_0^2$$

- The **wavelength**, λ, of a wave is defined as the shortest distance between two points moving in phase. (It is also the perpendicular distance between adjacent wavefronts.)

Wavelength is easily shown on a displacement–position graph (Figure 4.34). Wavelength is usually measured in metres.

- The (time) **period**, T, of a wave is defined as the time it takes for one complete wave to pass a given point (or the time for one complete oscillation of a point within the medium).

Period is easily represented on a displacement–time graph (Figure 4.35).

- The **frequency**, f, of a wave is defined as the number of waves that pass a given point in unit time (or the number of oscillations at a point within the medium in unit time).

- **Wave speed** is defined as the distance travelled by a wave (wavefront) in unit time.

Like other speeds, we use the symbol v for wave speed. However, the speed of electromagnetic waves has a symbol of its own, c. All wave speeds are usually measured in m s^{-1}.

4.4.8 **Derive** and **apply** the relationship between wave speed, wavelength and frequency.

Wave equation

We know that, in general, speed equals distance moved divided by time taken. A wavefront moves a distance of one wavelength in one time period, so that $v = \lambda/T$. Or, since $T = 1/f$,

$$v = f\lambda$$
velocity = frequency × wavelength

This equation is in the IB *Physics data booklet*.

This is a simple, but very important and widely used equation. It can be used for *all* types of wave.

It is important to realize that, once a wave has been created, its frequency cannot change. So if there is a change of speed, for example after entering a different medium, there will be a corresponding change of wavelength (slower speed, smaller wavelength or higher speed, greater wavelength).

Electromagnetic waves

4.4.9 **State** that all electromagnetic waves travel with the same speed in free space, and **recall** the orders of magnitude of the wavelengths of the principal radiations in the electromagnetic spectrum.

Light travels as a transverse wave and the **spectrum** of visible 'white' light, from red to violet, is a familiar sight (Figure 4.37). The wavelengths of light waves are very small and the different colours of the spectrum have different wavelengths. Red light has the longest wavelength (approximately 7×10^{-7} m) and violet has the shortest wavelength in the visible spectrum (approximately 4×10^{-7} m).

Figure 4.37 Spectrum of visible light produced by a prism

But light is only a very small part of a much larger group of waves, called the **electromagnetic spectrum**, the major sections of which are listed in Table 4.1. It is important to realize that the spectrum is *continuous* and that there is often no definite boundary between one section and another; the sections may overlap. For example, gamma rays from a radioactive source may be identical to X-rays from an X-ray tube.

We are surrounded by many of these waves all of the time and they come from a variety of very different sources. **Light** is just the name we give to those electromagnetic waves which human beings can detect with their eyes. The only basic difference between the different kinds of electromagnetic waves is their wavelength, so it is not surprising that, under appropriate circumstances, all of these transverse waves exhibit the same basic wave properties.

All electromagnetic waves can travel across empty (*free*) space (in other words, a vacuum) with exactly the same speed, $c = 3.00 \times 10^8\ \mathrm{m\,s^{-1}}$.

This constant is listed in the IB *Physics data booklet*.

Electromagnetic waves do not *need* a medium through which to travel, but when electromagnetic waves pass through other materials, their speeds are less than in vacuum (although their speed in air is almost identical to their speed in vacuum).

Table 4.1 gives a very brief summary of the major parts of the spectrum. Waves of wavelength smaller than about 10^{-8} m can cause damage to the human body, even at low intensity.

Table 4.1 The different sections of the electromagnetic spectrum

Name	A typical wavelength	Origins	Some uses
Radio waves	10^{2} m	Electronic circuits/aerials	Communications, radio, TV
Microwaves	10^{-2} m	Electronic circuits/aerials	Communications, mobile phones, ovens, radar
Infrared (IR)	10^{-5} m	Everything emits IR but hotter objects emit *much* more IR than cooler things	Lasers, heating, cooking, medical treatments, remote controls
Visible light	5×10^{-7} m	Very hot objects, light bulbs, the Sun	Vision, lighting, lasers
Ultraviolet (UV)	10^{-8} m	Sun, UV lamps	Fluorescence
X-rays	10^{-11} m	X-ray tubes	Medical diagnosis and treatment, investigating the structure of matter
Gamma rays	10^{-13} m	Radioactive materials	Medical diagnosis and treatment, sterilization of medical equipment

The range of different wavelengths, from 10^{-14} m or less, up to 10^{4} m or more, is enormous and it is important to remember an order of magnitude for the principal radiations.

The exact nature of these waves was a major puzzle in science for a long time because all other wave types need a medium in which to travel, whilst electromagnetic waves can transfer energy across free space.

We now know that these transverse waves are not carried by oscillating particles but by linked oscillating electric and magnetic fields (which do not need a medium), as shown in Figure 4.38, hence their name electromagnetic waves. However, electromagnetic waves also have some properties which can only be explained by thinking of them as 'particles' (and not waves), called **photons**. Each photon transfers an individual amount of energy dependent on its frequency. These ideas are discussed in more detail in Chapters 7 and 13 and a full understanding is not expected here.

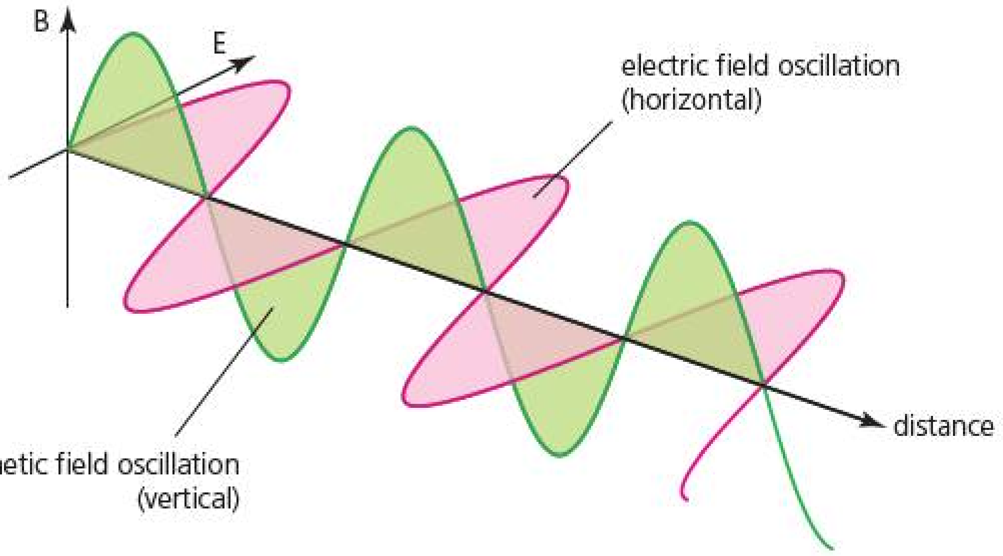


Figure 4.38 Electromagnetic waves are combined electric and magnetic fields

Worked example

6 The crests of waves passing into a harbour are 2.1 m apart and have an amplitude of 60 cm. Ten waves pass an observer every minute.

a What is their frequency?

b What is their speed?

c On another day, the waves have an amplitude of 1.8 m. How does the intensity of the waves compare on the two days?

a $T = 60/10 = 6.0$ s

So $f = 1/T = 1/6.0 = 0.17$ Hz

b $v = f\lambda$

$v = 0.17 \times 2.1 = 0.35$ m s^{-1}

c Amplitude has increased by a factor of 3 (1.8/0.60), so intensity has increased by a factor of $3^2 = 9$.

25 **a** Calculate the wavelength in air of a high pitched sound of frequency 2.20 kHz (speed of sound in air = 340 m s^{-1}).
b When the same sound passed into a metal its wavelength increased to 2.05 m. What was the speed of sound in the metal?
c Suggest a reason why sound travels faster in solids than in gases like air.

26 **a** What is the frequency of radio waves of wavelength 560 m (speed of electromagnetic waves = 3.0×10^8 m s^{-1})?
b How long would it take a radio signal to travel from:
i Delhi to Mumbai (1407 km)?
ii a satellite orbiting at a height of 33 000 km to the Earth's surface?
iii from Earth to Mars? (Use the Internet to help determine the maximum and minimum separations of the two planets.)

27 A power station on the sea shore transfers wave energy of water to electricity at a rate of 50 kW on a day when the average wave amplitude is 1.2 m.
a What will the output be if the wave amplitude doubles?
b What assumption did you make?
c What amplitude of waves might be expected to produce an output of 150 kW?
d What power could be transferred from waves of amplitude 1.0 m? (There is more about wave power generators in Chapter 8.)

28 **a** What is the frequency (in MHz) of a gamma ray which has a wavelength of 4.4×10^{-12} m?
b Suggest a possible reason why electromagnetic radiation of higher frequencies is more dangerous to humans than lower frequencies.

29 **a** Calculate the wavelength of a visible light of frequency of 5.5×10^{14} Hz.
b Suggest what colour the light may be.
c Some radiation from the Sun has a wavelength of 6.0×10^{-8} m. In what part of the electromagnetic spectrum is this radiation?

30 The total separation of five wave crests on a ripple tank is 5.6 cm. If the waves are made by a vibrator with a frequency of 12 Hz, how long will it take the waves to travel a distance of 60 cm?

31 As you are reading this, which electromagnetic radiations are in the room?

Additional Perspectives

Heinrich Hertz

The unit for frequency is named after the German physicist, Heinrich Hertz. His most famous achievement in 1886 was to prove experimentally for the first time that electromagnetic waves (radio waves) could be produced, transmitted and detected elsewhere. Although the distance involved was very small, it was the start of modern wireless communication. It was left to others (such as Guglielmo Marconi) to develop techniques for the transmission over greater and greater distances and then to modify the amplitude, frequency or phase of the radio waves to transfer information (for example, speech).

Figure 4.39 Heinrich Hertz

Hertz is often quoted as saying his discovery was 'of no use whatsoever' and he was not alone in expressing that opinion at the time. The history of science contains many such statements and predictions which later turn out to be incorrect. The importance of discoveries or inventions is often only realized many years later. This is one reason why most scientists think it would be foolish to limit research to those projects which have immediate and obviously useful applications.

Tragically, Hertz contracted a fatal disease and died at the age of 36, less than eight years after his discovery, which was a long time before the full implications of his work had become clear.

Question

1 Modern scientific research can be very expensive, so would countries be wiser to spend their money on other things (for example, improving general medical care), unless the research is obviously leading towards a definite and useful aim?

4.5 Wave properties

Changes in the speed of waves, or obstacles placed in their path, will change their shape and/or direction of motion, with very important consequences. These effects are called **reflection**, **refraction**, and **diffraction**. **Interference** happens when waves combine. We will discuss each of these four wave properties in this section.

Reflection

4.5.1 (part) **Describe** the reflection of waves at a boundary between two media.

When a wave meets a **boundary** between two different media usually some or all of the waves will be reflected back. Under certain circumstances some waves may pass into or through the second medium (we would then say that there was some **transmission** of the waves). An obvious example would be light waves passing through **transparent** materials, for example liquids or various kinds of glass. Figure 4.40 shows both reflection and transmission at the same window.

We will refer again to our two wave models (waves on springs and waves on water) to develop our understanding of reflection.

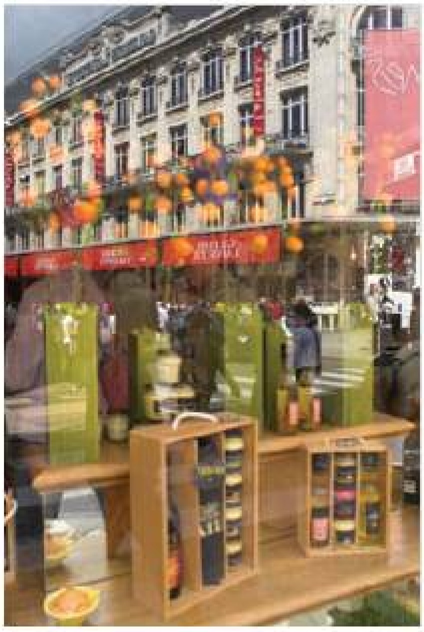


Figure 4.40 Light reflected off and passing through a window

Reflection in one dimension

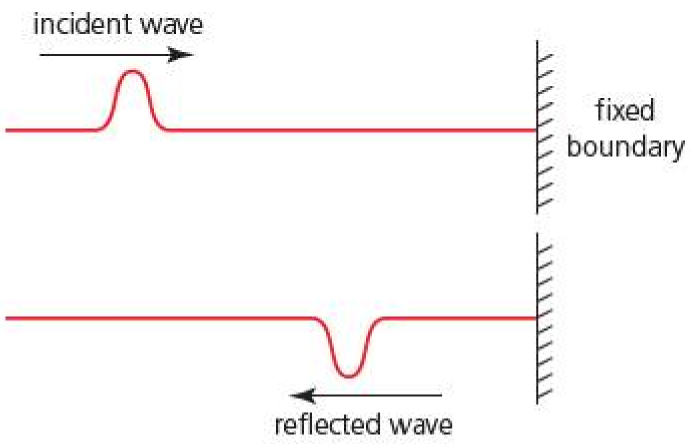


Figure 4.41 Reflection of a pulse off a fixed boundary

Figure 4.41 shows a single pulse on a spring or string travelling towards a fixed boundary where it is totally reflected, with no loss of energy. Note that the reflected wave is inverted. We say that it has undergone a phase change of π.

Pulses will change speed if they cross a boundary between springs/strings of different mass per unit length. In Figure 4.42 a transverse pulse is transmitted from a less 'dense' string to a more 'dense' string (greater mass per unit length), where it travels more slowly.

Note that there are now two pulses and both have reduced amplitude because the energy has been split between the two. The transmitted pulse will now have a slower speed but its phase has not changed. The reflected pulse returns with the same speed but has undergone a phase change of π.

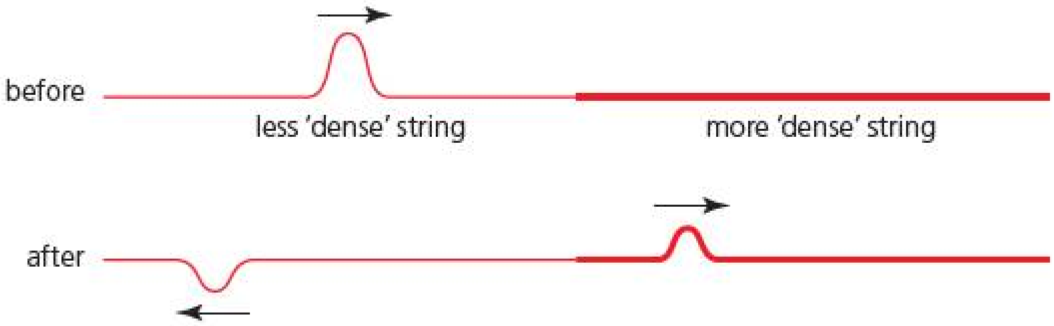


Figure 4.42 A pulse travelling into a 'denser' medium

In Figure 4.43 the transverse pulse is passing from a more 'dense' string to a less 'dense' string. This time there is no phase change for the reflected pulse.

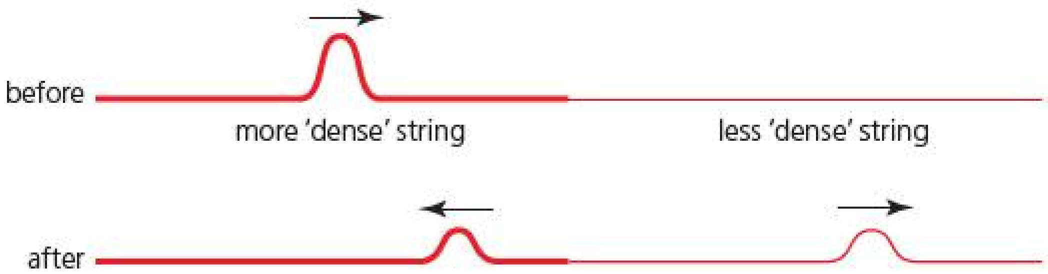


Figure 4.43 A pulse travelling into a less 'dense' medium

Longitudinal waves and pulses behave in a similar way to transverse waves.

Reflection in two dimensions

If plane waves reflect off a straight (plane) boundary between two media, they will reflect so that the reflected waves and the incoming (**incident**) waves make equal angles with the boundary. See Figure 4.44. The **angle of incidence**, *i*, is equal to the **angle of reflection**, *r*.

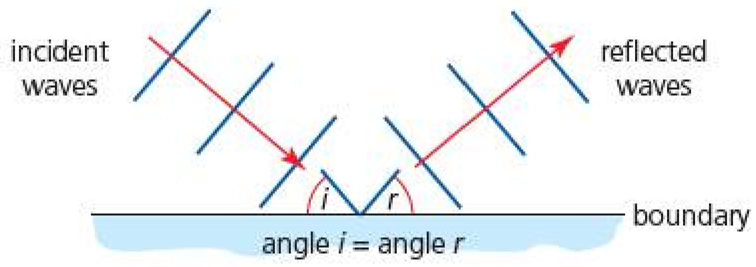


Figure 4.44 Reflection of plane waves from a straight boundary

When discussing the reflection of light, it is more common to represent reflection by a **ray diagram**, as shown in Figure 4.45.

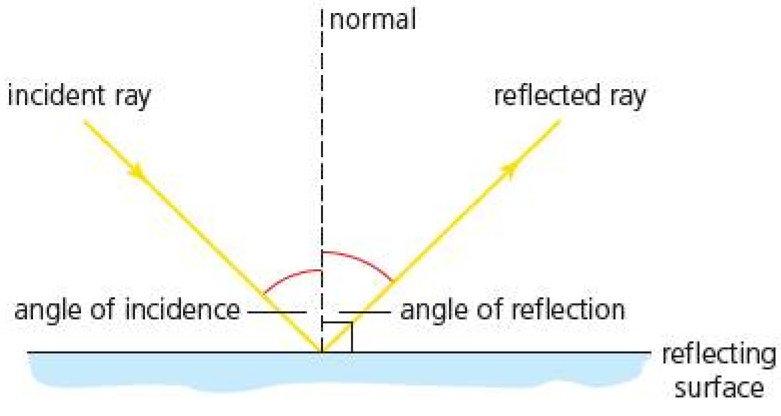


Figure 4.45 Reflection of rays from a reflecting surface

As before, the angle of incidence, *i*, is equal to the angle of reflection, *r*. But in this diagram the angles are measured between the rays and the 'normal'. The **normal** is a constructional line we draw on diagrams that is perpendicular to the surface.

In practice, as with the one-dimensional waves, some wave energy may also be transmitted as well as reflected. The transmitted waves may change direction. This effect is called refraction and it is discussed in the next section.

Additional Perspectives

Sound in large rooms

Sound reflects well off hard surfaces like walls, whereas soft surfaces, like curtains, carpets, cushions and clothes, tend to absorb sound. A sound that reaches our ears may be quite different from the sound that was emitted from the source because of the many and various reflections it may have undergone. Because of this, singing in the shower will sound very different from singing outdoors or singing in a furnished room.

Figure 4.46 A large auditorium designed for effective transmission of sound

In a large room designed for listening to music (such as an auditorium, Figure 4.46), sounds travel a long way between reflections and, since it is the reflections that are responsible for most of the absorption of the sound waves, it will take a longer time for a particular sound to fall to a level that we cannot hear. This effect is called **reverberation**. The longer reverberation times of bigger rooms mean that a listener may still be able to hear reverberation from a previous sound at the same time as a new sound is received. That is, there will be some 'overlapping' of sounds.

Reflections of sounds off the walls, floor and ceiling are also an important factor when music is being produced in a recording studio, although some effects can be added or removed electronically after the original sound has been recorded.

Questions

1 Suggest why soft surfaces are good absorbers of sound.
2 Will the acoustic engineers, who are responsible for the sound quality in an auditorium, aim for a short or a long reverberation time? How can the reverberation time be changed?
3 Find out what anechoic chambers are, and what they are used for.

Refraction

4.5.1 (part) **Describe** the transmission of waves at a boundary between two media.

Typically, waves will change speed when they travel into a different medium. The speed of water waves decreases as they pass into shallower water.

Figure 4.47 shows plane wavefronts arriving at a medium in which they travel slower (a 'denser' medium). The wavefronts are parallel to the boundary and the ray representing the wave motion is perpendicular to the boundary.

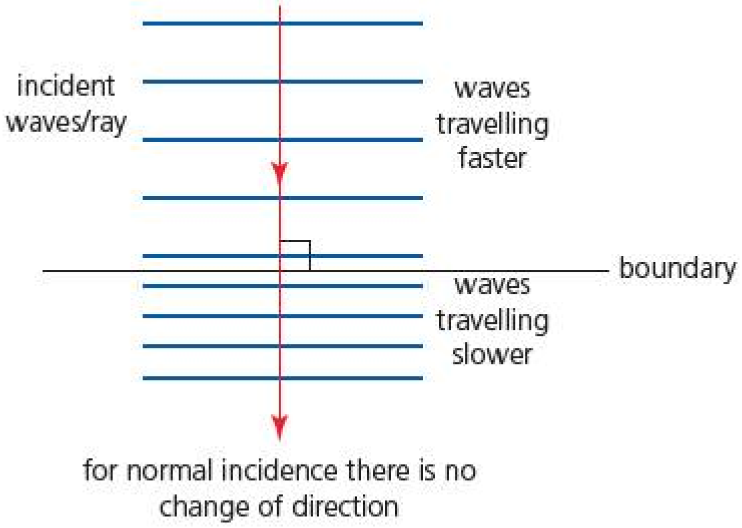


Figure 4.47 Waves slowing down as they enter a different medium

In this case there is no change of direction but, because the waves are travelling slower, their wavelength decreases, although their frequency is unchanged (consider $v = f\lambda$). Now consider what happens if the wavefronts are *not* parallel to the boundary, as in Figure 4.48.

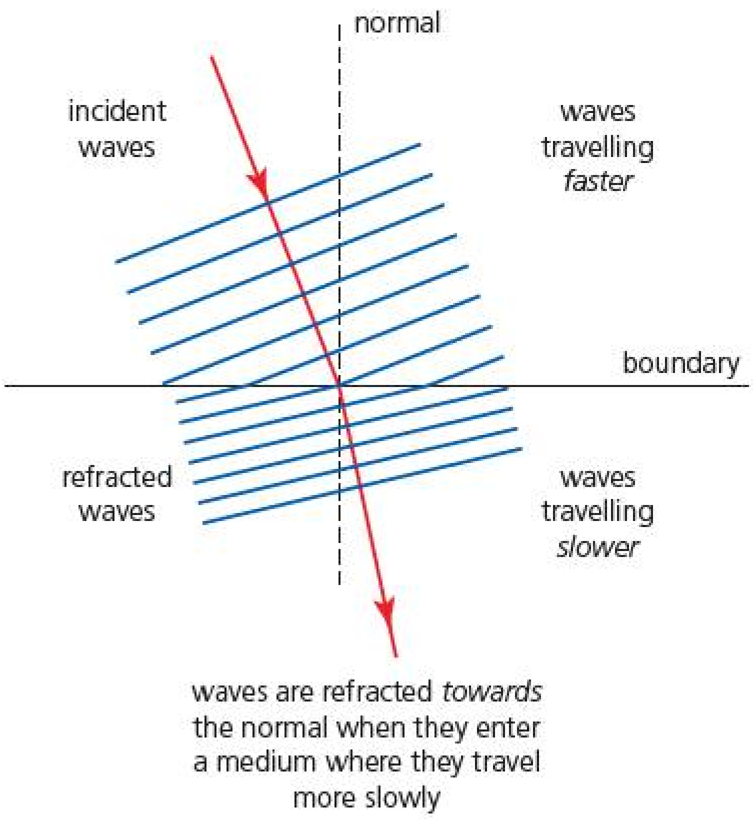


Figure 4.48 Waves refracting as they enter a denser medium

Different parts of the same wavefront reach the boundary at different times and consequently they change speed at different times. There is a resulting change of direction which we call **refraction**. The greater the change of speed, the greater the change of direction.

When waves enter a medium in which they travel more slowly, they are refracted *towards the normal*. Conversely, when waves enter a medium in which they travel faster, they are refracted *away from the normal*. This is shown in Figure 4.49, but note that this is just the reverse of Figure 4.48.

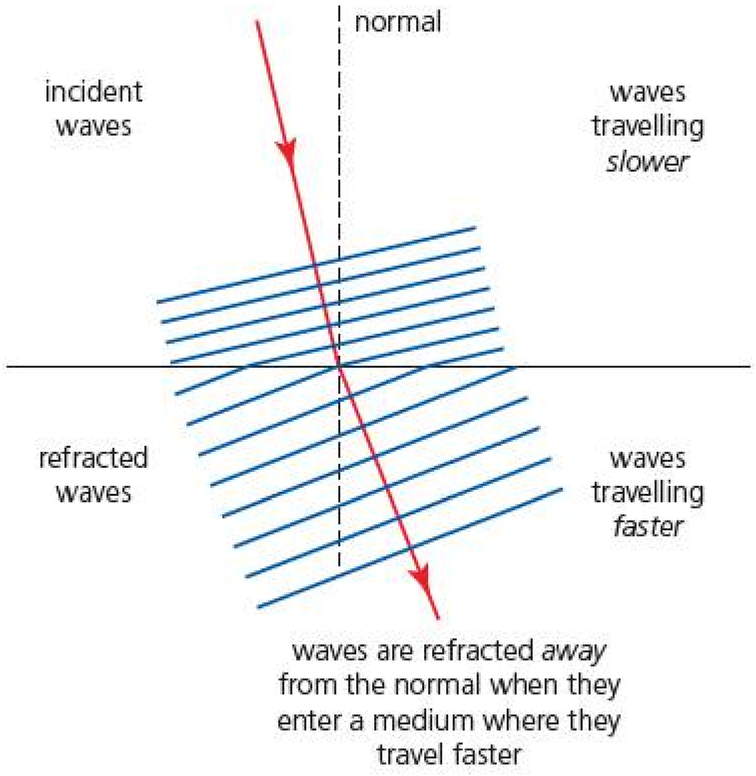


Figure 4.49 Waves refracting as they enter a less dense medium

The refraction of light is a familiar topic in the study of physics, especially in optics work on lenses and prisms, but all waves tend to refract when their speed changes. Often this is a sudden change at a boundary between media, but it can also be a gradual change, for example if the density of a medium changes gradually.

The photograph in Figure 4.50 shows the gradual refraction of water waves approaching a beach. It is possible to learn about how the water depth is changing by observing the refraction of waves in shallow water. Waves travel slower in shallower water, so the wave crests get closer together (smaller wavelength) because the frequency does not change.

Figure 4.50 Ocean waves refracting (and diffracting) as they approach a beach

Figure 4.51 A lens uses refraction to focus light

The focusing of a glass lens occurs because the light waves slow down and are refracted in a systematic way by the smooth curvature of the glass.

In Figure 4.52 light waves are refracted in a more disorganized way by the irregular changes in density of the hot air moving above the runway and behind the plane.

Figure 4.52 Refraction of light by gases

Snell's law of refraction

4.5.2 **State** and apply Snell's law.

Figure 4.53 shows a single ray of light representing the direction of waves being refracted when entering a medium where they travel slower (Figure 4.53a) and a medium where they travel faster (Figure 4.53b).

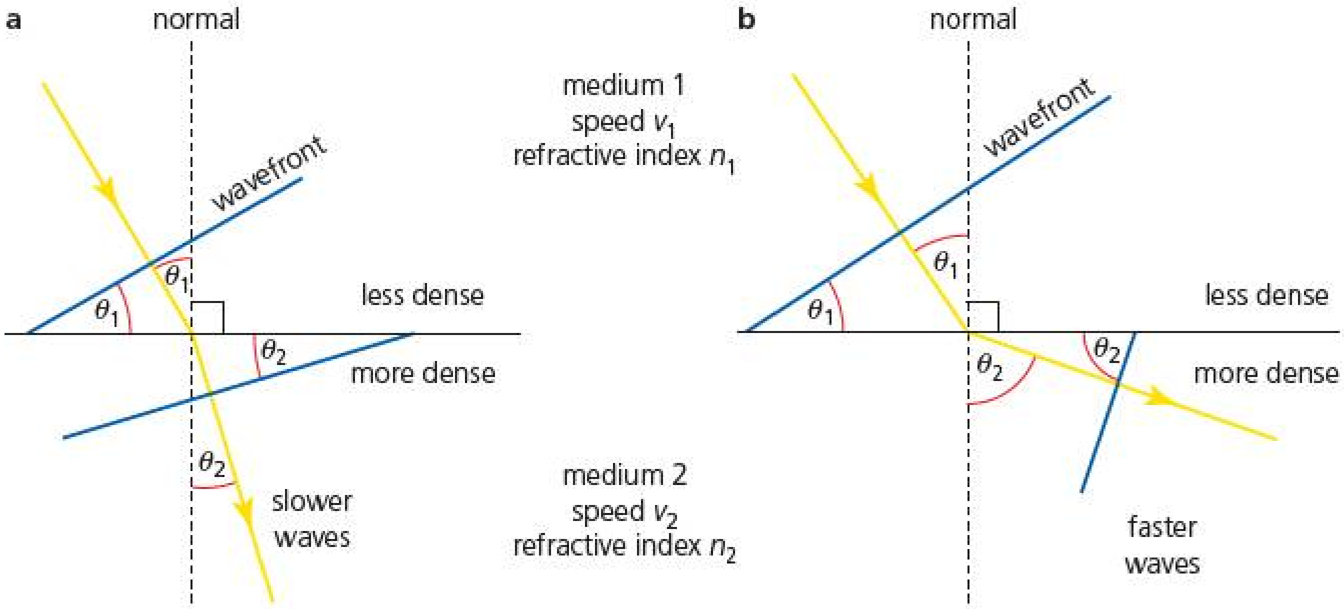


Figure 4.53 Light rays being refracted **a** towards the normal **b** away from the normal

For a boundary between any two given media, it was discovered that any angle of incidence, θ_1, was connected to the corresponding angle of refraction, θ_2, by the equation:

$$\frac{\sin\theta_1}{\sin\theta_2} = \text{a constant}$$

The constant is called the **refractive index** for light passing from medium 1 into medium 2. Refractive index is given the symbol n. To make it clear which media are involved, we write it in the form ${}_1n_2$, meaning the refractive index for light passing from medium 1 into medium 2. Since it is a ratio, there is no unit.

However, it is more common to refer to the refractive index of a single medium, such as water, or glass of a particular kind. The value that is quoted is the refractive index assuming that light is entering the medium from air (or a vacuum). That is, the refractive index of glass is:

$$n_{\text{glass}} = {}_{\text{vacuum}}n_{\text{glass}} = {}_{\text{air}}n_{\text{glass}}$$

Using trigonometry it is not difficult to show that the refractive index must also equal the ratio of wave speeds (v_1/v_2), so that:

$${}_1n_2 = \frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}$$

This is known as **Snell's law**.

It can also be easily shown that the refractive index for waves going from medium 1 into medium 2, ${}_1n_2$, equals n_2/n_1, so that the equation can be written as:

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1} = \frac{v_2}{v_1}$$

It is this form of Snell's law that is given in the IB *Physics data booklet*.

Worked examples

7 a If light waves represented by a ray making an angle of incidence of 60° enter into glass of refractive index 1.52, calculate the angle of refraction.

b What is the speed of light in this glass (speed of light in air = $3.0 \times 10^8\,\text{m}\,\text{s}^{-1}$)?

a ${}_{\text{air}}n_{\text{glass}} = \dfrac{\sin\theta_{\text{air}}}{\sin\theta_{\text{glass}}}$

$$1.52 = \frac{\sin 60}{\sin\theta_{\text{glass}}}$$

Angle of refraction = $\theta_{\text{glass}} = 35°$

b ${}_{\text{air}}n_{\text{glass}} = \dfrac{v_{\text{air}}}{v_{\text{glass}}}$

$$1.52 = \frac{(3.0 \times 10^8)}{v_{\text{glass}}}$$

$v_{\text{glass}} = 2.0 \times 10^8\,\text{m}\,\text{s}^{-1}$

8 a Figure 4.54 represents light waves passing out of water such that the ray has an angle of incidence of 40°. The refractive index of water is 1.3. Calculate the speed of light in water.

b What is the angle of refraction in air?

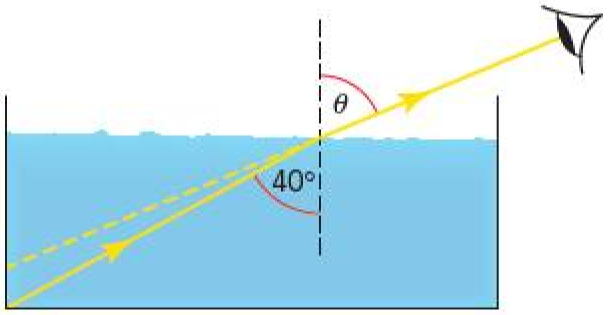


Figure 4.54

a ${}_{\text{air}}n_{\text{water}} = \dfrac{v_{\text{air}}}{v_{\text{glass}}}$

$$1.3 = \frac{(3.0 \times 10^8)}{v_{\text{water}}}$$

$v_{\text{water}} = 2.3 \times 10^8\,\text{ms}^{-1}$

b $_{air}n_{water} = \frac{\sin\theta_{air}}{\sin\theta_{water}}$

$$1.3 = \frac{\sin\theta_{air}}{\sin 40°}$$

$$\theta_{air} = 57°$$

An eye/brain receiving this light will 'assume' it came in a straight line and hence the water will look shallower than its real depth.

Dispersion of light

The speeds of different colours (wavelengths) of light in a particular medium (glass, for example) are not exactly the same. Red light travels the fastest and violet is the slowest. This means that different colours travelling in the same direction from the same source will not travel along exactly the same paths when they are refracted. When light goes through parallel-sided glass (like a window) the effect is not usually significant. However, when white light passes into and out of other shapes of glass (like **prisms** and **lenses**), or water droplets, it can be **dispersed** (separated into different colours). A triangular prism, as shown in Figure 4.55, is commonly used to disperse white light into a spectrum (as shown in Figure 4.37).

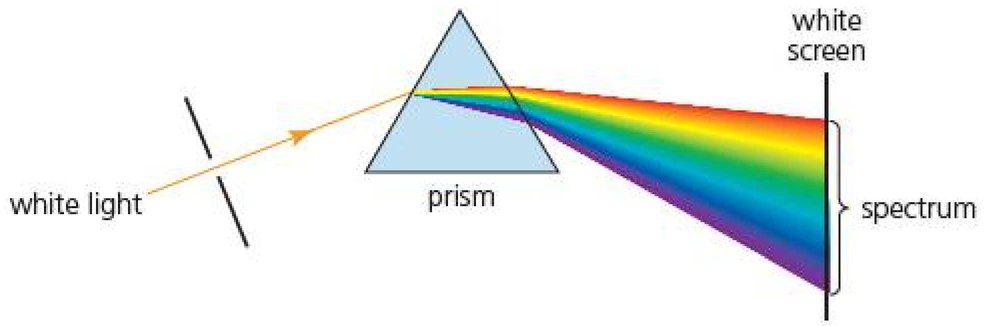


Figure 4.55 Using a prism to produce a spectrum of white light

Additional Perspectives

Ibn al-Haytham

Until recently the important role played by Islamic scientists has been somewhat neglected by other cultures. An 11th century scientist, Ibn al-Haytham (Figure 4.56), also known as Alhazen, is arguably one of the greatest physicists of all time. He was a pioneer of modern 'scientific method', with its emphasis on experimentation and mathematical modelling, but he lived hundreds of years before Galileo and the others who are widely credited with similar innovations.

His work was wide-ranging, but included experimentally and quantitatively investigating the refraction of light (similar to Snell's work centuries later). Incidentally, he is also credited by many with being the first to realize that the 'twinkling' of stars is due to refraction of light passing through the Earth's atmosphere.

Figure 4.56 Ibn al-Haytham

Questions

1 Find out the names of some other scientists and mathematicians from around the time of the 11th century, or a few hundred years before. Where did they live and what were their major achievements?

2 Are scientific developments sometimes achieved in isolation by lone individuals, or is collaboration, rather than secrecy, an important aspect in most research?

32 Light rays in air enter a liquid at an angle of 38°. If the refractive index of the liquid is 1.4, what is the angle of refraction?

33 Plane water waves travelling at $48\,\text{cm}\,\text{s}^{-1}$ enter a region of shallower water with the wavefronts at an angle of 34° to the boundary. If the waves travel at a speed of $39\,\text{cm}\,\text{s}^{-1}$ in the shallower water, predict the direction in which they will move.

34 Light rays travel at $2.23 \times 10^8\,\text{m}\,\text{s}^{-1}$ in a liquid and at $3.00 \times 10^8\,\text{m}\,\text{s}^{-1}$ in air.
a What is the refractive index of the liquid?
b Light rays coming out of the liquid into air meet the surface at an angle of incidence of 25°. What is the angle of the emerging ray to the normal in air?

35 A certain kind of glass has a refractive index of 1.55. If light passes into the glass from water (refractive index = 1.33) and makes an angle of refraction of 42°, what was the angle of incidence?

36 **a** Use trigonometry to show that the refractive index between two media is equal to the ratio of wave speeds (v_1/v_2) in the media.
b Show that the refractive index for waves going from medium 1 into medium 2, ${}_1n_2 = \frac{n_2}{n_1}$.

37 Consider again Figure 4.54.
a Calculate the angle of incidence in water needed for the refracted ray to be directed along the water surface (refractive index of water = 1.33).
b Suggest what will happen to incident rays striking the surface at even larger angles.

Diffraction

4.5.3 **Explain** and **discuss** qualitatively the diffraction of waves at apertures and obstacles.

When waves pass through gaps (apertures) or pass around obstacles in their path, they will tend to 'spread' or 'bend' around them. This important effect is called **diffraction** (a term which should not be confused with *refraction*). Waves often encounter objects in their path, and the study of diffraction is vital in appreciating how waves get from place to place. This has become especially important in the age of wireless communication.

All waves diffract under suitable conditions; the fact that something diffracts is clear evidence of its wave nature. Sometimes the effects of diffraction are very noticeable, as they usually are with water waves and sound waves, but sometimes diffraction is difficult to observe, as it is with light waves. The reason for this is because of the very important fact that the amount of diffraction is dependent on how the size of the wavelength compares to the size of the obstacle or gap. Diffraction is most important when the wavelength and the gap or obstacle are about the same size.

Figure 4.57 represents the two-dimensional diffraction of waves through **apertures** (gaps) and around obstacles. The diagrams can be applied to the diffraction of any waves, including in three dimensions. These diagrams are simplified.

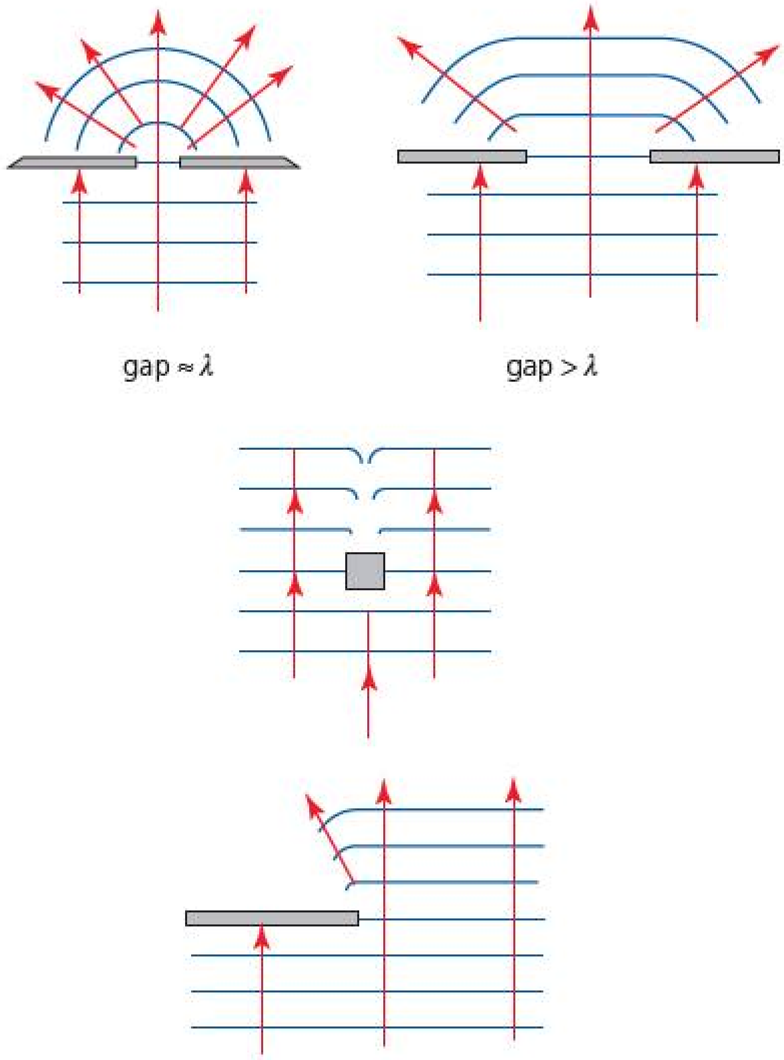


Figure 4.57 Diffraction of waves through gaps and around obstacles (reflected waves not shown)

Examples of diffraction

4.5.4 Describe examples of diffraction.

- **Sound** wavelengths typically vary from about 2 cm to 20 m. This means that sound is easily diffracted around corners, buildings, doors and furniture, for example. So, we can hear sounds even when we cannot see where they are coming from.
  Lower pitched sounds have longer wavelengths and diffract better around bigger things, like buildings, so we tend to hear them from further away. Low pitched sounds also spread away better from bigger (loud) speakers (often called 'woofers'), whilst high pitched sounds can use smaller speakers ('tweeters').

Figure 4.58 This large loudspeaker is good at emitting low frequencies at high volume

- All the various colours of **light** have wavelengths of less than 10^{-6} m (10^{-3} mm). This means that the diffraction of light tends to go unnoticed because only very small gaps diffract light. However, the diffraction of light at our eyes does limit our ability to see (resolve) details, and it also limits the resolution of telescopes and microscopes.
  Figure 11.21 in Chapter 11 is a photograph showing the diffraction of light. You can see some effects of diffraction by looking at a white surface through a narrow gap made between your fingers: dark lines are seen which are parallel to the length of the gap.
- **Radio waves** (including microwaves) have a very wide range of wavelengths, from a few centimetres up to one kilometre or more. When engineers design radio communication systems for radio, TV, satellite broadcasts and mobile phones, for example, they have to choose a suitable wavelength to use. This involves considering how far they want the waves to travel between transmitter and receiver, and whether there are obstacles such as buildings or hills in between. Ideally, the size of the transmitting and receiving aerials will also be approximately comparable to the size of the wavelength used, although cost and convenience may reduce aerial sizes. For example, the wavelengths used for mobile phones are typically a few centimetres.

Figure 4.59 Microwaves diffract when they are emitted from transmitting aerials

- Since diffraction effects (for a given wavelength) depend on the size of the diffracting object, it is possible to learn something about an object by observing and measuring how it diffracts a wave of known wavelength. This has many important applications, for example **X-rays** have wavelengths comparable to the size of atoms and the diffraction of X-rays has been very important for scientists learning about the spacing of atoms and how they are arranged in crystalline solids.

Additional Perspectives

Tsunamis

The consequences of the tsunamis following the massive earthquakes off the Indonesian island of Sumatra on 26 December 2004 and the north-eastern Japanese coast line on 11 March 2011 were tragic and overwhelming. Sudden and massive motion of the Earth's crust along a fault line passed energy to the ocean above, resulting in the movement of an enormous volume of water.

Any tsunami waves resulting from an earthquake travel at high speed (maybe for thousands of kilometres) with little loss of energy, and possible devastating consequences when they reach land.

Figure 4.60 The tsunami of December 2004

But why are some places affected much more badly than others? Of course the height of the land near the shore is a major factor, as is the distribution of homes and people. A full explanation must also take into account the refraction, reflection and diffraction of the incident waves as they approach a coast. Changes in depth of water and the orientation of such changes (relative to the coast) will affect the height and shape of the waves, and the direction in which they are moving. The shape of the coast line can result in reflections and diffractions which have a focusing effect.

Similar explanations can be used to show why some beaches are much better for surfing than others.

Question

1 What causes waves on the oceans and why do they always seem to come into shore (rather than travelling outwards)?

Interference

4.5.5 **State** the principle of superposition and **explain** what is meant by constructive interference and by destructive interference.

4.5.6 **State** and **apply** the conditions for constructive and for destructive interference in terms of path difference and phase difference.

4.5.7 **Apply** the principle of superposition to determine the resultant of two waves.

We are surrounded by many kinds of waves and, of course, the paths of these waves cross each other all the time. When different waves cross, or 'meet', they usually pass through each other without any significant effect, but if the waves are similar to each other (in amplitude and wavelength), then the results can be important. This effect is known as the **interference** of waves.

Superposition of waves

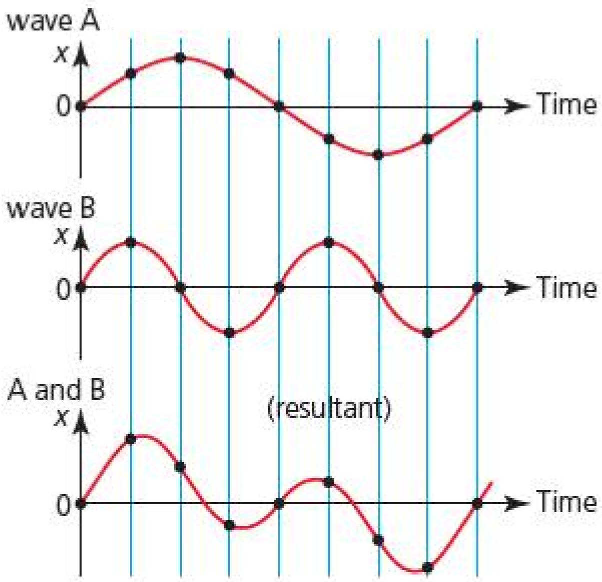


Figure 4.61 Adding wave displacements using the principle of superposition

In general, we can predict what will happen when waves meet by using the principle of **superposition**.

The principle of superposition states that at any moment the overall displacement at any point will be the vector sum of all the individual wave displacements.

This is illustrated by Figure 4.61. If wave A and wave B meet at a point, the resulting disturbance at any time is determined by adding the two individual displacements at that moment.

In this example waves A and B have different frequencies, but in the rest of this section we will only deal with the combination of two waves of the same frequency.

Constructive and destructive interference

Consider Figure 4.62. Suppose that waves of the same frequency are emitted from sources A and B in phase, or with a constant phase difference. Such sources are described as being **coherent**.

If these waves travel equal distances and at the same speed to meet at a point such as P_0, which is the same distance from both sources, they will arrive *in phase* and, using the principle of superposition, we know that if they have the same amplitudes, the result will be an oscillation at P_0 which has twice the amplitude of the individual oscillations. This is an example of **constructive interference**.

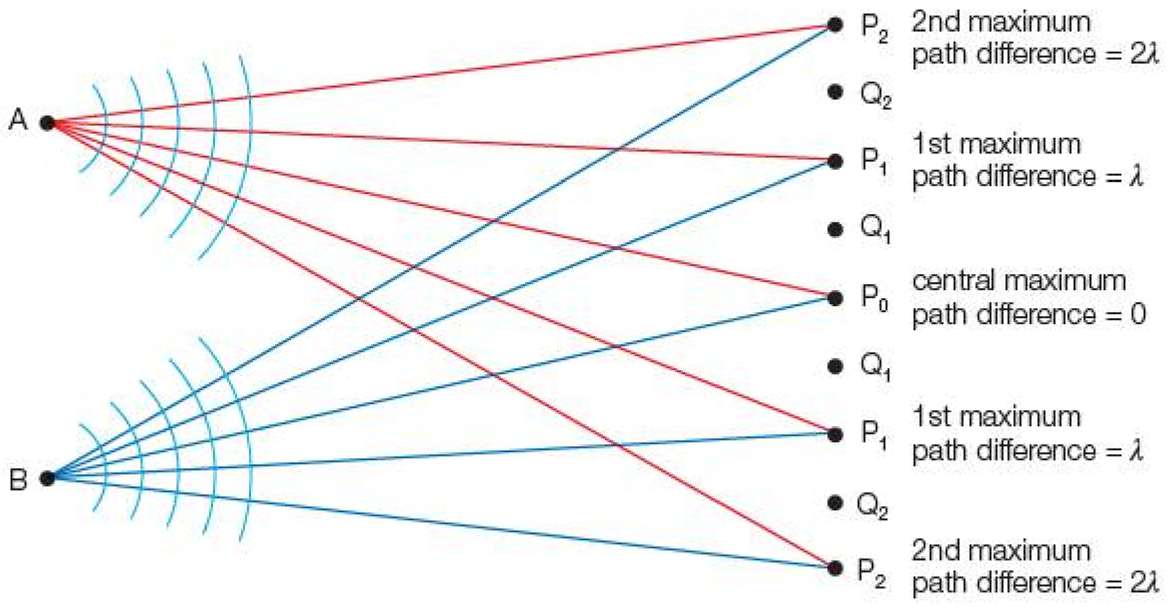


Figure 4.62 Interference and path difference

Similarly, there will be other places, such as P_1, P_2, etc. (Figure 4.63a), where the waves are *in phase* and interfere constructively because one wave has travelled one wavelength further than the other, or two wavelengths further, or three wavelengths, for example.

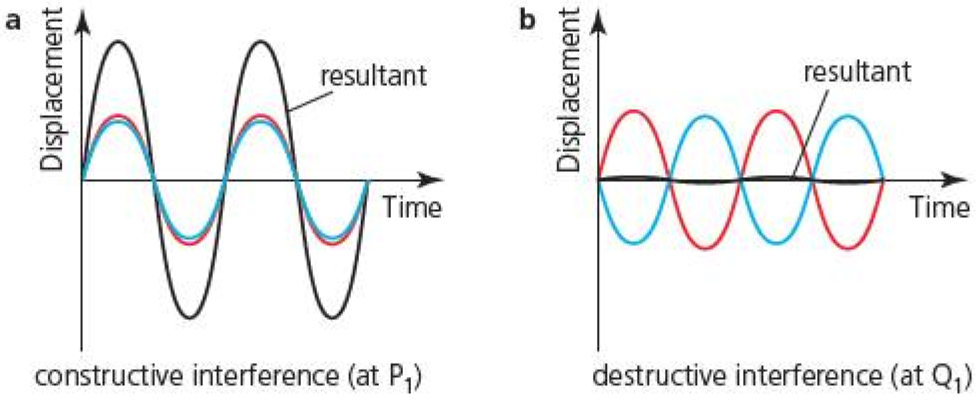


Figure 4.63 Constructive and destructive interference

In general, we can say that, under these conditions, constructive interference will occur anywhere where there is a **path difference** equal to a whole number of wavelengths. Path difference is the difference in the distance travelled by waves from two separate sources which arrive at the same point.

In other places, such as at points Q_1 and Q_2 for example, the waves will arrive *exactly out of phase* because one wave will have travelled half a wavelength further than the other, or one and a half wavelengths, or two and a half wavelengths, and so on. In these places the result will be a minimal oscillation, an effect which is called **destructive interference** (as shown in Figure 4.63b). The resultant will not be zero because one wave has a greater amplitude than the other, since they have travelled different distances. The overall pattern will appear as shown in Figure 4.64 on the next page.

The fact that there are places where waves can come together to produce (almost) no waves is especially important because *only* waves can show this behaviour. For example, when it was discovered that light can interfere, there was only one possible conclusion – light must travel as a wave.

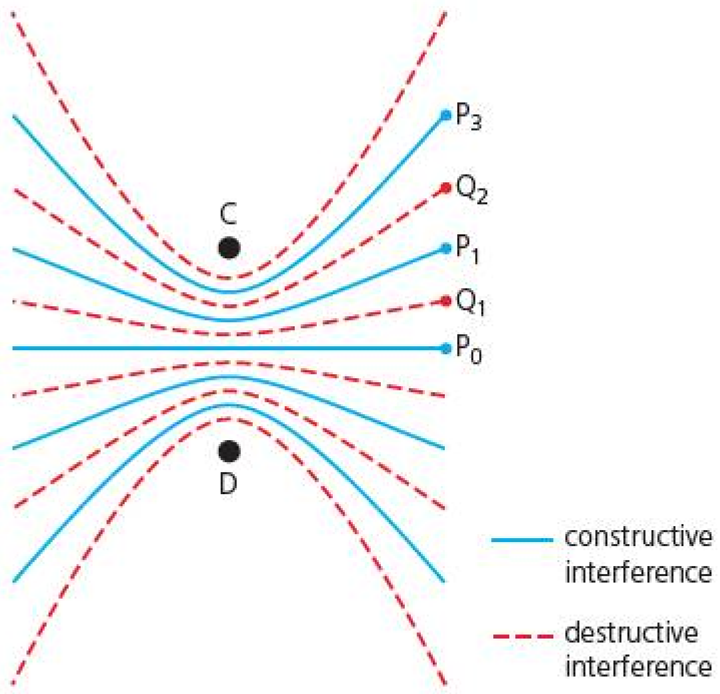


Figure 4.64 Interference pattern produced by waves from two sources, C and D

Two waves combining to give no waves at particular places may seem to contradict the principle of conservation of energy, but the 'missing' energy appears at other places in the interference pattern where there is constructive interference, giving twice the amplitude. (Remember that doubling the amplitude of an oscillation implies four times the energy.)

Summary of conditions for interference

The condition for constructive interference is that coherent waves arrive at a point in phase. This will occur when the path difference is equal to a whole number of wavelengths.

For constructive interference, path difference = $n\lambda$ (where n is an integer, 1, 2, 3, … etc).

The condition for destructive interference is that coherent waves arrive at a point exactly out of phase. This will occur when the path difference equals an odd number of half wavelengths.

For destructive interference, path difference = $\left(n + \frac{1}{2}\right)\lambda$ (where n is an integer, 1, 2, 3, … etc).

These two conditions are given in the IB *Physics data booklet*.

At most places in an interference pattern there is neither perfectly constructive nor perfectly destructive interference, but something in between these two extremes.

Examples of interference

Figure 4.65 Interference of microwaves

Different sources of waves (light waves, for example) are not normally coherent because waves are not produced in a coordinated way. So, although in principle all waves can interfere, in practice, examples are limited to those waves which can be made to be coherent. This normally means using a single source of waves and splitting it into two.

- *Interference of microwaves*. In Figure 4.65 a single microwave source is being used, but the wavefronts are split into two by the slits between the aluminium sheets and, if the receiver is moved around, it will detect an interference pattern.

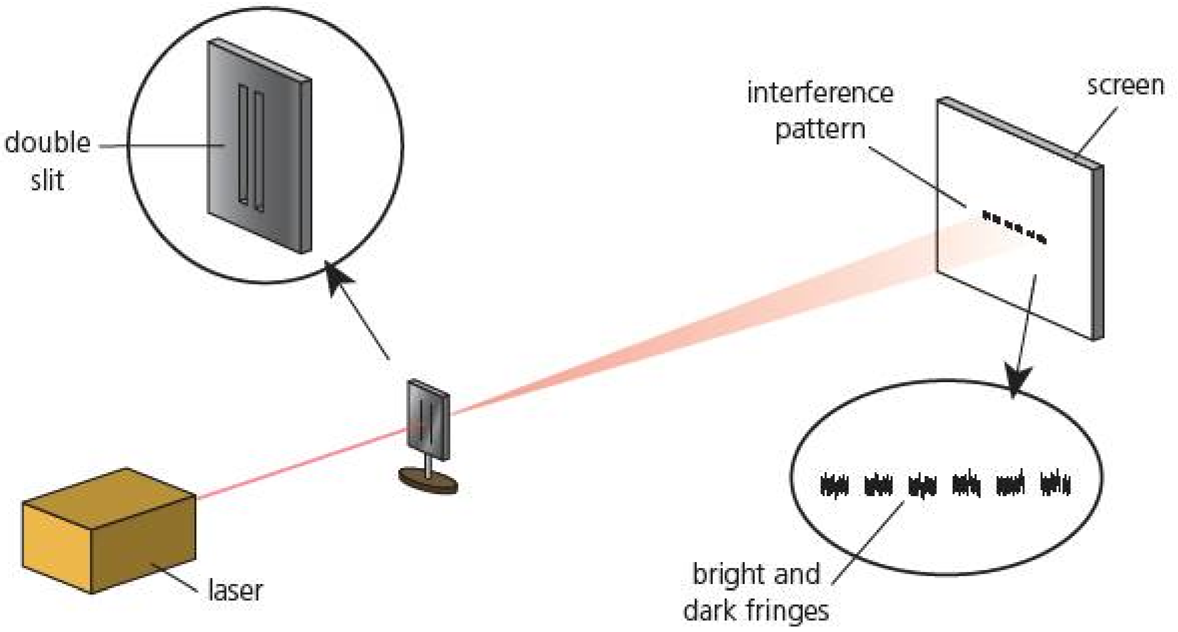


Figure 4.66 Interference of light waves

- *Interference of light waves*. Lasers provide coherent light of a single wavelength (monochromatic) and are ideal for demonstrating the interference and diffraction of light waves. See Figure 4.66.
- *Interference of sound waves*. Identical waves can be produced by two sources driven by the same electronic signal, such as with radio or microwave transmitters, or with sound loudspeakers. In Figure 4.67 the listener hears the sound intensity changing as she walks past the speakers.

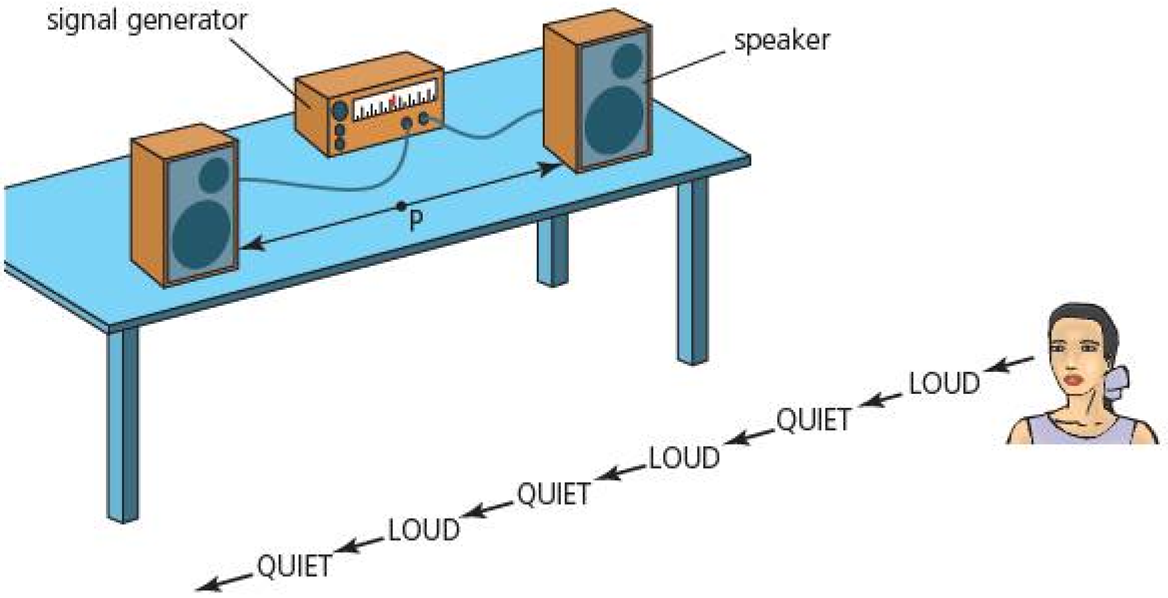


Figure 4.67 Interference of sound waves

38 Why are microwaves often used in school laboratories to demonstrate the interference of electromagnetic waves?

39 The girl shown in Figure 4.67 found that when she moved in the direction shown by the arrows, there was approximately 50 cm between successive positions where the sound was loud. The speakers were 120 cm apart and her closest distance to point P was 80 cm. Estimate the approximate wavelength and frequency of the sources.

40 Figure 4.68 shows two sources of waves on a ripple tank. The diagram is one quarter of full size. Each source produces waves of wavelength 2.8 cm. Take measurements from the diagram to determine what kind of interference will occur at P.

41 An observer stands halfway between two speakers which are facing each other and he hears a loud sound of frequency 240 Hz.
- **a** Explain why the intensity of the sound decreases if he walks slowly in any direction.
- **b** What is the shortest distance he would have to move to hear the sound rise to a maximum again? (The speed of sound in air = 340 m s^{-1}.)

42 Explain why no interference pattern is seen when the light beams from two car headlights cross over each other.

43 In Figure 4.65 the centres of the slits were 6 cm apart.
- **a** Suggest why the width of the two slits was chosen to be about the same size as the wavelength.
- **b** The receiver detected a maximum signal when it was 45 cm from one slit and 57 cm from the other. Suggest possible values for the wavelength of the microwaves.
- **c** How could you determine the actual wavelength?

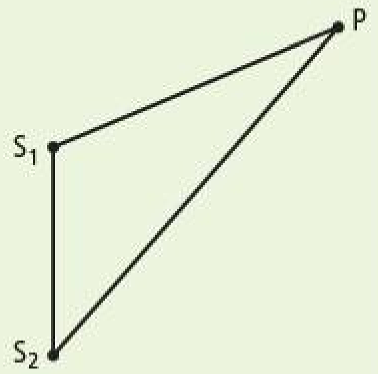


Figure 4.68

44
- **a** Sketch a displacement–time graph for 1 s for an oscillation of amplitude 4 cm and frequency 2 Hz.
- **b** On the same axes draw a graph representing an oscillation of amplitude 2 cm and frequency 4 Hz.
- **c** Use the principle of superposition to draw the resultant of these two waves.

TOK Link: Useful but incorrect theories

In 1801 Thomas Young was the first to show that light could be made to interfere. This discovery was in conflict with the theory of light at that time, generally known as Newton's 'corpuscular theory'. Newton's theory explained the then known properties of light (like reflection and refraction) by assuming it to consist of tiny particles. However, particle theory cannot be used to explain interference. Figure 4.66 shows a modern version of Young's experiment.

If light interfered it must have wave-like properties, but all the other known waves needed a medium to travel through. For example, sound can travel through air but not through a vacuum, because in a vacuum there can be no oscillating molecules to transfer the waves.

About 200 years before, in the 17th century, Descartes is credited with developing the concept of the 'ether', a mysterious, undetectable substance that was everywhere, filling all of space. Scientists adopted this idea to help explain how light can travel across space. (Descartes had earlier proposed the 'ether' to help explain the idea that forces (magnetic, gravitational, electrical) can act 'at a distance' between objects with nothing in the intermediate space.)

For about 100 years, the ether was a less than totally convincing, but widely accepted, scientific theory, but it was finally discredited by the work of Einstein in the early 20th century following the earlier discovery by Michaelson and Morley that the speed of light was the same in all directions relative to the motion of the Earth.

Question

1 Can it be good science to use a theory to explain something and make predictions, although it is known that the theory is incomplete and may even be wrong? Or, more problematically, is it acceptable to use a theory that is known to be incorrect because it can be used to make what appear to be useful predictions?

SUMMARY OF KNOWLEDGE

4.1 Kinematics of simple harmonic motion

- When objects move backwards and forwards about the same place they are said to oscillate. Oscillations occur because there are restoring forces pulling the objects back to their equilibrium positions. There are many different examples of oscillations.
- The displacement, x, of an oscillator is defined as the distance from its equilibrium position (in a specified direction). The maximum displacement, x_0, is called the amplitude of the oscillation.
- The time for one complete oscillation is called the period, T, and the number of oscillations in unit time is known as the frequency, f (unit: hertz, Hz). $f = 1/T$.
- The simplest kind of oscillation is idealized simple harmonic motion (SHM), in which the force is always proportional to the displacement (but in the opposite direction).
- SHM is defined as an oscillation in which the acceleration, a, of an object is proportional to its displacement, x, and in the opposite direction: $a = -\omega^2 x$, where ω is the angular frequency (unit: rad s^{-1}). $\omega = \Delta\theta/\Delta t = 2\pi/T$. The negative sign in the SHM equation indicates that the acceleration is in the opposite direction to the displacement.
- If oscillators are doing exactly the same thing at the same time, we say that they are in phase with each other. If they are not in phase, the phase difference is defined as the angle (usually in radians) between the oscillations.
- Graphs of displacement–time, velocity–time and acceleration–time can be drawn to represent the characteristics of SHM.
- These graphs are sinusoidal in shape. The equations for displacement–time graphs are $x = x_0 \sin \omega t$ and $x = x_0 \cos \omega t$ (depending on the starting conditions).
- The equations for velocity–time graphs are $v = v_0 \cos \omega t$, and $v = -v_0 \sin \omega t$ (depending on the starting conditions).
- The velocity of an oscillator at any position can be determined from its amplitude and angular frequency: $v = \pm\omega\sqrt{(x_0^2 - x^2)}$.

4.2 Energy changes during SHM

- All mechanical oscillations involve a continual interchange between potential energy and kinetic energy. Graphs of potential energy, kinetic energy and total energy against displacement and time represent these energy changes.
- Kinetic energy of a simple harmonic oscillator, $E_K = \frac{1}{2}m\omega^2(x_0^2 - x^2)$; total energy, $E_T = \frac{1}{2}m\omega^2 x_0^2$; potential energy, $E_P = \frac{1}{2}m\omega^2 x^2$.

4.3 Forced oscillations and resonance

- When they are disturbed, most objects and structures have a natural frequency (or frequencies) at which they will oscillate.
- Frictional forces reduce the speeds and amplitudes of oscillation and dissipate energy. This process is called damping, and it can occur to very different degrees in different systems. It may be desirable or unwanted. Critical damping occurs when a system returns relatively quickly to its equilibrium position without passing through it.
- Objects are commonly exposed to forces from external vibrations. If the external force has the same frequency as the natural frequency, energy is efficiently transferred and amplitude increases. This is called resonance.
- The degree of damping determines whether any resonance effects are significant. Frequency–response graphs can be drawn to represent resonance in systems with different degrees of damping.
- There are many examples of useful resonance but also examples of destructive resonance.

4.4 Wave characteristics

- Oscillations (which may be approximately simple harmonic) in one part of a medium can transfer energy to their surroundings and, in this way, a progressive (travelling) wave or pulse can be propagated away from their source and through the medium. Such a wave will transfer energy, but there is no net motion of the medium itself.
- In transverse waves the oscillations are perpendicular to the direction of energy transfer. In longitudinal waves the oscillations are parallel to the direction of energy transfer. Electromagnetic waves and waves on strings are transverse. Sound is a longitudinal wave.
- The tops and bottoms of transverse waves are called crests and troughs. In a longitudinal wave the locations where the medium is squashed (compressed) are called compressions and places where the medium is stretched, or at low pressure, are called rarefactions.
- Wavelength, λ, is the shortest distance between two points moving in phase.
- On paper we can represent waves with wavefronts, and the direction in which they are moving is shown by rays. A wavefront is a line joining adjacent points that are moving in phase. Adjacent wavefronts are one wavelength apart.
- Displacement–time and displacement–position graphs can be drawn to represent wave motion.
- The equation $v = f\lambda$ can be used with all kinds of waves.
- The intensity of a wave is proportional to its amplitude squared. (More generally, intensity = power/area.)
- The electromagnetic spectrum contains waves of very different wavelengths and properties, but they all consist of oscillating electric and magnetic fields that do not need a medium to travel through. All electromagnetic waves travel across free space at the same speed. The main sections of the electromagnetic spectrum, with their approximate wavelengths are often displayed in an ordered list.

4.5 Wave properties

- All waves reflect, refract, diffract and interfere under suitable conditions.
- Reflection occurs when a wave meets a boundary between two different materials. Sometimes some of the wave energy is also transmitted into the second material. Waves are reflected with a phase change of π from a boundary with a medium in which they travel slower.
- When waves are transmitted obliquely into a different medium the change of speed results in a change of direction called refraction.
- Snell's law ($n_1/n_2 = \sin\theta_2/\sin\theta_1 = v_2/v_1$) can be used to predict the direction of a wave after refraction. When a wave passes into a medium in which it has a slower speed, it is refracted towards the normal and its wavelength is reduced.
- Diffraction is the spreading of waves as they pass gaps and obstacles. This is most significant when the gap or obstacle size is similar in magnitude to the wavelength. It is possible to provide examples of the diffraction of all kinds of waves.
- When two (or more) waves come to the same point at the same time, the result can be predicted using the principle of superposition. But the effect will not be significant unless the wave sources are coherent (a constant phase difference).
- When coherent waves arrive at a point in phase they will interfere constructively. If they are perfectly out of phase destructive interference will occur. Path difference = $n\lambda$ is the condition for constructive interference. Destructive interference occurs if the path difference = $\left(n + \frac{1}{2}\right)\lambda$.

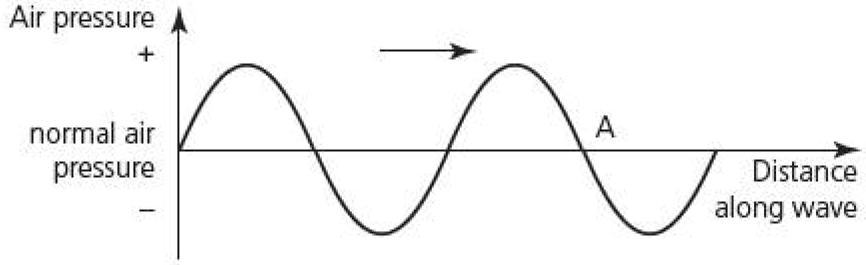
Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 A transverse wave is travelling along a stretched rope. Consider two points on the rope which are exactly half a wavelength apart. The velocities of the rope at these two points are always

A the same as each other
B constant
C in opposite directions
D in a direction which is parallel to the direction in which the wave is travelling.

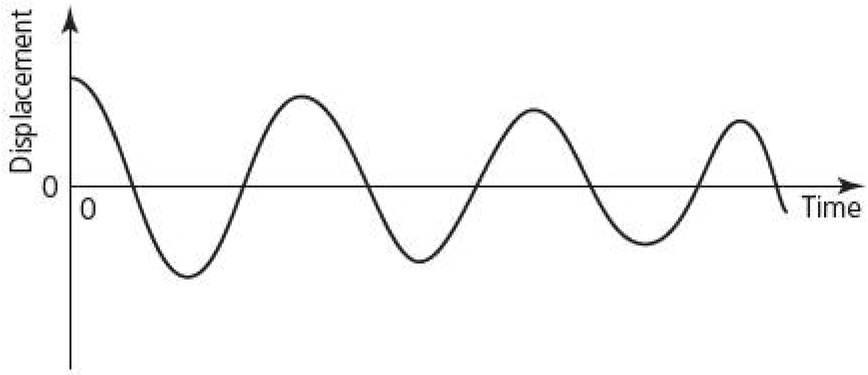
Q2 The graph represents a sound wave moving through air. The wave is moving in the direction shown by the arrow. Consider point A.



The pressure of the air at A is:

A zero
B always constant
C about to increase
D about to decrease.

Q3 The motion of an object oscillating with a constant time period is shown in the graph below.



From this graph it is possible to determine that

A the oscillations are not damped.
B the oscillations are lightly damped.
C the oscillations are critically damped.
D the oscillations are over-damped.

Q4 During one complete oscillation, the amplitude of a **damped** harmonic motion changes from 1.5 cm to 0.30 cm. The total energy at the end of the oscillation is E_2 and the total energy at the beginning of the oscillation is E_1. The ratio $\frac{E_2}{E_1}$ is

A $\frac{1}{5}$

B $\frac{1}{25}$

C 5

D 25

Standard Level Paper 1, May 09 TZ2, Q13

Q5 A transverse progressive wave is being transmitted through a medium. The wave speed is best described as

A the speed at which energy is being transferred.
B the average speed of the oscillating particles in the medium.
C the speed of the medium.
D the speed of the source of the waves.

Q6 Light waves of wavelength λ, frequency f and travelling at speed c, enter a transparent material of refractive index 1.6. Which of the following describes the wave properties of the transmitted light waves?

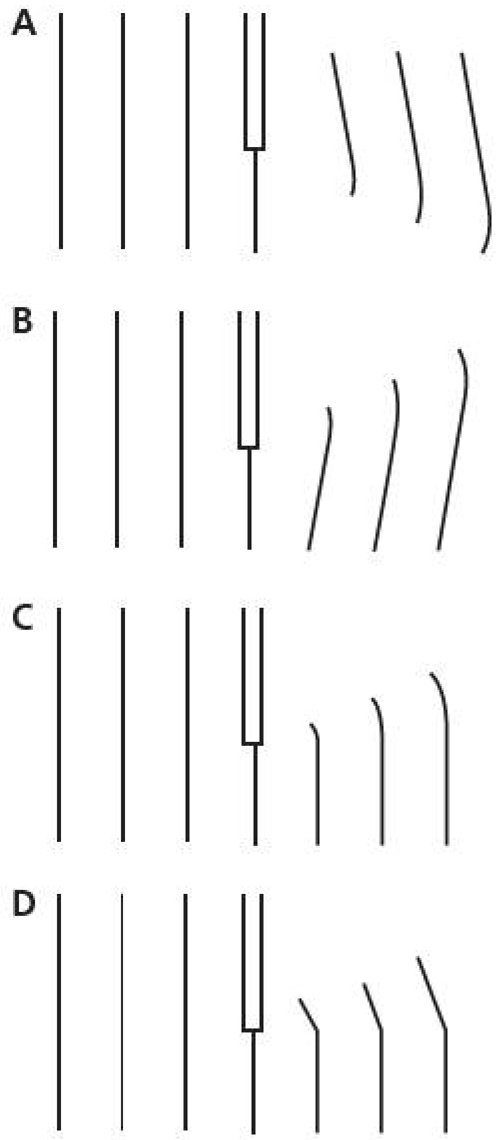
A wavelength 1.6λ; speed c

B frequency $1.6f$; speed $\frac{c}{1.6}$

C wavelength $\frac{\lambda}{1.6}$; frequency f

D wavelength λ; speed $\frac{c}{1.6}$

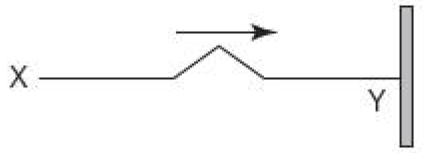
Q7 When plane wavefronts meet an obstacle they may be diffracted. Which of the following diagrams is the best representation of this effect?



Q8 Which of the following is a possible definition of the wavelength of a progressive (travelling) transverse wave?

A the distance between any two crests of the wave
B the distance between a crest and an adjacent trough
C the amplitude of a particle in the wave during one oscillation of the source
D the distance moved by a wavefront during one oscillation of the source

Q9 The diagram shows a pulse travelling along a rope from X to Y. The end Y of the rope is tied to a fixed support.



When the pulse reaches end Y it will

A disappear.

B cause the end of the rope at Y to oscillate up and down.

C be reflected and be inverted.

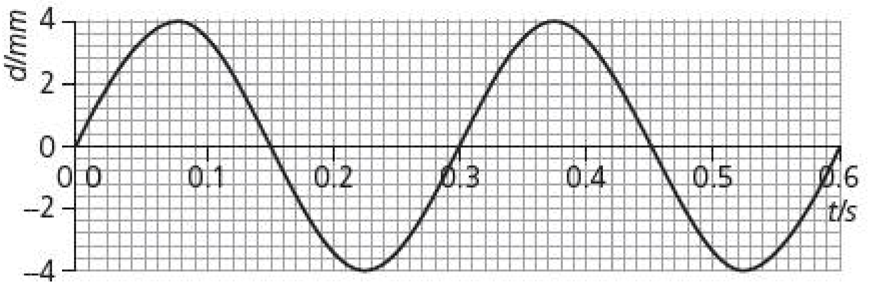
D be reflected and not be inverted.

Standard Level Paper 1, Specimen Paper 09, Q15

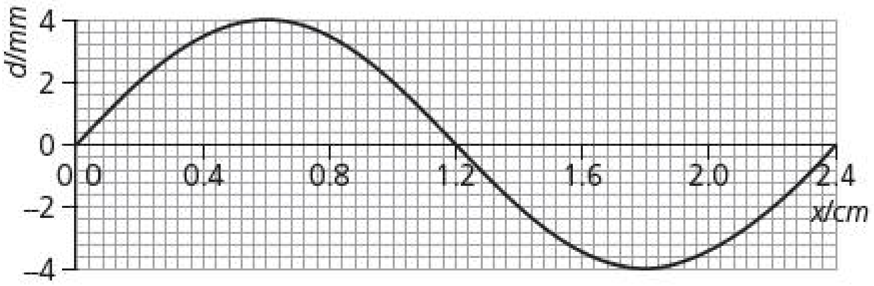
Paper 2 IB questions and IB style questions

Q1 a Graph 1 below shows the variation of time t with the displacement d of a travelling (progressive) wave. Graph 2 shows the variation with distance x along the same wave of its displacement d.

Graph 1



Graph 2



i State what is meant by a travelling wave. [1]

ii Use the graphs to determine the amplitude, wavelength, frequency and speed of the wave. [4]

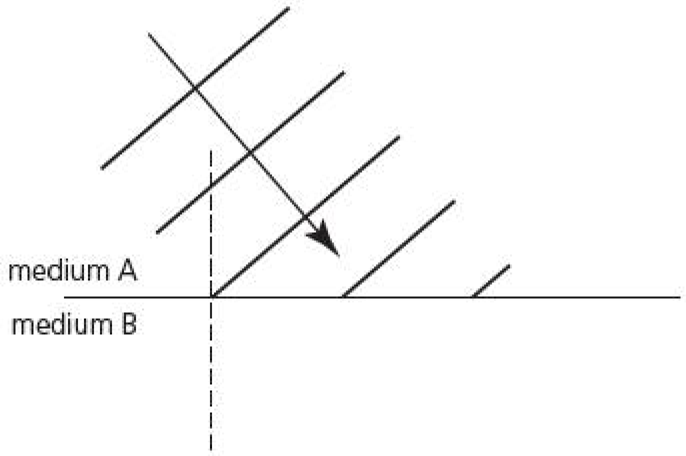
b The diagram shows plane wavefronts incident on a boundary between two media A and B.

The ratio $\dfrac{\text{refractive index of medium B}}{\text{refractive index of medium A}}$ is 1.4.

The angle between an incident wavefront and the normal to the boundary is 50°.

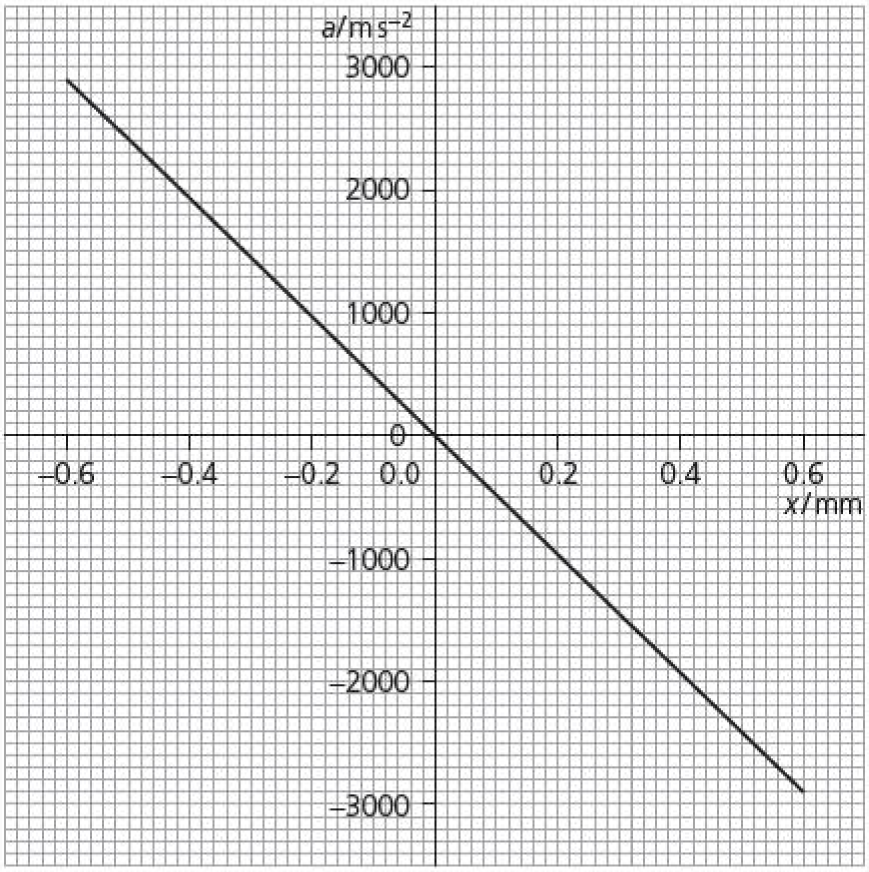
i Calculate the angle between a refracted wavefront and the normal to the boundary. [3]

ii Copy the diagram and construct **three** wavefronts to show the refraction of the wave at the boundary. [3]



Standard Level Paper 2, Nov 06, QB2 (Part 1)

Q2 An object is vibrating in air. The variation with displacement x of the acceleration a of the object is shown below.



a State and explain **two** reasons why the graph indicates that the object is executing simple harmonic motion. [4]

b Use data from the graph to show that the frequency of oscillation is 350 Hz. [4]

c The motion of the object gives rise to a longitudinal progressive (travelling) sound wave.

i State what is meant by a longitudinal progressive wave. [2]

ii The speed of the wave is $330\,m\,s^{-1}$. Using the answer in **b**, calculate the wavelength of the wave. [2]

Standard Level Paper 2, May 09, TZ2, QB2 (Part 1)

5 Electric currents

STARTING POINTS

- Atoms have a central nucleus containing protons and neutrons. Surrounding the nucleus there are electrons.
- Protons are positively charged. Electrons are negatively charged. Neutrons have no charge (they are neutral). There are forces between charged particles. Opposite charges attract each other. Like charges repel each other.
- Atoms, molecules and ions in solids vibrate about fixed positions.
- Kinetic energy can be calculated from $\frac{1}{2}mv^2$
- Power is the rate of transfer of energy. Power = energy transferred/time taken.

5.1 Electric potential difference, current and resistance

Charge and current

Electric charge is a fundamental property of some sub-atomic particles. All **protons** have a positive charge of $+1.60 \times 10^{-19}$ C and all **electrons** have a negative charge of -1.60×10^{-19} C. The charge on an electron is commonly represented by the letter *e*. This value for the basic quantity of charge is given in the IB *Physics data booklet*. One coulomb of negative charge is carried by 6.2×10^{18} electrons $[1/(1.60 \times 10^{-19})]$.

Charge is generally given the symbol *Q*, but *q* is also commonly used to represent the charge on particles. Charge is measured in coulombs, C. (Charge is covered in greater detail in Chapter 6.)

If large numbers of charged particles can be made to *flow* in one direction, they can be used to transfer energy. A flow of electric charge is called an electric **current**. Currents can be made to flow around complete electrical **circuits**. The wires and components of an electrical circuit are good **conductors** of electricity because they have many **'free' electrons**. Atoms, ions and molecules cannot move around freely in solids, so a flow of positively charged protons is not possible, but 'free' electrons are not attached to any particular atom.

When electrons are added or removed from a neutral atom or molecule, it is called an **ion**. Ions can flow through liquids and gases and this property can be very useful.

Figure 5.1 shows free electrons moving through a metal. This represents an electric current flowing through a conductor. Even when a current is not flowing, the free electrons are still moving around randomly at very high speeds, like molecules in a gas.

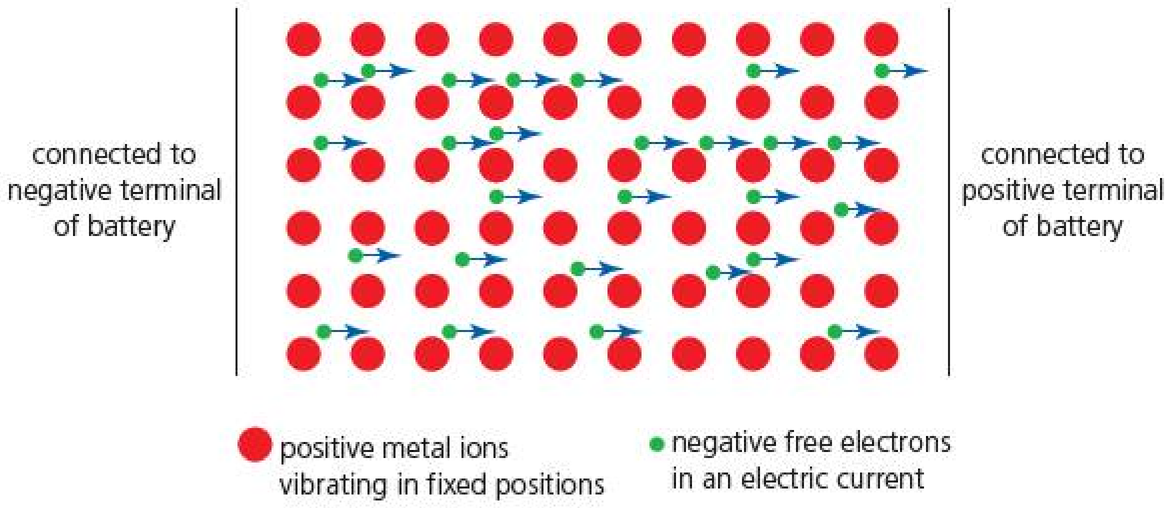


Figure 5.1 Electric current flowing through a metal

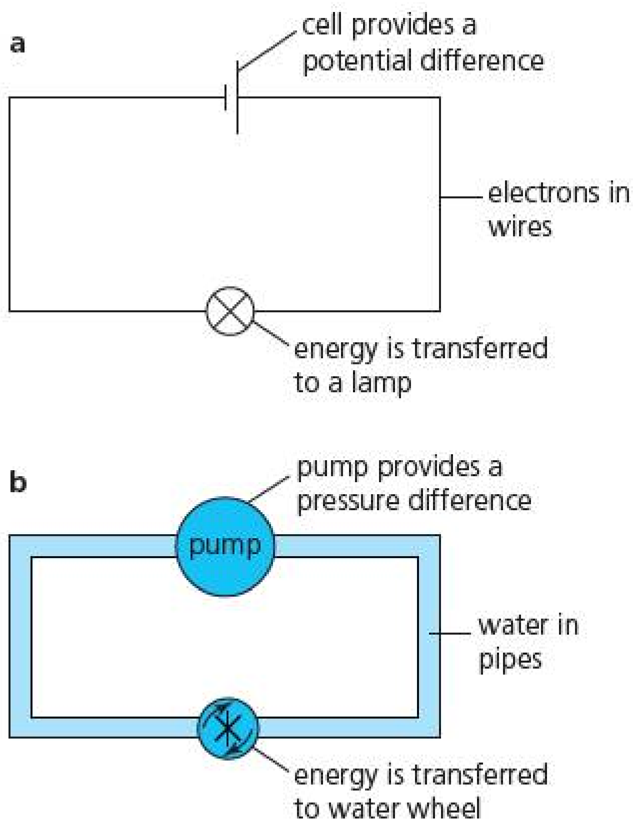


Figure 5.2 a An electric circuit can be compared to **b** a water circuit

In all electrical circuits, electric potential energy is transferred along wires by moving electrons. The energy is transferred from the energy source (such as a cell or battery) to one or more components in the circuit. These components transfer the electrical energy into another form of energy, which can be put to good use. For example, a battery in a torch transfers chemical energy to electric potential energy, and that energy is then transferred to light and thermal energy in the torch bulb. Remember that electrons are present all around circuits all the time; they do not originate in the battery.

It may be useful to compare a battery in an electrical circuit to a pump forcing water around pipes in a water circuit, as shown in Figure 5.2b. The pump provides a *pressure* difference which moves the water. The water is then able to do useful work (maybe turning a water wheel) as it moves around the circuit. In an electrical circuit (Figure 5.2a), the battery is considered to provide a *potential* difference.

Electric potential difference

5.1.1 Define *electric potential difference.*

The amount of energy transferred to or from charges (electrons) in any part of a circuit is represented by the **potential difference (p.d.)**. The symbol for p.d. is V. Potential difference is a key concept in all work on electrical circuits. It is also commonly called **voltage**. We refer to the p.d. (voltage) *across* devices, not through them, because it tells us about the difference between two points in a circuit.

The electric potential difference between two points is defined as the electric potential energy transferred as a unit charge moves between the points:

$$\text{potential difference} = \frac{\text{energy transferred (work done)}}{\text{charge}}$$

$$V = \frac{W}{q}$$

The unit of p.d. is the **volt**, V ($1\,\text{V} = 1\,\text{J}\,\text{C}^{-1}$).

Batteries, and other electrical energy sources, provide potential differences across the circuits in which they are connected. The positive terminal of a battery attracts electrons and the negative side repels them. A battery is the name we give to two or more **cells** joined together.

In Figure 5.3, a 1.5 V cell is transferring 1.5 J of energy to each coulomb of charge that flows through it. If the battery is connected to identical two lamps, each will have a p.d. of 0.75 V across it and energy will be transferred from electricity to light and thermal energy at a rate of 0.75 J for every coulomb flowing through them.

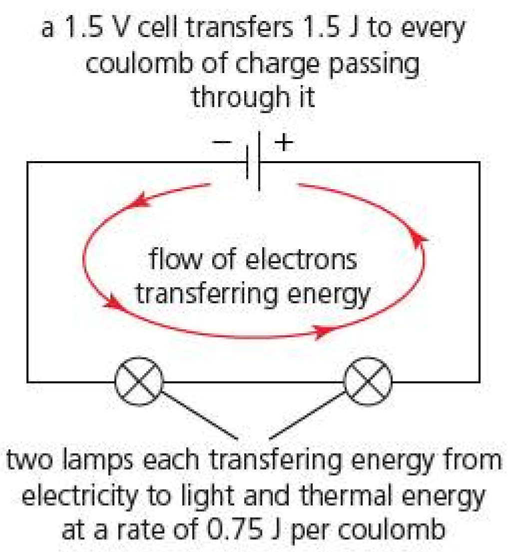


Figure 5.3 Energy transfers in a simple circuit

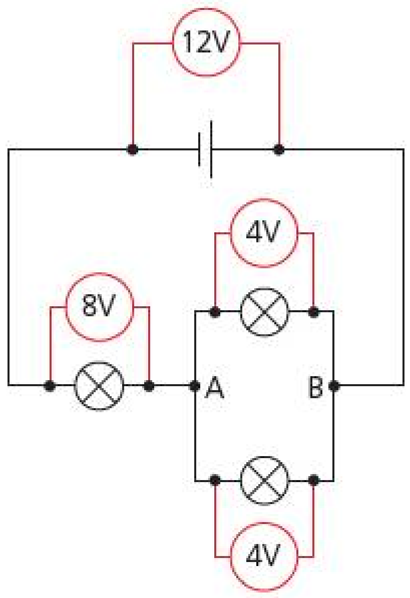


Figure 5.4 Measuring potential differences with voltmeters

5.1.4 Solve problems involving electric potential difference.

There will be a potential difference across any component in a circuit which is transferring energy. To measure a potential difference between any two points in an electrical circuit, a **voltmeter** is placed *in parallel* across the component, as shown in Figure 5.4.

The two voltmeters reading 4 V in Figure 5.4 are measuring the same p.d. (the p.d. between points A and B), so one of the voltmeters is not needed. The battery is transferring 12 J to every coulomb, so the same amount of energy (per coulomb) must be transferred in the circuit. This means that the sum of the p.d.s between points around the circuit must add up to the p.d. of the battery. There is no significant p.d. across the connecting wires (leads) because they are designed not to transfer energy to other forms (such as thermal energy).

1 a How much energy is supplied to 3.0 C of charge as it flows through a 12 V battery?
b If the charge flows around the circuit at a rate of 1.2 C s^{-1}, how much energy is transferred to the circuit in exactly one minute?

2 a What is the p.d. across a water heater if 44 000 J of energy is transferred to internal energy when 200 C of charge flows through it?
b What charge has passed through a 1.5 V cell if 60 J of energy were transferred?

3 What are the readings on voltmeters P and Q in Figure 5.5? (The lamps are not identical.)

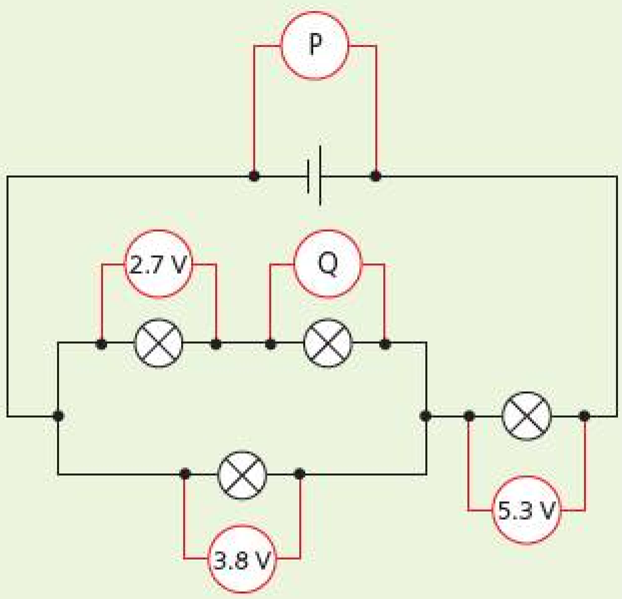


Figure 5.5

Accelerating charged particles across a vacuum

5.1.2 Determine the change in potential energy when a charge moves between two points at different potentials.

Charges may move between potential differences under a wide range of circumstances, not just around circuits with good conductors. For example, electrons (emitted from a very hot conductor) can be made to accelerate across a vacuum by a potential difference, as shown in Figure 5.6. (In later chapters we will discuss the properties and behaviour of electrons and other sub-atomic particles moving in particle beams.) As an electron moves from A (called the **cathode**) to B (the **anode**), it loses electric potential energy and gains kinetic energy $\left(\frac{1}{2}mv^2\right)$.

From the definition of potential difference we can see clearly that:

energy transferred or work done = potential difference × charge

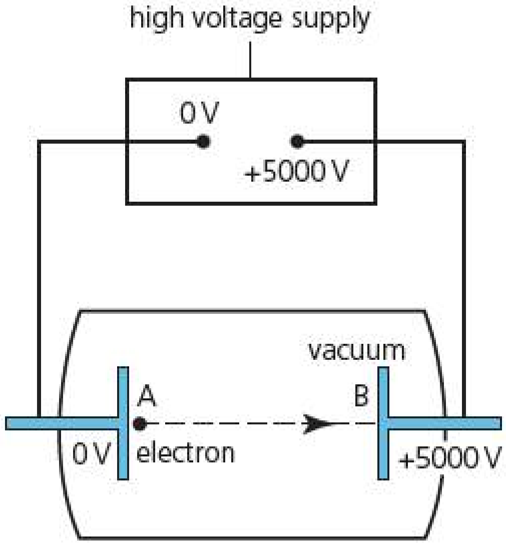


Figure 5.6 Accelerating electrons in a vacuum

$W = Vq$ in general

$Vq = \frac{1}{2}mv^2$

For an election $W = Ve$

so that

$$Ve = \frac{1}{2}mv^2$$

This equation is is given in the IB *Physics Data booklet*.

So, the change of electrical potential energy of a single electron accelerated by a p.d. of +5000 V, for example, equals $5000 \times (-1.6 \times 10^{-19}) = -8.0 \times 10^{-16}$ J. In a vacuum it can be assumed that the total energy of the electron remains constant, so that the gain of kinetic energy is $+8.0 \times 10^{-16}$ J.

The electronvolt: a unit of energy

5.1.3 Define the *electronvolt*.

When referring to the energies of particles, physicists often avoid the kind of calculation performed above (as well as the small numerical values of particle energies in joules) by simply saying that the electron has gained 5000 **electronvolts** of kinetic energy. When an electron, or any other particle with the same sized charge (for example, a proton), is accelerated by 5000 V it gains 5000 electronvolts of kinetic energy. A *doubly* charged ion accelerated by 5000 V would gain 10 000 electronvolts. The electronvolt is a widely used unit of energy when discussing sub-atomic particles of any kind (not just electrons).

One electronvolt is defined as the kinetic energy that would be gained by an electron if it was accelerated by a potential difference of one volt.

The symbol for the electronvolt is **eV**; keV and MeV are also in common use.

energy transferred (J) = energy transferred (eV) $\times (1.6 \times 10^{-19})$

Worked examples

1 How much energy is gained by a proton accelerated by a p.d. of 3500 V? Give your answer in:
 a electronvolts
 b joules.

 a 3500 eV
 b $3500 \times (1.6 \times 10^{-19}) = 5.6 \times 10^{-16}$ J

2 How many keV are transferred to an ion of charge 3.2×10^{-19} C when it is accelerated by a p.d. of 4000 V?

$$\left(\frac{3.2 \times 10^{-19}}{1.6 \times 10^{-19}}\right) \times 4.0 = 8.0\,\text{keV}$$

3 An electron of mass 9.1×10^{-31} kg is accelerated by a p.d. of 3000 V.
 a Calculate its final speed.
 b What assumption(s) did you make?

 a Energy transferred to electron = kinetic energy

$$Ve = \frac{1}{2}mv^2$$

$$3000 \times (1.6 \times 10^{-19}) = \frac{1}{2} \times (9.1 \times 10^{-31}) \times v^2$$

$$v = 3.2 \times 10^7\,\text{m s}^{-1}$$

 b This calculation assumed that the electron started with negligible kinetic energy and that no kinetic energy was transferred away from it during the acceleration (because it was in a vacuum).

5.1.4 Solve problems involving electric potential difference.

4 An electron is moving directly towards a negatively charged plate with a speed of $1.30 \times 10^7\,\text{m s}^{-1}$.
 a What is the kinetic energy of the electron in:
  i joules
  ii electronvolts?
 b If the voltage on the plate is –500 V, will the electron reach the plate? Explain.

5 **a** A particle emitted during radioactive decay had an energy of 2.2 MeV. What was its energy (in joules)?
 b If its mass was 6.8×10^{-27} kg, what was its speed?

6 Calculate the gain in kinetic energy (in joules) of a proton which is accelerated across a vacuum by a p.d. of 5000 V.

Electric current and resistance

5.1.5 Define *electric current*.

Electric current

As we have already discussed, an electric current exists wherever there is a flow of charged particles. More precisely, current (given the symbol *I*) is the rate of flow of charge:

$$\text{current} = \frac{\text{charge flowing past a point}}{\text{time}}$$

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

This equation is given in the IB *Physics data booklet*.

The unit of current is the **ampere**, A, which is usually shortened to **amp** ($1\,A = 1\,C\,s^{-1}$). That is, if one coulomb of charge passes a point in one second, the current is one amp. mA and μA are also in common use.

The ampere is one of the fundamental units of the SI system. It is defined in terms of the force per unit length between current-carrying conductors (see Chapter 6).

The coulomb is defined in terms of the amp: one coulomb is the charge flowing per second if the current is one amp.

A current which always flows in the same direction around a circuit is called a **direct current** (**dc**). Batteries and cells supply dc. If a current repeatedly changes direction, it is called an **alternating current** (**ac**). Often dc is more useful than ac, but most electrical energy is transferred around countries using alternating currents and high voltages because, in that way, less energy is wasted to thermal energy (see Chapter 12).

We often place an arrow on a circuit showing the direction of flow of a direct current but, obviously, this cannot be done for alternating currents. For reasons of consistency, direct electric current is *always* shown flowing around the circuit from the positive terminal of the energy source to the negative terminal, which would be the direction in which positive charges move. Since electric currents in solids are flows of negatively charged electrons, the arrows are in the *opposite direction to the movement of electrons* (as shown in Figure 5.7).

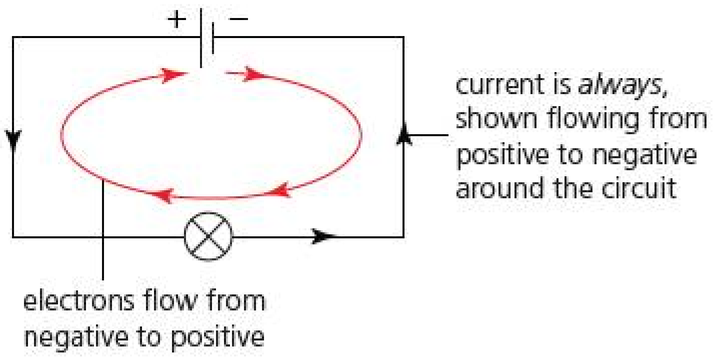


Figure 5.7 Conventional current flow is from positive to negative around the circuit

TOK link: Positive and negative

'Atoms consist of a central nucleus containing positively charged protons, with negatively charged electrons around the outside'. But, is this a scientific *fact*? If, instead, scientists said that electrons were positively charged and protons were negatively charged, it would not make any difference to our understanding of the universe. Negative protons and positive electrons *can* exist – they are called antimatter.

The use of the terms *positive* and *negative* is just a way of describing the way in which they interact (such as positive attracts negative); reversing the labels does not change the meaning. But without using terms like these we would not be able to easily transfer our knowledge of fundamental physics to other people.

By agreement, the 'conventional' direction of electric current was chosen to be from positive to negative more than 200 years ago, a long time before the nature of an electric current was understood. After the discovery that the currents in metallic conductors are actually a flow of negatively charged electrons (from negative to positive) there was no need to change the conventional direction that had been in use for a long time, and the same convention remains in use today, despite the fact that it is not strictly 'true'. (But, remember that some electric currents *are* a flow of positive ions.)

Question

1 Suggest a possible reason why scientists might need to agree on a direction of flow for electric currents.

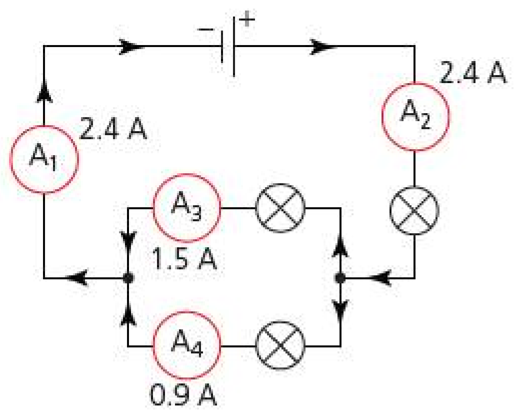


Figure 5.8 Measuring currents with ammeters

To measure the currents passing through any points in a circuit, they must be made to pass through **ammeters**, as shown in Figure 5.8. The current arriving at any point must be equal to the current leaving the same point.

This means that the readings on ammeters A_1 and A_2 must be the same as each other, which must be the same as the sum of the readings on ammeters A_3 and A_4.

5.1.11 Solve problems involving potential difference, current and resistance.

7 a How much charge flows through a lamp in one second if the current is 0.25 A?
b How many electrons will have flowed through the lamp in this time?

8 a How much charge has passed through a battery in one hour if the current is 120 mA?
b If the battery supplies a p.d. of 1.5 V, how much energy is transferred in exactly one hour?

Resistance

5.1.6 Define *resistance*.

We have seen that a potential difference from an energy source is needed to make a current flow around a circuit. The next question to ask is, 'how does the size of a current depend on the size of the p.d?' The simple answer is that the current for a particular p.d. depends on the **resistance** of the component or the wire.

Resistance is defined as the ratio of the potential difference across a conductor to the current through it. Resistance is given the symbol R.

$$\text{resistance} = \frac{\text{potential difference}}{\text{current}}$$

$$R = \frac{V}{I}$$

Figure 5.9 Fixed resistors

The unit of resistance is the **ohm**, Ω. If a p.d. of 1 V produces a current of 1 A, then the resistance is 1 Ω; kΩ and MΩ are also in common use.

A component that has been made to have a certain resistance is called a **resistor**. Resistors may have a fixed value or can be variable. Some examples of fixed resistors are shown in Figure 5.9.

To determine the resistance of a component or wire it is necessary to measure the current in it for a known potential difference. A calculation made using alternating currents and voltages will produce an answer with the same resistance as a calculation made with direct currents.

Worked example

4 a Calculate the resistance of a component if a p.d. of 8.7 V across it produces a current of 0.15 A through it.
b If the p.d. is increased to 12.4 V, the current rises to 0.17 A. What is the new resistance?

a $R = \frac{V}{I}$

$R = \frac{8.7}{0.15}$

$R = 58\,\Omega$

b $R = \frac{V}{I}$

$R = \frac{12.4}{0.17}$

$R = 73\,\Omega$

5.1.11 Solve problems involving potential difference, current and resistance.

9 What voltage is needed to make a current of 560 mA pass through a 670 Ω resistor?

10 What is the resistance of a 230 V domestic water heater if the current through it is 8.4 A?

11 What current flows through a 37 kΩ resistor when there is a p.d. of 4.5 V across it?

12 What is the p.d. across a 68.0 Ω resistor if 120 C of charge flows through it in 60.0 s?

Factors which affect resistance

5.1.7 Apply the equation for resistance in the form $R = \frac{\rho L}{A}$ where ρ is the resistivity of the material of the resistor.

A material has electrical resistance because any electrons flowing through it have to pass atoms or ions vibrating in their path (see Figure 5.1). The longer the conductor, the more 'collisions' are likely to occur and so the greater the resistance. The wider the conductor, the easier it is for electrons to flow and so the lower the resistance. More precisely, the resistance of a metal wire is proportional to its length, L, and is inversely proportional to its cross-sectional area, A:

$$R \propto \frac{L}{A}$$

But, resistance also depends on the particular material being used. The more free electrons (per unit volume) in a material, the better it will be at conducting an electric current. We cannot simply look up the resistance of a material (like copper, for example) in a table of data, because resistance depends on shape as well as the material itself. Instead, we look up a material's **resistivity**, which is the resistance of a specimen of length 1 m and cross-sectional area 1 m^2.

Not surprisingly (for a theoretical wire of cross-sectional area 1 m^2), the resistivity of a good conductor has a very low numerical value. For example, the resistivity of copper is 1.7×10^{-8} Ω m at 20 °C.

Resistivity is given the symbol ρ. It has the unit ohm metre, Ω m (*not* ohms per metre).

$$R = \frac{\rho L}{A}$$

This equation is listed in the IB *Physics data booklet*.

Resistance and resistivity usually change with temperature, sometimes significantly, and we should be careful about describing a material as an insulator or a conductor without also stating the temperature. Materials can vary enormously in their resistivities, as is shown by Table 5.1. The semiconducting elements silicon and germanium are in the middle of the range.

Table 5.1 Resistivities of various substances at 20 °C

Material	Resistivity/Ω m
Copper	1.7×10^{-8}
Aluminium	2.8×10^{-8}
Iron	1.0×10^{-7}
Nichrome (used for electric heaters)	1.1×10^{-6}
Carbon (amorphous: non-crystalline)	$\approx 10^{12}$
Germanium	4.6×10^{-1}
Sea water	$\approx 2 \times 10^{-1}$
Silicon	6.4×10^{2}
Glass	$\approx 10^{12}$
Quartz	$\approx 10^{17}$
Teflon (PTFE)	$\approx 10^{23}$

Worked examples

5 a Calculate the resistance of a 1.80 m length of iron wire of cross-sectional area 2.43 mm².
b A current of 2.4 A flowed through a 83 cm length of metal alloy wire of area 0.54 mm² when a p.d. of 220 V was applied across its ends. What was the resistivity of the alloy?

a $R = \dfrac{\rho L}{A}$

$R = \dfrac{(1.0 \times 10^{-7}) \times 1.80}{(2.43 \times 10^{-6})}$

$R = 7.4 \times 10^{-2}\,\Omega$

b $R = \dfrac{V}{I} = \dfrac{220}{2.4} = 91.7\,\Omega$

$\rho = \dfrac{RA}{L}$

$\rho = \dfrac{91.7 \times (5.4 \times 10^{-7})}{0.83}$

$\rho = 6.0 \times 10^{-5}\,\Omega\,\text{m}$

6 Suggest possible reasons why aluminium is usually used for the cables carrying large electric currents around the world (rather than copper).

The resistivities of these two metals are in the ratio of 2.8/1.7 = 1.6. This means that an aluminium cable conducts equally well as a copper wire if its cross-sectional area is 1.6 times greater. Aluminium is a better choice because it is much cheaper than copper, and aluminium cables (of equal resistance) are lighter in weight.

Copper has other properties which make it a better choice for household wiring, such as its flexibility. Some more expensive metals like silver, gold and platinum, are better conductors than copper and they are sometimes used where a very low resistance wire or connection is needed (for example, in some sound reproduction systems).

5.1.11 Solve problems involving potential difference, current and resistance.

13 a What length of nichrome wire of cross-sectional area 0.0855 mm² is needed to make a 15.0 Ω resistor?
b What value resistor would be made with nichrome wire of double the length and half the diameter?

14 Suggest the possible relationship between the resistivity of a metal and the number of free electrons in it per cubic millimetre.

15 The nichrome heating element in an electric kettle has an overall length of 5.32 m. If its resistance is 25 Ω, what is the diameter of the wire?

16 Calculate the resistivity of a metal wire if a voltage of 2.5 V is needed to make a current of 26 mA pass through a wire of diameter 0.452 mm and length 745 cm.

Additional Perspectives

How fast does electricity flow?

This interesting question has more than one answer. A current in a metal wire is a flow of free electrons, so one way to answer this question is by considering the speed of the electrons. Even in a metal without a current, the free electrons move randomly like the molecules of a gas, only faster. A typical random speed might be $10^6\,\text{m s}^{-1}$. When a current is flowing the electrons also move along the wire. In order to link the speed along the wire to the size of the current, we need to consider how many free electrons are available and the width of the wire. Figure 5.10 shows a current, I, flowing through part of a wire of cross-sectional area A.

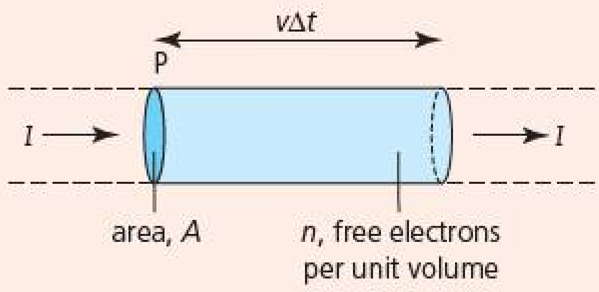


Figure 5.10 Deriving $I = nAve$

During time Δt an electron travelling at a net speed v along the wire will move a distance of $v\Delta t$. The number of electrons which flow past point P in time Δt is equal to the volume $v\Delta tA$ (shown by the shaded area in Figure 5.10) multiplied by the number of free electrons in unit volume, n. (Every material has a specific value for n which is closely linked to its resistivity.)

Since current = charge/time, the current in the wire equals the number of electrons passing a point in unit time multiplied by the charge on each electron, e.

$$I = \frac{v\Delta tA \times n \times e}{\Delta t}$$

$$I = nAve$$

Questions

1 Calculate the net speed of free electrons in a wire of cross-sectional area 1 mm^2 if the current is 1 A and the metal has 10^{28} free electrons per cubic metre.

2 Consider a series circuit which contains conductors of different thickness and different materials.

   a What must happen to the speed of free electrons as they move from a thicker wire to a thinner wire of the same material? Explain your answer.

   b Compare the value of n for a good conductor with that for a semiconductor. If the resistivity of a particular semiconductor is one million times greater than copper, compare the probable speeds of free electrons passing through these materials for the same current. What assumption(s) did you have to make?

Typically calculations like these show that the net speed of electrons in currents in metal wires is less than a millimetre per second, which is tiny fraction of their random speed. The low speed is surprising because we are used to electrical devices working immediately when we turn them on. But remember that we should not think of the current as starting at the battery or the switch. It is better to think of all the free electrons, wherever they are in the circuit, starting and stopping at the same time. However, in reality, this cannot be perfectly true. When a switch is turned on, the battery sets up an electric field which moves around the circuit at a speed close to the speed of light ($3 \times 10^8\,m\,s^{-1}$). As the electric field reaches individual electrons they experience a force which starts their net motion.

3 Does the speed of electricity in circuits have any significant effect on the speed at which a computer can process data?

Heating effect of a current

When a current passes through a conductor the free electrons transfer some energy to the ions or atoms with which they 'collide'. Any increase in the amplitude of vibrations of the ions or atoms is an increase in the internal energy and temperature of the conductor.

In electrical heaters this effect is very useful but in most other devices and in all connecting wires the effect is unwanted and a waste of useful energy. Connecting wires should be made of a very good metallic conductor (usually copper) and thick enough that there is no significant heating effect in normal use.

When a resistor heats up, the increased vibrations of the ions or atoms will tend to increase the resistance, especially for good conductors like metals. This is also true for poor conductors and semiconductors, but for these materials a higher temperature will probably release significantly more free electrons, resulting in an overall decrease in resistivity and making them better conductors.

Ohm's law

5.1.8 **State** Ohm's law.
5.1.9 **Compare** ohmic and non-ohmic behaviour.

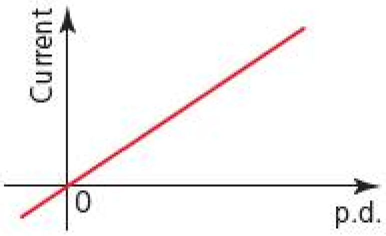


Figure 5.11 Ohm's law for an ohmic resistor

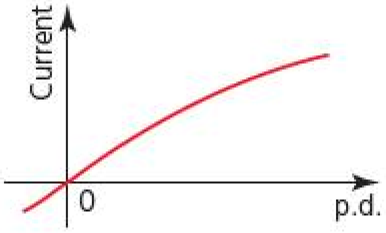


Figure 5.12 A current–p.d. graph for a filament lamp

If the temperature of a metallic conductor is kept constant, its resistance will not change; any increase in the potential difference across it will produce a proportional increase in current, so that V/I ($= R$) is constant, as shown in Figure 5.11. This relationship was discovered by the German physicist Georg Ohm in 1826. (Note that the graph has been extended some way into negative values as an indication that the same behaviour is seen if p.d. and current both reverse directions.)

Ohm's law states that the current through a metallic conductor is proportional to the potential difference across it, if the temperature is constant.

$I \propto V$ at constant temperature

If the temperature of a metallic conductor does not change very much, it can usually be assumed that the resistance is constant and Ohm's law can be used. But, if there are large changes in temperatures, Ohm's law is not useful. For example, the operating temperature of a filament lamp may be 2500 °C and its resistance at that temperature will be significantly greater than its resistance at room temperature. Figure 5.12 shows how the increasing resistance of a filament lamp might affect its I–V graph.

Ohm's law is a starting point for understanding electrical conduction. It works equally well for dc and ac circuits, but there are many circuit components to which it cannot be applied. In order to investigate the electrical behaviour of a specific component, either of the circuits shown in Figure 5.13 could be used. In circuit **a**, a variable electrical supply is used to obtain different voltages across the component. Circuit **b** uses a battery of fixed voltage and a **variable resistor** (a **rheostat**) to change the overall resistance of the circuit, therefore changing the current and voltage across the component.

If the current through an electrical component is proportional to the p.d. across it, then it is described as being **ohmic**, because it 'obeys' Ohm's law. All other components (including those not being used at approximately constant temperature) are described as **non-ohmic**.

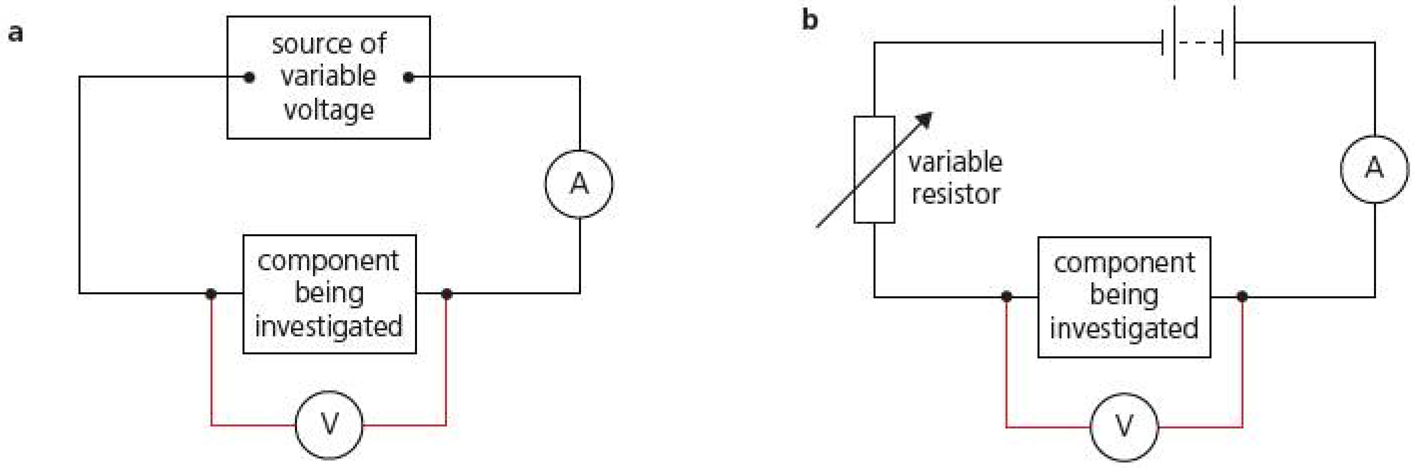


Figure 5.13 Two methods to investigate the I–V characteristic of an electrical component

Power dissipation in resistors

5.1.10 **Derive** and apply expressions for electrical power dissipation in resistors.

If the current through a resistor is, for example, 3 A, then 3 C of charge is passing through it every second. If there is a potential difference across the resistor of 6 V, then 6 J of energy is being transferred by every coulomb of charge (to internal energy). The rate of transfer of energy is $3 \times 6 = 12$ joules every second (watts).

More generally, we can derive an expression for the power dissipated to internal energy in a resistor by considering the definitions of p.d. and current, as follows:

$$\frac{\text{energy transferred}}{\text{time}} = \frac{\text{energy transferred in resistor}}{\text{charge flowing through resistor}} \times \frac{\text{charge flowing through resistor}}{\text{time}}$$

$$\frac{W}{t} = \frac{W}{q} \times \frac{q}{t}$$

power = potential difference × current
$P = VI$

Since $V = IR$, this can be rewritten as $P = (IR)I$, or

$$P = I^2R$$

Alternatively, $P = V(V/R)$, or

$$P = \frac{V^2}{R}$$

These three forms of the same equation are all shown in the IB *Physics data booklet*.

To calculate the total energy transferred in a given time, we know that energy = power × time, so that:

energy = VIt

Worked example

7 An electric iron is labelled as 220 V, 1200 W.
a Explain what the label means.
b What is the resistance of the heating element of the iron?
c Explain what would happen if the iron was used in a country where the mains voltage was 110 V.

a The label means that the iron is designed to be used with 220 V and, when correctly connected, it will transfer energy at a rate of 1200 joules every second.

b $P = \frac{V^2}{R}$

$1200 = \frac{220^2}{R}$

$R = 40.3\,\Omega$

c $P = \frac{V^2}{R}$ so $P = \frac{110^2}{40.3}$

$P = 300\,\text{W}$

The iron would transfer energy at $\frac{1}{4}$ of the intended rate and would not get hot enough to work properly ($P = VI$, and both the p.d. and the current will be halved, assuming the resistance is constant).

If an iron designed to work with 110 V was plugged into 220 V it would begin to transfer energy at four times the rate it was designed for; it would overheat and be permanently damaged.

5.1.11 Solve problems involving potential difference, current and resistance.

17 A 12 V potential difference is applied across a 240 Ω resistor.
a Calculate: i the current ii the power iii the total energy transferred in 2 minutes.
b What value resistor would have twice the power with the same voltage?

18 A 2.00 kW household water heater has a resistance of 24.3 Ω.
a What current flows through it?
b What is the mains voltage?

19 a What would be the rate of production of thermal energy if a current of 100 A flowed through an overhead cable of length 20 km and resistance 0.001 ohm per metre?
b Comment on your answer.

20 a What power heater will raise the temperature of a metal block of mass 2.3 kg from 23 °C to 47 °C in 4 minutes (specific heat capacity = 670 J kg^{-1} °C^{-1})?
b Draw a circuit diagram to show how the heater should be connected to a 12 V supply and suitable electrical meters, so that the power can be checked.

21 a An electric motor is used to raise a 50 kg mass to a height of 2.5 m in 74 s. The voltage supplied to the motor was 240 V but it was only 8% efficient. What was the current in the motor?
b Suggest two reasons why the motor has a low efficiency.

22 a What value resistance would be needed to make a 1.25 kW water heater in a country where the mains voltage is 110 V?
b What current flows through the heater during normal use?
c Suggest why a greater current flows when it is first turned on.

Additional Perspectives

Paying for electricity

When we buy a battery or pay for the 'mains' electricity connected to our homes, we are really buying energy. In most countries mains electrical energy is sold by the *kilowatt hour*. 1 kW h is the amount of energy transferred by a 1 kW device in one hour, that is, 1000 J s^{-1} for 3600 s, or 3.6 MJ.

Questions

1 a Use energy = VIt to estimate how much useful energy can be transferred from a single AA sized cell.
b What is the approximate cost of that energy per MJ?
2 a How much do householders have to pay for 1 kW h in your country?
b Use the Internet to compare your prices with those in some other countries.
c What is the cost of 1 MJ of mains electrical energy in your country? By comparison, it should be clear that buying disposable batteries is a very expensive way of paying for energy and convenience. The disposal of the batteries is also a pollution problem.

In Chapter 8 we will discuss the effect of the generation and use of electrical energy on the environment. There are enormous differences in how much electrical energy is used in different countries. One way of discouraging the excessive use of electricity in developed countries is for governments to make the cost of a kW h greater for those homes that use more energy.

3 Discuss the advantages and disadvantages of the pricing system for electricity mentioned in the last paragraph.

5.2 Electric circuits

Internal resistance and electromotive force

5.2.1 **Define** *electromotive force* (*emf*).
5.2.2 **Describe** the concept of internal resistance.

Cells and batteries use chemical reactions to provide potential differences and energy to circuits. There are other, less commonly used, devices which can also provide p.d.s from different processes, for example, **photovoltaic cells** which transfer light energy to electrical energy. Our homes are provided with a p.d. and energy by **generators** at electrical power stations.

The **electromotive force** (emf) of a battery, or other source of electrical energy, is defined as the *total* energy transferred in the source per unit charge passing through it.

Most of this energy should be supplied to the circuit, but some energy will be transferred to internal energy within the battery itself.

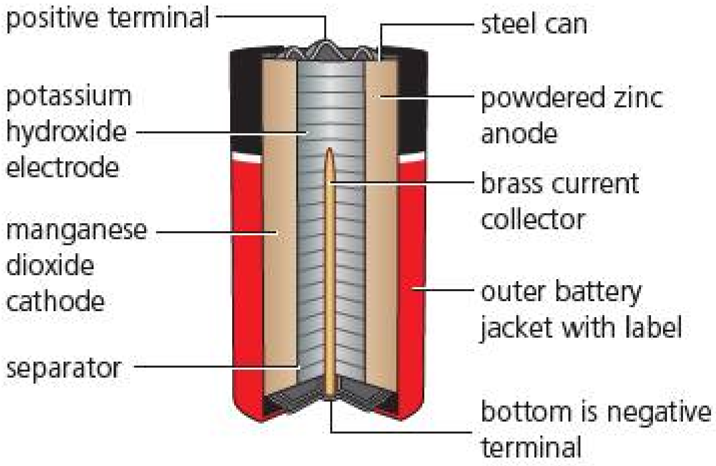


Figure 5.14 The chemicals inside a battery provide internal resistance

Electromotive force is given the symbol $\mathcal{E}$ and its unit is the volt, V. Electromotive force is a potential difference (voltage), so the name can cause confusion because it is not really a force at all. For this reason it is commonly called **emf**. The emf of a battery can be thought of as the p.d. across it when it is not being used to supply a current, so that no energy is being transferred inside it.

Cells and batteries and other sources of electrical energy are not perfect conductors of electricity. They all have resistance, called their **internal resistance**, which is given the symbol r. For example, the internal resistance of a new battery bought for a torch might be about one ohm.

If the internal resistance of a battery (Figure 5.14) is much less than the rest of the circuit, its effect can usually be ignored and, as a result, many examination questions refer to batteries or cells of 'negligible internal resistance'. But in other examples, the internal resistance of an energy source can have a significant effect on the circuit.

The value of the internal resistance of a battery may vary when different currents flow through it, but we usually assume that it is constant.

Consider Figure 5.15. As current flows through the cell, energy will be transferred to internal energy because of its internal resistance, so that:

total energy transferred by the cell (per coulomb) = energy transferred to the circuit (per coulomb) + energy transferred inside the cell (per coulomb)

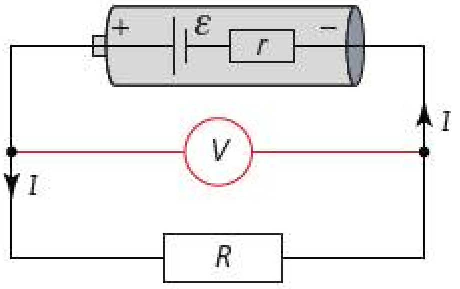


Figure 5.15 A cell in a simple circuit

Emf of cell, $\mathcal{E}$ = terminal p.d. across circuit, V + 'lost volts'

The same current flows through both resistances and using $V = IR$ gives:

$$\mathcal{E} = IR + Ir = I(R + r)$$

This equation is in the IB *Physics Data booklet*.

The voltmeter shown in Figure 5.15 is *not* measuring the emf of the cell; it is measuring the p.d. across the terminals of the cell, which is the same as the useful p.d. supplied to the circuit.

Worked example

8 a If the cell shown in Figure 5.15 has an emf of 1.5 V and an internal resistance of 0.90 Ω, calculate the current in the circuit if the resistor has a value of 5.0 Ω.

b What is the p.d. across the circuit?

a $\mathcal{E} = I(R + r)$

$1.5 = I(5.0 + 0.90)$

$I = \dfrac{1.5}{5.90} = 0.25\text{ A}$

b $V = IR$

$V = 0.25 \times 5.0 = 1.3\text{ V}$

The voltmeter will read 1.3 V, but the cell has an emf of 1.5 V. 0.2 V has been 'lost' because of the internal resistance. The 'lost volts' cannot be measured directly with a voltmeter, but can be calculated from Ir. This shows us that the value of the 'lost volts' will increase when a larger current, I, is flowing through the cell. If the power source has significant internal resistance the useful p.d. supplied to the circuit will decrease if a larger current flows.

To measure the emf of a source of electrical energy, a voltmeter can be placed across it when it is *not* supplying a current to a circuit. (An ideal voltmeter needs to have a very high resistance so that no significant current flows through it.)

5.2.8 Solve problems involving electric circuits.

23 a When a battery of emf 4.5V and internal resistance 1.1 Ω was connected to a resistor the current was 0.68 A. What was the value of the resistor?
b If the resistor was replaced with another of twice the value, what would the new current be?
c What assumption(s) did you make?

24 When a cell of internal resistance 0.24 Ω was connected to a lamp, the current was 0.72 A and the p.d. across the lamp was 2.8V.
a Calculate the resistance of the lamp.
b What was the emf of the cell?
c Calculate the rate of energy transfer (power) in:
i the lamp
ii the cell.

25 A high resistance voltmeter shows a voltage of 12.5V when it is connected across the terminals of a battery which is not supplying a current to a circuit. When the battery is connected to a lamp a current of 2.5 A flows and the reading on the voltmeter falls to 11.8V.
a What is the emf of the battery?
b Calculate the internal resistance of the battery.
c What is the resistance of the lamp?

26 The engine of a car is started at night with the car's headlights on. Starting a car requires a high current from a 12 V battery and the headlights become momentarily dim.
a Why is a high current needed?
b Suggest why the headlights become dim.

27 If a connecting wire is connected by mistake across a battery, or power supply, it is an example of a 'short circuit'.
a Calculate the current which flows through a battery of emf 12.0V and internal resistance 0.25 Ω if it is accidentally 'shorted'.
b What assumptions did you make?
c Calculate the power transferred in the battery.
d Suggest what will happen to the battery.

28 Explain why it is not possible to measure the internal resistance of a battery in the same way as the resistance of a resistor in a circuit.

Additional Perspectives

Batteries

Question

1 The battery in Figure 5.16 has an emf of 6.0 V and an internal resistance of 0.45 Ω. The variable resistor, *R*, can be changed to any value between 0 Ω to 3 Ω.
a Set up a spreadsheet to calculate the current in the circuit for a range of suitable values of *R*.
b Add columns to calculate the p.d. across *R* and the power generated in the resistor.
c Draw a graph showing how the power varies with the value of *R*.
d Explain why the power is low for both high and low values of *R*.
e Under what conditions can maximum power be obtained from the battery?

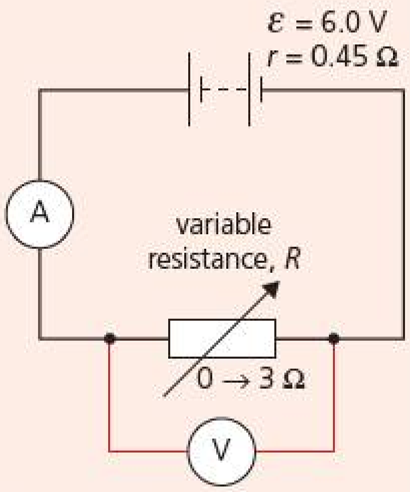


Figure 5.16

When we replace a battery in electrical or electronic equipment we usually have no choice about which kind to use – that decision was made during the design process. The question above shows that the internal resistance of a cell (and how it compares with the rest of the circuit) is important, as is the choice of which voltage to use. But there are other factors as well, including the size and weight of the battery, how much energy is stored in it and whether the voltage stays reasonably constant during use.

If batteries are thrown away they can cause a pollution problem because of the chemicals they contain. This is one reason why rechargeable batteries are preferable.

The amount of energy stored in a battery is often given in the unit amp hour, A h (or watt hour, W h), rather than joules. For example, a battery rated at 10 A h should be able to provide 1 A for 10 h or 0.01 A for 1000 h, etc. If it is a 12 V battery then the nominal total energy stored will be $12 \times 10 \times 3600 = 4.3 \times 10^5$ J. In the modern world much attention has turned to improving battery design and getting the maximum possible energy from batteries that are as small and lightweight as possible, for use in mobile phones and notebook computers for example.

Figure 5.17 A battery in a mobile phone

A typical mobile phone battery (Figure 5.17) is lithium ion in design, and might be rated at 3.7 V, 3 W h and approximately 1 kJ can be stored in every gram of chemicals in the battery.

Question

2 A major attraction of electric cars is that they are non-polluting in use, but the energy stored in their batteries had to be transferred from somewhere else. Much research is continuing into the development of better batteries for electric cars to make them able to supply the energy needed for greater distances and make their recharging quicker. Use the Internet to find out the latest developments.

Resistors in series and parallel

5.2.3 **Apply** the equations for resistors in series and parallel.

Two or more resistors (or other components) are said to be **in series** if they are connected one after another, so that the same current flows through them all. Resistors (or other components) are said to be in **parallel** if the current divides and can take two or more different routes between the same two points. A circuit may contain a combination of series and parallel arrangements.

Figure 5.18 shows three different resistors in series. Because of the law of conservation of charge (see Chapter 6), the charge per second (current) flowing into each resistor must be the same as the current, I, flowing out of it and into the next resistor.

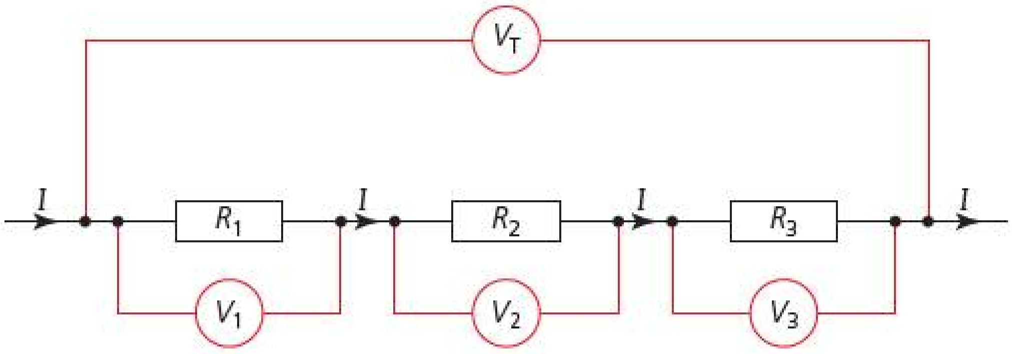


Figure 5.18 Three resistors in series

The sum of the separate potential differences must equal the potential difference across them all, V_T, so that

$$V_T = V_1 + V_2 + V_3$$

Using $V = IR$ for the individual resistors, we get $IR = IR_1 + IR_2 + IR_3$, so that we can derive an equation for the single resistor, R, which has the same resistance as the combination:

$$R = R_1 + R_2 + R_3$$

This equation is listed in the IB *Physics data booklet*.

Figure 5.19 show three resistors in parallel with each other. Because they are all connected between the same two points, they all have the same potential difference, V, across them. The law of conservation of charge means that:

$$I_T = I_1 + I_2 + I_3$$

Applying $V = IR$ throughout gives

$$\frac{V}{R} = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3}$$

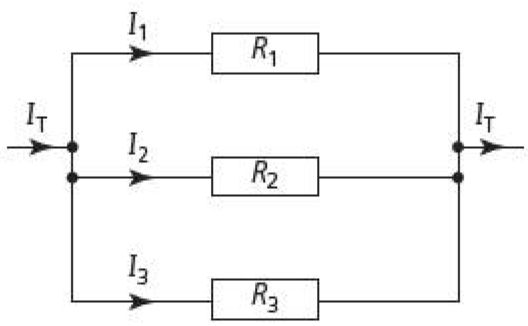


Figure 5.19 Three resistors in parallel

Cancelling the Vs gives us an equation for the single resistor, R, which has the same resistance as the combination:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

This equation is shown in the IB *Physics data booklet*.

All the electrical equipment in our homes is wired in parallel because, in that way, each device is connected to the full supply voltage and can be controlled with a separate switch.

Circuit symbols

5.2.4 Draw circuit diagrams.

Figure 5.20 summarizes the standard circuit symbols used so far, together with a few others that need to be known for the rest of the work in this chapter.

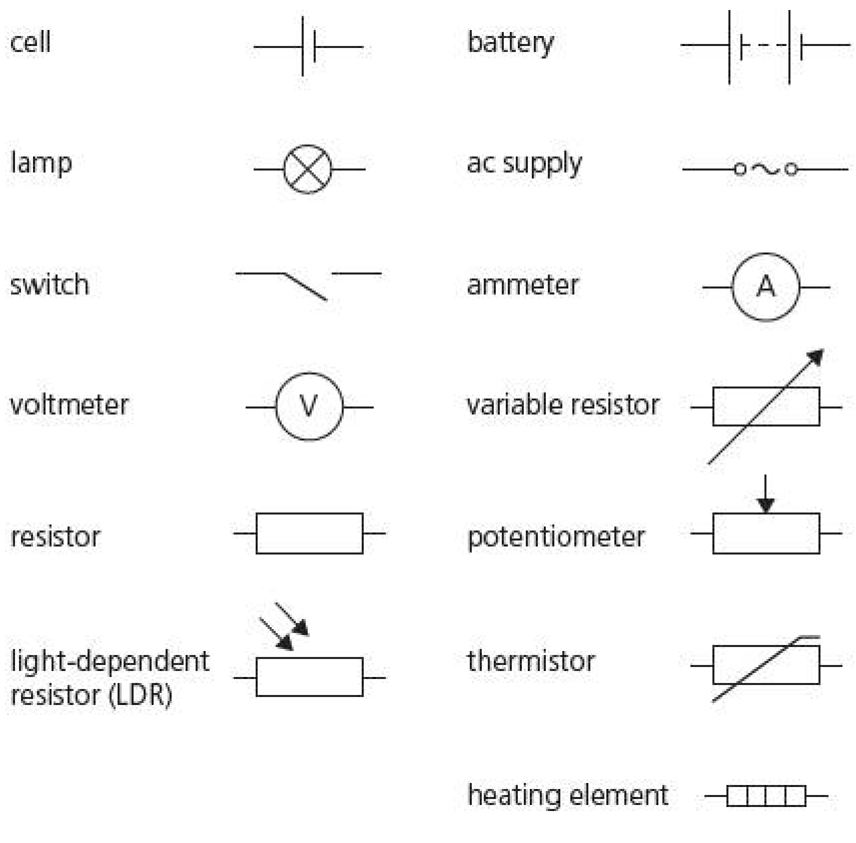


Figure 5.20 Summary of the circuit symbols used in this course

Resistance of electrical meters

5.2.5 Describe the use of ideal ammeters and ideal voltmeters.

Ammeters are always connected in series in the circuit which they are measuring. An ideal ammeter will have a very low resistance so that it does not reduce the current that it is trying to measure. Although ammeters do have some resistance, this is usually very small and generally not a problem.

Voltmeters are connected in parallel across the component being tested. An ideal voltmeter will have a resistance which is much higher than the component across which it is connected. This is so that very little current flows through it, so it does affect the voltage that it is trying to measure.

Modern digital voltmeters usually have very high resistances, but the resistance of other (analogue) voltmeters may be much lower and their use with high resistance circuits may give misleading results because they significantly reduce the voltage they are trying to measure.

Worked examples

9 A 5000 Ω resistor and 8000 Ω resistor are connected in series.

a What is their combined resistance?

b What is the current through each of them if they are connected to a 4.5 V battery of negligible internal resistance?

c What is the potential difference across the 5000 Ω resistor?

d Repeat these three calculations for the same resistors in parallel with each other.

a $5000 + 8000 = 13\,000\,\Omega$

b $I = \frac{V}{R} = \frac{4.5}{13\,000} = 3.5 \times 10^{-4}$ A, through both resistors

c $V = IR = 3.5 \times 10^{-4} \times 5000 = 1.7$ V

d $\frac{1}{5000} + \frac{1}{8000} = \frac{13}{40\,000}$

$R = \frac{40\,000}{13} = 3080\,\Omega$

Both resistors have a p.d. of 4.5 V across them.

Current through 5000 Ω, $I = \frac{V}{R} = \frac{4.5}{5000} = 9.0 \times 10^{-4}$ A

Current through 8000 Ω, $I = \frac{V}{R} = \frac{4.5}{8000} = 5.6 \times 10^{-4}$ A

10 The lamps shown in Figure 5.21 are all the same and the battery has negligible internal resistance.

a Compare the brightness of all the lamps.

b If all the lamps have the same constant resistance of 2 Ω, what is the total resistance of the circuit?

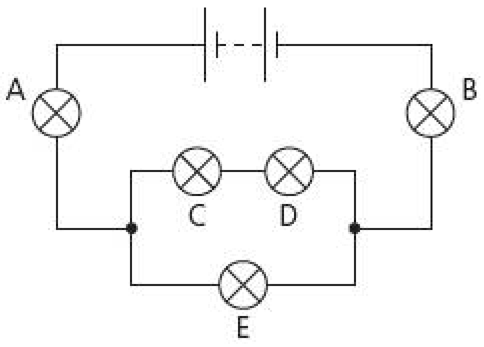


Figure 5.21

a Lamps A and B will have the same brightness because the same current flows through them both. That same current will be split between lamp E and lamps C and D, so that these three must all be dimmer than lamps A and B. Lamps C and D will have the same brightness because they are in series with each other. Lamp E will be brighter than lamps C or D because a greater current will flow through it.

b C and D together will have a resistance of $2 + 2 = 4\,\Omega$

E in parallel with C/D will have a resistance of $1.3\,\Omega \left(\frac{1}{R} = \frac{1}{2} + \frac{1}{4}\right)$

Total resistance $= 2 + 2 + 1.3 = 5.3\,\Omega$

11 Consider the circuit shown in Figure 5.22 before the voltmeter is connected.

a What is the current in the circuit before the voltmeter is connected?

b What is the voltage across the 2000 Ω resistor?

c If voltmeters with resistances of

i 5000 Ω and

ii 50 000 Ω are connected across the 2000 Ω resistor, what voltages will be measured?

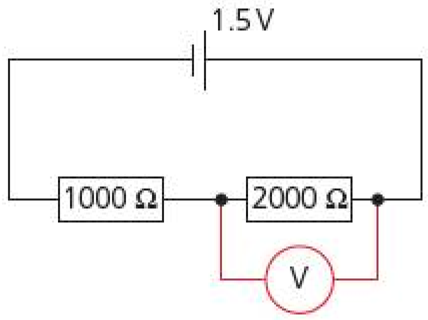


Figure 5.22

a $I = \dfrac{V}{R}$

$I = \dfrac{1.5}{(1000 + 2000)}$

$I = 5 \times 10^{-4}\,\text{A}$ (0.5 mA)

b $V = IR$

$V = (5 \times 10^{-4}) \times 2000$

$V = 1.0\,\text{V}$

c i First calculate the combined resistance of the 2000 Ω resistor and the 5000 Ω voltmeter in parallel:

$\dfrac{1}{R} = \dfrac{1}{R_1} + \dfrac{1}{R_2}$

$\dfrac{1}{R} = \dfrac{1}{2000} + \dfrac{1}{5000}$

$R = 1430\,\Omega$

Then find the voltage across a resistance of 1430 Ω connected in series with a 1000 Ω resistance:

$V = 1.5 \times \dfrac{1430}{1430 + 1000}$

$V = 0.88\,\text{V}$ (instead of the 1.0 V predicted in **b**)

ii Repeating the calculation with a 50 000 Ω voltmeter gives $V = 0.99\,\text{V}$, which is better because it is much closer to the value predicted without the voltmeter.

Additional Perspectives

Lighting our homes

The first inventions which produced light from electricity used the heating effect of a current in a metal wire to make it so hot that it emitted visible light. The same principle is still used today; the wire is called a filament and the lamps are described as being incandescent. Incandescent lamps (Figure 5.23a) can be designed for use with a wide range of different voltages and powers. They can be made big enough to light a room or small enough for use in a pocket torch.

Incandescent lamps are easy to manufacture and inexpensive, but less than 10% of the electrical energy transferred to them becomes visible light (the rest is thermal energy). When the total number of filament lamps in use around the world is considered, this is obviously an enormous waste of energy and the generation of all that wasted power has a significant environmental effect on the planet.

An electric current passing through a gas at low pressure can also produce visible light, but a significant amount of ultraviolet radiation may also be emitted. If the inside of the glass container is given a suitable coating then most of the ultraviolet radiation can be changed into visible light. The coating and the lamp are described as fluorescent. Fluorescent lamps are typically much more efficient at producing light than incandescent lamps, but they have the disadvantages of being bigger, more expensive and they cannot be used with low voltages. Large fluorescent tubes (Figure 5.23b) have been considered the best choice for lighting shops, offices, schools and advertising displays, for example, for many years. However, in recent years smaller fluorescent lamps have become cheaper and more widely available for home use, so that the widespread use of incandescent lamps for general household lighting is now generally discouraged in most countries; it is likely that more and more countries will ban the use of incandescent lamps.

Apart from being much more efficient than incandescent lamps, compact fluorescent lamps (CFLs) (Figure 5.23c) usually have a longer 'lifetime' before they need to be replaced, and this is important when considering the financial costs and the effects on the environment of their manufacture and disposal.

Figure 5.23 Four kinds of lamp: **a** incandescent, **b** fluorescent tube, **c** CFL, **d** LED

Light emitting diodes (LEDs) are beginning to provide another alternative for lighting our homes. They have been used for a long time as indicator lamps to show when electrical devices are on or off. They have been useful for this purpose because of their small size, low power, low cost, reliability and long lifetime. Early designs of LEDs emitted green or red light and it is only in recent years that LEDs have begun to be used for more general lighting purposes. This is because of the inventions in the 1990s of LEDs that emit blue light and then white light (Figure 5.23d). LEDs that emit different colours can be combined to create a vast range of different coloured effects.

A lot of research is still going on into improving LED design, especially into improving their overall efficiency in the large-scale conversion of electrical energy from a mains supply into visible light. Currently, the overall efficiency of using LEDs for home lighting is about the same as for filament (incandescent) lamps, but they are much more expensive to buy. LEDs offer much greater possibilities for interesting and innovative lighting design and they are more reliable and longer lasting.

Questions

1 Make a checklist of the properties of a lamp which make it a suitable choice for lighting the homes of people around the world.

2 Use the Internet to research into the latest information on incandescent lamps, CFLs and LEDs. Use your checklist from 1 to compare the advantages and disadvantages of the three types of lamp.

5.2.8 Solve problems involving electric circuits.

29 a Three cells each of emf 1.5 V and internal resistance 0.80 Ω are joined in series. What is their combined emf and internal resistance?

b What is the combined emf and internal resistance if they are connected in parallel?

c What is the current through the battery and the p.d. across the circuit when the cells are connected to a 3.2 Ω resistor

i in series and

ii in parallel?

30 The two lamps shown in Figure 5.24 both have a resistance of 4.0 Ω when cold and 6.0 Ω when working normally with a p.d. of 6.0 V across them. The battery has negligible internal resistance.

a What must be the value of the resistor R for the lamps to operate normally?

b What is the p.d. across the lamps when the switch is first turned on?

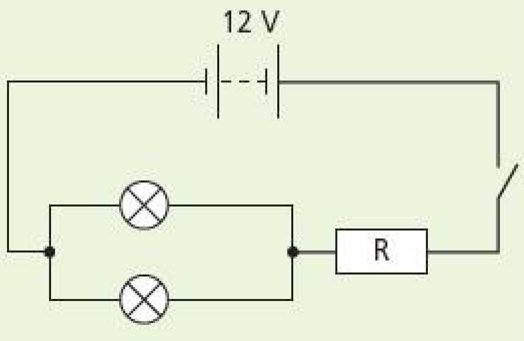


Figure 5.24

31 A fan heater, which is operated from the electricity mains, contains a heating element, a fan and a lamp to indicate when the heating element is on. The fan must always be on when the heater is on, but the fan can also be switched on if the heater is off. Draw a circuit with a two-way switch to show how they are all connected.

32 Consider the circuit shown in Figure 5.25. In which resistor will power be dissipated at the greatest rate? Explain your answer.

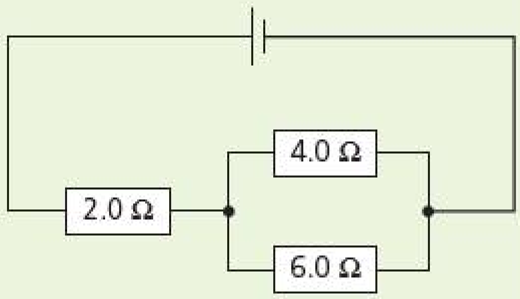


Figure 5.25

33 A 220 V mains electric heater has three settings on its switch: high, medium and low. Inside the heater there are two 40.0 Ω heating elements.

- **a** Explain, with diagrams, how the use of just two heating elements can produce three different power outputs.
- **b** Calculate the power output on the three different settings.

34 Refer back to Figure 5.25.

- **a** What is the combined resistance of the three resistors?
- **b** If the cell has an emf of 1.48 V and an internal resistance of 0.64 Ω, calculate the current flowing through the cell.
- **c** What is the voltage across the 6.0 Ω resistor?

35 Figure 5.26 shows a simple circuit in which the ammeter and voltmeter have been connected in the wrong positions. The battery has negligible internal resistance.

- **a** What (approximate) readings would you expect to see on the meters? Explain your answer.
- **b** When the positions of the meters were swapped to their correct positions, what readings would you expect to see on the meters?
- **c** The actual reading on the voltmeter was 3.9 V. Suggest a reason why it was lower than expected.
- **d** Calculate the resistance of the voltmeter.
- **e** What assumption(s) did you make about the ammeter?

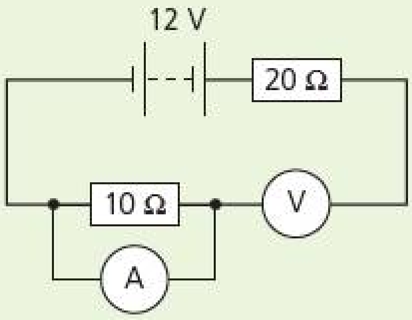


Figure 5.26

36 In the circuit shown in Figure 5.27 the battery has negligible internal resistance and, when the switch is closed, a current of 830 mA flows through the lamp.

- **a** What is the p.d. across the lamp?
- **b** Calculate the resistance and power of the lamp under these conditions.
- **c** If the switch is opened, explain what happens to the brightness of the lamp.
- **d** Calculate the new p.d. across the lamp.

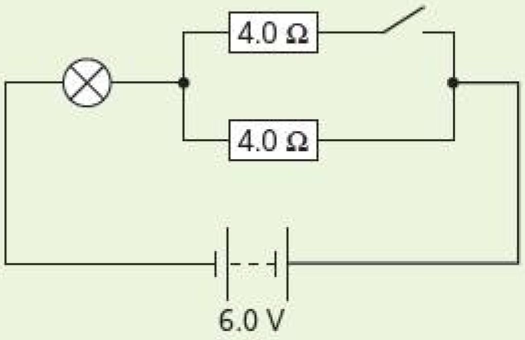


Figure 5.27

Circuits with potential dividers

5.2.6 Describe a potential divider.

When two (or more) resistors are connected in series they share the total potential difference across them (as shown in Worked example 9).

For resistors in series, the potential difference will be shared in the same ratio as the resistances. When this kind of arrangement of resistors is used deliberately for the control of p.d.s around circuits, it is called a **potential divider**.

Typically, one of the two resistors will be variable, as shown in the following example.

Worked example

12 The value of the variable resistor in Figure 5.28 can be changed continuously from 1 kΩ to 10 kΩ. What is the maximum and minimum potential difference, V_{out}, that can be obtained across R? Assume that the battery has negligible internal resistance.

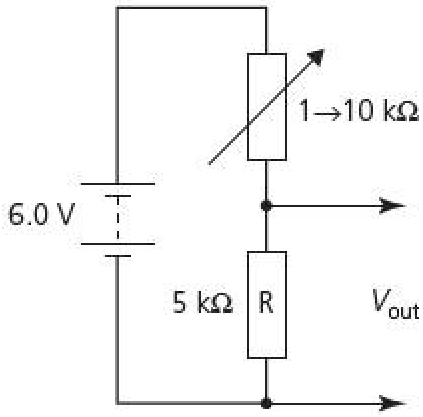


Figure 5.28

If the variable resistance is set to 1 kΩ, $V_{out} = \frac{6 \times 5}{5+1} = 5\,V$

If the variable resistance is set to 10 kΩ, $V_{out} = \frac{6 \times 5}{5+10} = 2\,V$

V_{out} may be a voltage which controls another part of the circuit, perhaps some kind of electronic switch that might be on if the voltage was, for example, greater than 4 V and off if it was less than 3 V. In this way, another circuit can be switched on or off by changing the value of the variable resistor in this potential divider.

Sensors

5.2.7 Explain the use of sensors in potential divider circuits.

Potential dividers are often used in *automatic control systems* in which a sensor is used as one of the two resistors. A **sensor** is an electrical component which responds to a change in a physical property with a corresponding change in an electrical property (usually resistance).

Sensors are commonly available for the detection of most of the physical properties discussed in this book, but we shall limit our discussion to three sensors in particular:

- **light-dependent resistors (LDRs)** (Figure 5.29)
- temperature dependent resistors (**thermistors**) (Figure 5.30)
- **strain gauges**.

Figure 5.29 An LDR

Figure 5.30 A thermistor

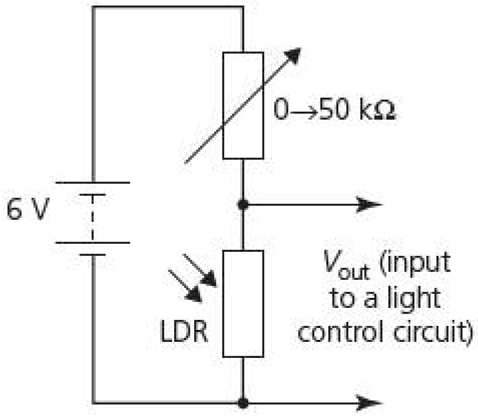


Figure 5.31 An LDR in a potential divider circuit

The resistance of LDRs and thermistors both decrease when increased light intensity or higher temperatures provide the energy needed to release more free electrons in their semiconducting materials. Strain gauges affect resistance when a change of shape occurs.

Figure 5.31 shows an LDR connected as part of a potential divider. As the light intensity increases, the resistance of the LDR decreases and, therefore, the voltage across it, V_{out}, also goes down. V_{out} can be used to control an electronic switch which turns off lights when the light intensity rises to a certain level. The light level at which the lights are turned on or off can be changed by adjusting the value of the variable resistor.

A thermistor placed in a similar circuit could be used to turn a heater on when the temperature falls below a certain value, or turn a heater off when it gets too hot. Used in this way to control temperature, the thermistor would be part of a **thermostat**. Figure 5.32 shows how the resistance of a certain kind of semiconducting thermistor changes with temperature. This is called a **negative temperature coefficient (NTC) thermistor** because its resistance decreases as the temperature rises.

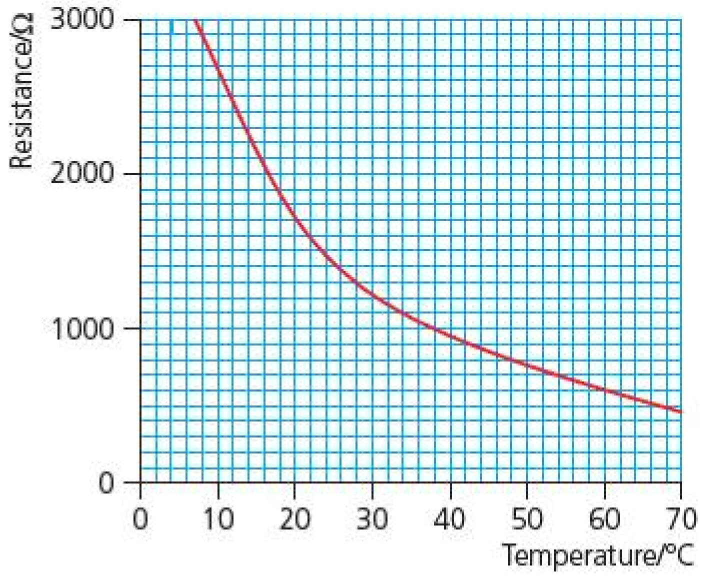


Figure 5.32 Resistance–temperature graph for an NTC thermistor

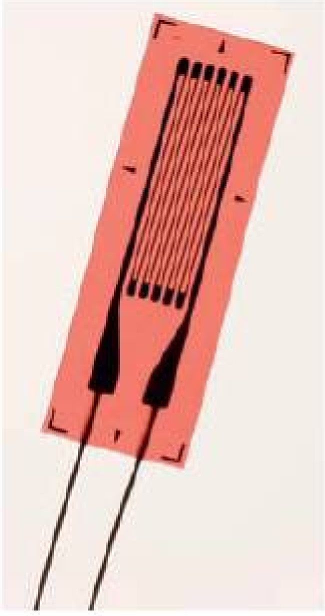


Figure 5.33 A strain gauge

Figure 5.33 shows a strain gauge with a thin wire that is used to detect when a structure undergoes a small change of shape. It contains a thin wire which gets very slightly longer and thinner when strained, so that its resistance increases a little (remember $R = \rho L/A$).

Worked example

13 A battery providing a potential difference of 6.0 V is connected across a thermistor and a 2000 Ω fixed resistor connected in series. Use the graph in Figure 5.32 to predict the temperature at which the p.d. across the thermistor becomes 2.5 V.

The ratio of the resistances will equal the ratio of the voltages:

$$\frac{R_T}{2000} = \frac{2.5}{6.0 - 2.5}$$

$$R_T = \frac{2000 \times 2.5}{3.5} = 1430\,\Omega$$

Using the graph, we can see that the resistance is 1430 Ω when the temperature is about 25 °C.

5.2.8 Solve problems involving electric circuits.

37 **a** What is the potential difference across points A and B in Figure 5.34?

b A student wishes to connect a lamp which is rated at 15 W, 6 V. What is the working resistance of the lamp?

c The student thinks that the lamp will work normally if he connects it between A and B, in parallel with X. Explain why the lamp will not work as he hopes.

d Another student thinks that the lamp will work if the resistor X is removed (while the lamp is still connected between A and B). Calculate the voltage across the lamp. Will the lamp work normally now?

e Suggest how the lamp could work normally using a 12 V battery.

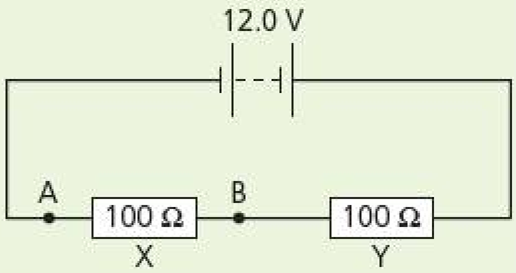


Figure 5.34

38 **a** Draw a potential dividing circuit which could be used to control the temperature in a refrigerator.

b Make a list of electrical devices which have thermostats inside them.

39 Figure 5.35 shows how a particular semiconducting LDR responds to changes in light intensity. The LDR has a resistance of about 1300 Ω in normal room lighting (400 lux).

a Explain why the resistance decreases as the light intensity increases.

b The scales are logarithmic. Explain why this type of scale is used.

c Write an equation for the line.

d Calculate the resistance of the LDR when the light intensity is 200 lux.

e If the LDR was connected as shown in Figure 5.31, what value of the variable resistance would produce a p.d. of 1.2 V across the LDR under normal room lighting?

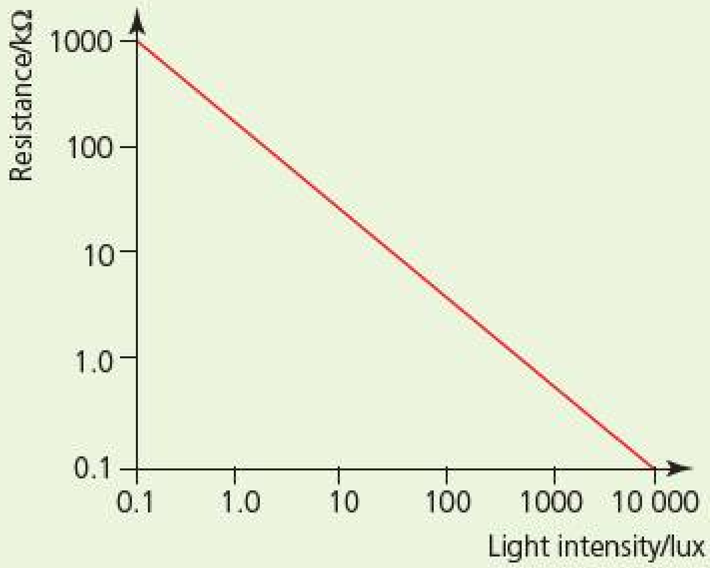


Figure 5.35 Resistance–intensity graph for an LDR

40 a If the length of a certain wire was increased by 1.00%, what percentage change would you expect in its resistance? (Assume that its volume remains constant.)

b Figure 5.36 shows a strain gauge connected in a potential dividing circuit. Calculate the reading on the voltmeter. What assumption(s) did you have to make about the voltmeter?

c If the resistance of the strain gauge increases by 0.570%, what will be the new reading on the voltmeter?

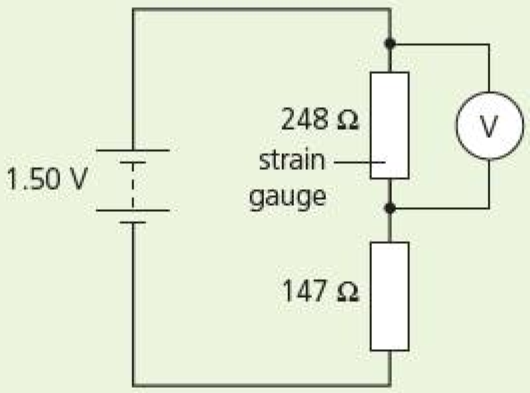


Figure 5.36

Using a variable resistor as a potential divider

Variable resistors (Figure 5.37) are made in many different shapes and sizes. Most have three terminals, one at each end of the resistor and one sliding contact which can be moved along the resistor anywhere between the other two contacts.

Figure 5.37 A selection of variable resistors

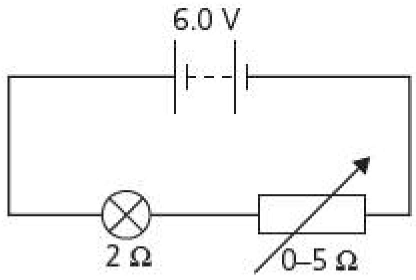


Figure 5.38 Circuit using a variable resistance to change current

The brightness of the bulb in Figure 5.38 can be adjusted with the variable resistance. Until now, we have only considered using two terminals. Used in this way, it may be best to think of a variable resistor as a way of changing the current, by changing the resistance in a series circuit – the p.d. is divided between the lamp and the variable resistor. Of course, the circuit shown in Figure 5.38 is also a potential dividing circuit. The range of the variable resistance must be chosen carefully for its intended use.

As an example, consider a 0–5 Ω variable resistor used to control a lamp of fixed resistance 2.0 Ω designed for normal use when there is a p.d. of 6 V across it. When the resistance is set to 0 Ω, the lamp will get the full 6 V and when the resistance is set to its maximum of 5 Ω, the lamp will have a p.d. of $(2.0/7.0) \times 6.0 = 1.7$ V across it. It is not possible to reduce the p.d. across the lamp to zero. Lower p.d.s would be possible with a variable resistance that had a much greater range, but then making adjustments for higher voltages would become more difficult.

Alternatively, a variable resistor can also be connected across a battery using all three terminals, as shown in Figure 5.39. Used in this way it can provide a potential difference, V_{out}, to another part of the circuit which varies continuously from zero to the full p.d. of the battery, V_{in}. The maximum voltage will be obtained with the sliding contact at the top of the variable resistor, and the voltage will be zero with the contact at the bottom. A variable resistor used in this way is called a **potentiometer**. It is essentially just a potential divider with the sliding contact on the variable resistor dividing it into two resistors of variable sizes (as shown by R_1 and R_2 where $R_1 + R_2$ is constant).

When a potentiometer is connected as the input into a circuit, the value of V_{out} cannot be calculated without considering the effect of the resistance of the rest of that circuit. Generally, the resistance of the circuit should be much greater than the resistance of the potentiometer.

A potentiometer provides the best way of varying the voltage to a component in order to investigate its *I–V* characteristic (see Figure 5.40).

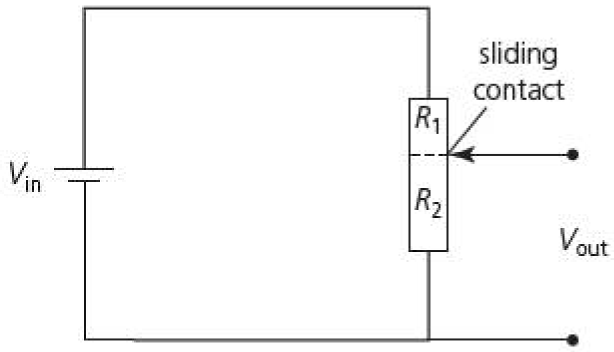


Figure 5.39 A variable resistor used as a potentiometer.

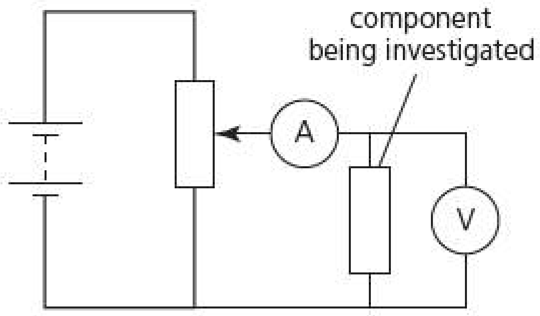


Figure 5.40 A circuit for investigating *I–V* characteristics of electrical components

41 a Draw a circuit which uses a potentiometer and a 12 V battery to provide an input that varies from 0 to 12 V for a small lamp rated at 12 V, 1.5 W.
b Label the lamp with its resistance (assume it is constant).
c Suggest a suitable resistance range for the potentiometer.
d If a 0–200 Ω potentiometer was used, estimate the p.d. across the lamp when the potentiometer was set to the middle of its range.
e Explain why the answer to **d** is not 6 V.

Additional Perspectives

Touchscreens

Touchscreens (Figure 5.41) have become very popular in devices such as mobile phones and tablet computers because they enable the user to control the display directly without the need for a keypad or mouse.

Figure 5.41 Touch-screen tablet

There are several different technologies used in touchscreens; one of these uses the resistive properties of the display screen. When any object, such a finger, touches the screen the resistance at that point changes and the location is calculated by the central controller. If the finger moves, new locations are detected – the speed and direction of the movement is calculated and used to control the device. Multipoint screens can respond to two or more touches at the same time. One problem with resistive touchscreens is the fact that the screen reduces the quality of the image.

Another common touchscreen technology responds to changes in capacitance (capacity for storing charge) at the point where the screen is touched. These screens need to be touched with a conductor, like a finger.

Questions

1 Choose a particular touchscreen phone, tablet computer (or other device) and find out what technology is used for the screen.
2 Can you think of any disadvantages of using touchscreen devices?

SUMMARY OF KNOWLEDGE

5.1 Electric potential difference, current and resistance

- Electric potential energy is transferred in circuits by moving electrons, which each carry a negative charge of -1.60×10^{-19} C. (6.24×10^{18} separate electrons carry a total charge of 1 C.)
- The energy transferred as one unit of charge moves between two points is called the potential difference (p.d.), commonly called voltage. That is, p.d. = energy/charge; $V = W/q$. One volt means one joule per coulomb.
- The p.d. between two points in a circuit can be measured with a voltmeter connected between the points (in parallel across the component being checked).
- Change in electric potential energy = p.d. × charge. In atomic physics the electronvolt, eV, is often used as the unit of energy (rather than the joule). It is defined as the kinetic energy gained (or potential energy lost) by a charge of 1.6×10^{-19} C accelerated by one volt. Energy transferred (J) = energy transferred (eV) × (1.6×10^{-19}).
- $Ve = \frac{1}{2}mv^2$
- An electric current is a flow of charged particles. Direct currents (dc) always flow in the same direction around a circuit, which is shown with an arrow that always points from the positive terminal (of the energy source) to the negative terminal. Alternating currents (ac) repeatedly change direction.
- Current = charge/time. $I = \Delta q/\Delta t$. Current is measured in amps. $1\,\text{A} = 1\,\text{C}\,\text{s}^{-1}$. The amp is a fundamental unit in the SI system, defined by the force per unit length between two parallel conductors 1 m apart each carrying a current of 1 A.
- Electric currents are measured by ammeters placed in series in the circuit.
- A p.d. is needed across any conductor in order to make a current flow through it. A knowledge of how different p.d.s affect the current is important information about any component. Circuits which use variable rheostats, or potentiometers, can be designed to investigate the I–V characteristic (graph) of a component.
- The ratio of p.d. to current is called resistance, R. $R = V/I$. The unit of resistance is the ohm, Ω. $1\,\Omega = 1\,\text{V}\,\text{A}^{-1}$. The resistance of a component may vary for a number of different reasons. In particular, significant temperature changes usually affect the resistance of a component.
- For some materials (like metals) the current through them is proportional to the p.d. across them, provided that the temperature is constant. This is known as Ohm's law. Components which obey Ohm's law are called ohmic.
- An ohmic component will have constant resistance and a linear I–V characteristic passing through the origin. Non-ohmic components (like a filament lamp) will not have linear I–V characteristics because their resistance changes.
- The resistance of a length of metal wire is proportional to its length, L, and inversely proportional to its cross-sectional area, A (at constant temperature). In general, for a regularly shaped specimen of the same material, $R = \rho L/A$, where ρ is a constant, called resistivity, which represents the resistive properties of the material (unit: Ω m).
- A component which is designed to have a certain resistance is called a resistor. Variable resistors are widely used for varying currents and voltages in circuits.
- When an electric current flows through a resistance, some electric potential energy is dissipated to internal energy, and the resistance may get hotter. The defining equations for p.d. and current can be combined to show that the power dissipation in a resistance, $P = VI$. Using $R = V/I$, it can be shown that the power is also given by: $P = I^2R = V^2/R$.

5.2 Electric circuits

- Electrical power sources (such as batteries) are not perfect conductors. They have internal resistance, r, and when a current flows through them some energy is dissipated to internal energy.

- If the internal resistance is much less than the resistance of the circuit, it may be described as negligible but, in some circuits, a value for the internal resistance of the source must be included in a calculation of the total resistance of the circuit (= $R + r$).
- The electromotive force (emf), $\mathcal{E}$, of a source is defined as the total energy transferred per unit charge passing through it. In practice, emf is equal to the p.d. across the terminals of a source when no current is flowing (so that no energy is transferred to internal energy). However, when a current flows some volts are 'lost' (Ir) because of the internal resistance and are not available to the rest of the circuit.
- emf = voltage delivered to circuit + lost volts. $\mathcal{E} = IR + Ir$.
- The total resistance of resistors in series, $R = R_1 + R_2 + \ldots$
- The total resistance of resistors in parallel can be found from $1/R = 1/R_1 + 1/R_2 + \ldots$
- Ideal ammeters have negligible resistance and ideal voltmeters have very large resistance (compared to the other components). In circuits where this is not true, the resistances of the meters have to be included in circuit calculations.
- Sensors are components that respond to a difference in a physical property with a corresponding change in resistance (or other electrical property). The resistance of a light-dependent resistor (LDR) decreases with light intensity; the resistance of most thermistors decreases as the temperature increases; the resistance of a strain gauge increases as it gets longer.
- Sensors are often connected in series with another resistor and a power source (battery), so that they share (divide) the total p.d. When the resistance of the sensor changes, so too do the p.d.s, and the changing p.d.s can be used to turn another part of the circuit on or off. Such arrangements are called potential dividers.
- If all three terminals of a variable resistor are used in a control circuit, it is called a potentiometer.

Examination questions – a selection

Paper 1 IB questions and IB style questions

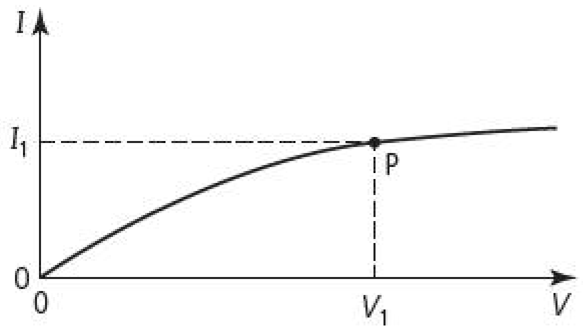
Q1 What are the ideal electrical properties of ammeters and voltmeters?

A ammeter zero resistance; voltmeter zero resistance
B ammeter zero resistance; voltmeter infinite resistance
C ammeter infinite resistance; voltmeter zero resistance
D ammeter infinite resistance; voltmeter infinite resistance

Q2 A voltage of 3 kV is used to accelerate an electron from rest in a vacuum. What is the maximum kinetic energy of the electron?

A 3 eV
B 3 keV
C 4.8 eV
D 4.8 keV

Q3 The graph shows the I–V characteristics of a certain electrical component.



Which of the following shows how the resistance of the component at point P may be calculated?

A $\dfrac{V_1}{I_1}$
B $\dfrac{I_1}{V_1}$
C the gradient of the line at P
D $\dfrac{1}{\text{gradient at } P}$

Q4 The SI unit of current is based on:

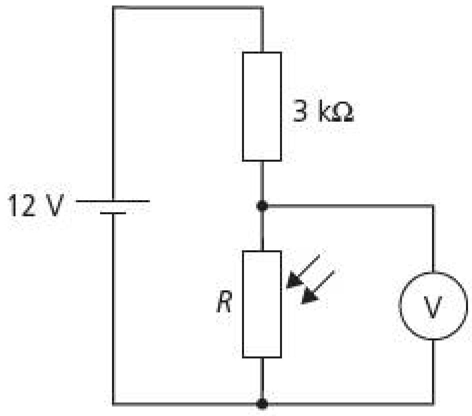
A the charge passing a point in one second.

B the power transformed by a p.d. of one volt.

C the force on a conductor in a permanent magnetic field.

D the force between parallel current-carrying conductors.

Q5 The diagram shows a potential divider circuit.



In order to increase the reading on the voltmeter the

A temperature of R should be increased.

B temperature of R should be decreased.

C light intensity on R should be increased.

D light intensity on R should be decreased.

Standard Level Paper 1, May 09 TZ1, Q18

Q6 An electric heater has two elements, both connected in parallel to the same voltage supply. The first element has a resistance of R and a power output of P. The second element has a power output of $4P$. What is its resistance?

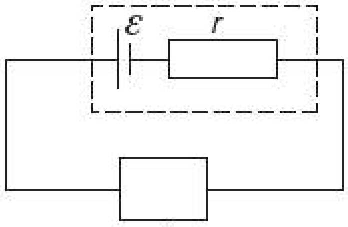
A $2R$

B $4R$

C $\frac{R}{2}$

D $\frac{R}{4}$

Q7 A cell of emf ε and internal resistance r delivers current to a small electric motor.



450 C of charge flows through the motor and 9000 J of energy are converted in the motor. 1800 J are dissipated in the cell. The emf of the cell is

A 4.0 V

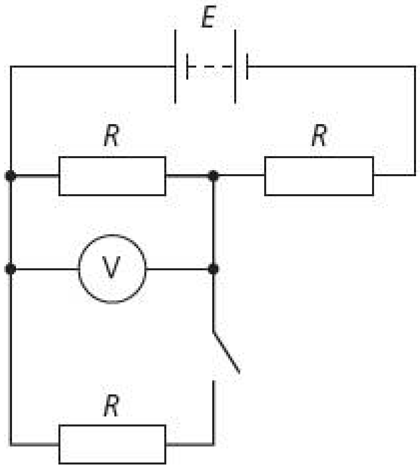
B 16 V

C 20 V

D 24 V

Standard Level Paper 1, Nov 09, Q16

Q8 The diagram shows a circuit containing three equal resistances, a battery of emf. E and negligible internal resistance, and an ideal voltmeter.



When the switch is closed the reading on the voltmeter

A rises from zero to $\frac{E}{2}$

B rises from zero to $\frac{E}{3}$

C falls from $\frac{E}{2}$ to $\frac{E}{3}$

D falls from $\frac{E}{2}$ to $\frac{E}{4}$

Q9 A wire is made of a metal of resistivity ρ. The wire has length L, radius r and resistance R. What will be the resistance of another wire of resistivity 2ρ, length $L/2$ and radius $2r$?

A R

B $\frac{R}{2}$

C $2R$

D $\frac{R}{4}$

Q10 Which of the following statements about the resistance of a conductor is a correct interpretation of Ohm's law?

A Resistance is proportional to temperature.

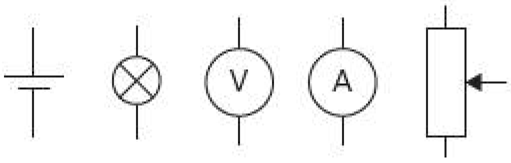
B Resistance is constant if temperature is constant.

C Resistance is proportional to potential difference if temperature is constant.

D Resistance is proportional to current if temperature is constant.

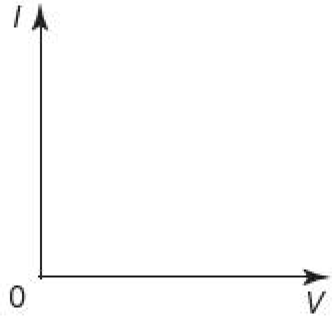
Paper 2 IB questions and IB style questions

Q1 The components shown below are to be connected in a circuit to investigate how the current I in a tungsten filament lamp varies with the potential difference V across it.



a Construct a circuit diagram to show how these components should be connected together in order to obtain as large a range as possible for values of potential difference across the lamp. [4]

b Copy the axes below and use them to sketch a graph of I against V for a filament lamp in the range $V = 0$ to its normal working voltage. [2]



Standard Level Specimen Paper A2

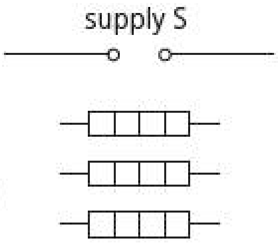
Q2 a A heating coil is to be made of wire of diameter 3.5×10^{-4} m. The heater is to dissipate 980 W when connected to a 230 V dc supply. The material of the wire has resistivity $1.3 \times 10^{-6}\,\Omega$ m at the working temperature of the heater.

i Define *electrical resistance*. [1]

ii Calculate the resistance of the heating coil at its normal working temperature. [2]

iii Show that the length of wire needed to make the heating coil is approximately 4 m. [2]

b Three identical electrical heaters each provide power P when connected separately to a supply S which has zero internal resistance. Copy the diagram and complete the circuit by drawing two switches so that the power provided by the heater may be *either* P *or* $2P$ *or* $3P$. [2]



Standard Level Paper 2, May 09 TZ2, QA2

6 Fields and forces

STARTING POINTS

- The motion of objects can be described and predicted using Newton's three laws of motion.
- Density is defined as mass per unit volume ($\rho = m/V$).
- The acceleration due to gravity is given the symbol g. It is also equal to the ratio of weight to mass.
- The centripetal force required for circular motion is given by $F = mv^2/r$.
- Electric charge is measured in coulombs, C.
- An atom contains protons and neutrons in its nucleus and electrons surrounding it.

6.1 Gravitational force and field

Universal gravitation and the inverse square law

Isaac Newton was the first to realize that if the force of gravity makes objects (like apples) fall to the Earth and also keeps the Moon in orbit around the Earth, then it is reasonable to assume that the force of gravity acts between *all* masses. This is why he called it *universal* gravitation. Newton believed that the size of the gravitational force between two masses increases with the sizes of the masses, and decreases with increasing distance between them – following an *inverse square relationship*.

The distance between the Earth and the Moon is equal to 60 Earth radii and Newton was able to prove that the centripetal acceleration of the Moon towards the Earth (from v^2/r) was equal to $g/60^2$ (see Figure 6.1).

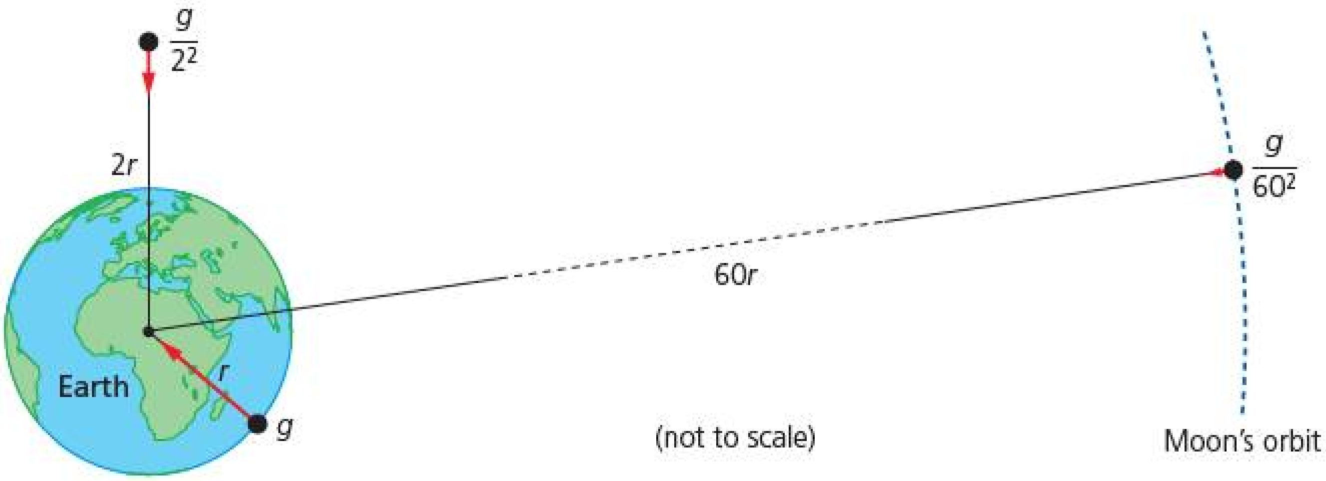


Figure 6.1 How the acceleration due to gravity varies with distance from the Earth

■ **Additional Perspectives**

Inverse square relationships

Inverse square law relationships play an important part in physics. Almost any kind of force or energy that spreads out evenly in all directions from a point obeys an inverse square law. For example, consider light radiating from a point source without being absorbed (see Figure 6.2). The further from the source, the greater the area over which the light is spread, and the fainter it becomes. The amount of light falling on each unit of area decreases with the square of the distance. This is because at twice the distance away from the source the light has to spread to cover four times the area, and at three times the distance it has to spread to cover nine times the area.

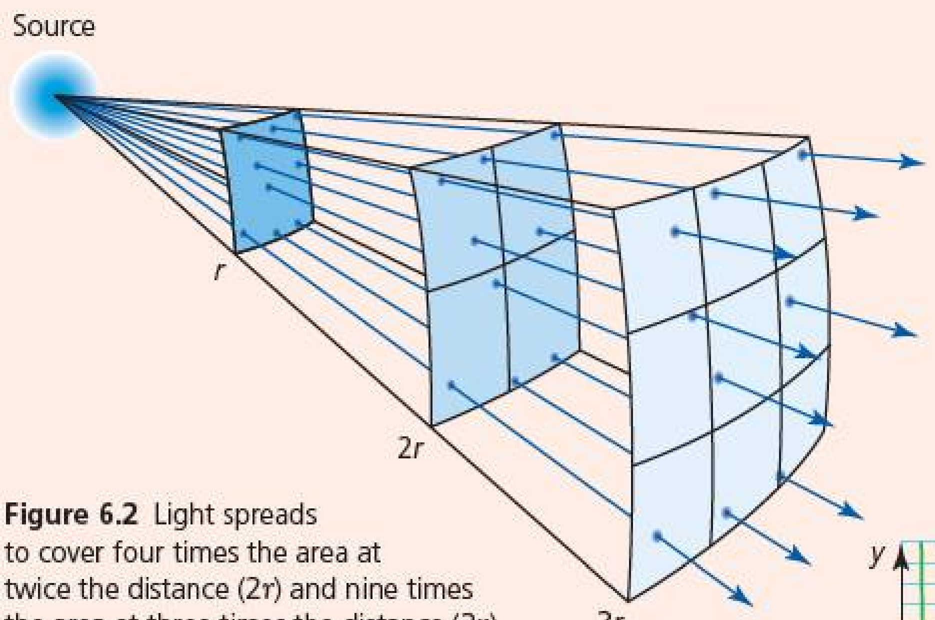


Figure 6.2 Light spreads to cover four times the area at twice the distance (2*r*) and nine times the area at three times the distance (3*r*)

Consider the data for x and y shown in Table 6.1. If $y \propto 1/x^2$, then x^2y = a constant.

It is important to know how an inverse square relationship appears on a graph. If a graph of y against x is drawn for the data in Table 6.1, its shape should be similar to Figure 6.3. Note that the line does not cross the axes. However, it is not easy to be sure from a graph like this if there is an inverse square relationship between two variables, so a more detailed check is needed.

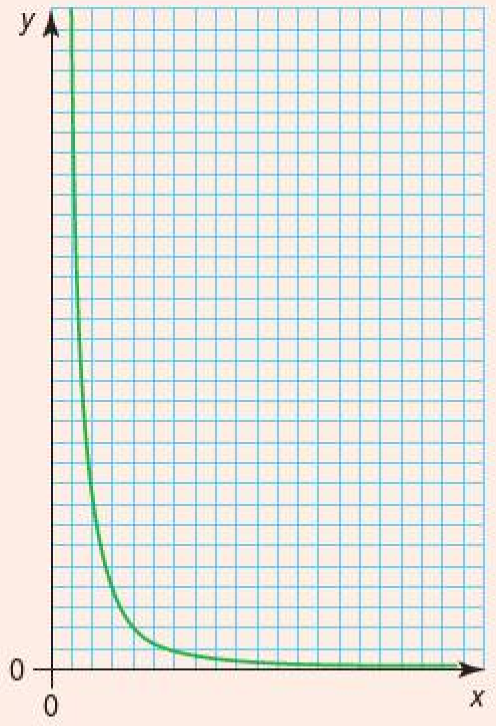


Figure 6.3 An inverse square relationship

Table 6.1 Are y and x^2 inversely proportional?

x	y
1.34	8.8
0.96	17
0.81	24
0.70	32
0.64	38
0.59	45
0.55	52
0.52	59
0.49	66

Questions

1 Use a spreadsheet to calculate values for x^2, $1/x^2$ and x^2y for the data in Table 6.1. Are x^2 and y inversely proportional?

2 Another way of checking data to see if there is certain relationship is to draw a suitable graph to see if it produces a *straight line*. Use the values you calculated in question **1** to draw a graph of y against $1/x^2$. Does it produce a straight line through the origin (which would confirm inverse proportionality)?

Newton's universal law of gravitation

6.1.1 State Newton's universal law of gravitation.

The forces acting between two **point masses** (m_1 and m_2) are proportional to the product of the masses and inversely proportional to their separation (r) squared.

$$F \propto m_1 m_2$$
$$F \propto \frac{1}{r^2}$$

Putting a constant of proportionality into the relationship, we get Newton's universal law of gravitation:

$$F = \frac{Gm_1m_2}{r^2}$$

This equation is given in the IB *Physics data booklet*.

G is known as the **universal gravitation constant**. Its value of $6.67 \times 10^{-11}\,\text{N}\,\text{m}^2\,\text{kg}^{-2}$ is given in the IB *Physics data booklet*. Its small value reflects the fact that gravitational forces are small unless one (or both) of the masses is very large. G is a *fundamental* constant which, as far as we know, always has exactly the same value everywhere in the universe. It should not be confused with g, the acceleration due to gravity which varies with location.

The relationship between force and distance is illustrated in Figure 6.4. Note that exactly the same force always acts on both masses (but in opposite directions), even if one mass is larger than the other. This is an example of Newton's third law of motion.

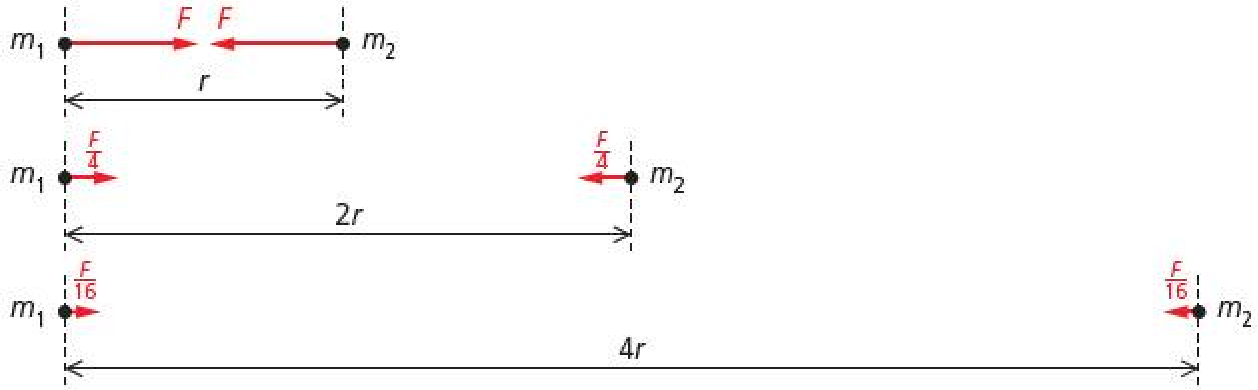


Figure 6.4 The gravitational force between point masses m_1 and m_2 decreases with increasing separation (the vectors are not drawn to scale)

Of course, the mass of an object is not all located at one point, but this does not mean that Newton's equation cannot be used for real masses. The forces between two spherical masses of uniform density located a long way apart are the same as if the spheres had all of their masses concentrated at their centre points.

Worked example

1 Calculate the gravitational force acting between the Earth and a 1.0 kg book on the Earth's surface. (The Earth's mass is 6.0×10^{24} kg and its radius is 6.4×10^6 m.)

$$F = \frac{Gm_1m_2}{r^2}$$

$$F = \frac{(6.67 \times 10^{-11}) \times 1.0 \times (6.0 \times 10^{24})}{(6.4 \times 10^6)^2}$$

$$F = 9.8\text{N}$$

This is the **weight** of a 1.0 kg mass on the Earth's surface. The book attracts the Earth up towards it with an equally sized force, which, of course, has a negligible effect on the Earth.

6.1.5 **Solve** problems involving gravitational forces and fields.

1 Estimate the gravitational force between you and your pen when you are 1 m apart.

2 What is the gravitational force between two steel spheres each of radius 45 cm and separated by 10 cm? (The density of steel is 7900 kg m^{-3}.)

3 Calculate the average gravitational force between the Earth and the Sun. (You will need to research the relevant data.)

4 A proton has a mass of 1.7×10^{-27} kg, and the mass of an electron is 9.1×10^{-31} kg. Estimate the gravitational force between these two particles in a hydrogen atom, assuming that they are 5.3×10^{-11} m apart.

Gravitational fields

A region (around a mass) in which another mass would experience a gravitational force is called a gravitational **field**. We all live in the gravitational field of the Earth, while the Earth moves in the gravitational field of the Sun. Gravitational forces can be very small if the masses are small or far apart, in which case the fields may be totally insignificant but, in theory, they never reduce completely to zero.

We often want to represent a gravitational field on paper or on a computer screen, and this can be done with gravitational **field lines** as shown in Figure 6.5. The lines and arrows show the direction of the gravitational force that would be experienced by a mass placed at a particular place in the field. Figure 6.5a represents the spreading **radial** field lines around the Earth.

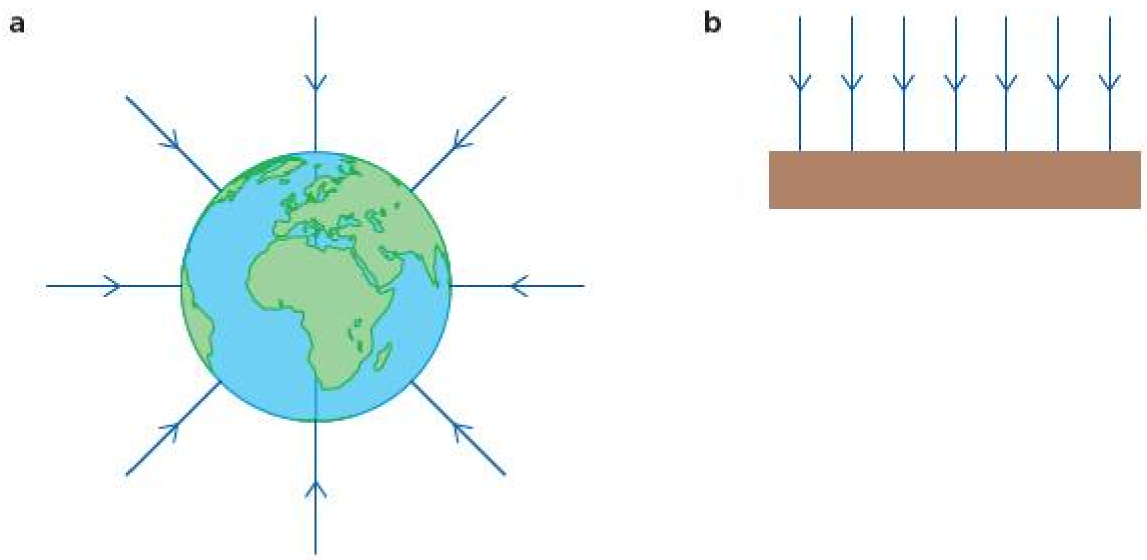


Figure 6.5 Field lines are used to represent gravitational fields on paper or on screen: **a** radial field; **b** uniform field

The lines are closer together nearer to the Earth, which shows that the field is stronger. Field lines never cross each other; this would mean that gravitational force is acting in two different directions at the same place. The parallel lines in Figure 6.5b represent a **uniform** gravitational field, such as in a small region of the Earth's surface where variations in field are negligible.

TOK Link: Fields

The fact that one mass can affect another mass without any contact or other matter between them is very difficult to understand and explain: how can a mass 'know' that there is another mass affecting it when there is no connection between them? (This is sometimes called 'action at a distance'.) Similar comments can be made about the forces between electric charges. Adding to the puzzle, there seems to be no time delay between the movement of a mass and the gravitational effect of that movement somewhere else.

By giving a force a name (for example, 'gravitational force') and calling the space in which the force can be detected a 'field', we may feel that we understand it better, but do we really? This raises the interesting question of what we mean when we say that we 'understand' something.

As always, the real usefulness of physics is in making calculations and predictions, and Newton's law (and Coulomb's law, see later) is fundamentally and extremely useful in this respect. These laws predict that the fields which they describe extend indefinitely. In practice, of course, forces become immeasurably small if the distances involved get very large.

Einstein's theory of general relativity has provided a different interpretation of gravitational force and field in terms of the curvature of space–time, but that does not reduce the importance of Newton's universal law of gravitation any more than it reduces the usefulness of any of Newton's laws of motion. Quantum electrodynamics (QED) removes the need for the concept of electric field, instead using the idea of virtual particles to cause the forces.

There are certain phenomena in physics for which there seems to be no simpler explanation and we may be tempted to think 'that is just the way the universe is', not delving any deeper. However, the search for fundamental 'truths' is one of the defining features of physics.

The study of gravitational, electromagnetic and nuclear forces, and the fields we use to describe them, is a central part of physics because these forces have produced the world and the universe that we see around us. Trying to see similarities between these fundamental forces and fields has long been an issue for physicists trying to develop the concept of a single unified force.

Questions

1 Which of the physics theories you have studied can be described as truly fundamental?

2 G is described as a *fundamental* constant. What does this mean? Give some other examples.

Gravitational field strength

6.1.2 Define *gravitational field strength.*

We may want to ask the question, 'if a mass was put in a particular place, what would be the gravitational force on it?' The answer, of course, depends on the mass, so it is more helpful to generalize and ask, 'what is the force on a unit mass (1 kg)?' If we know this, then we can easily calculate the gravitational force on any other mass.

Gravitational field strength is defined as the force per unit mass that would be experienced by a small test mass placed at that point.

Reference is made to a 'small test mass' because a large mass (compared to the mass creating the original field) would have a significant gravitational field of its own.

Gravitational field strength is given the symbol g and has the unit newtons per kilogram, $\mathrm{N\,kg^{-1}}$. Field strength is a vector quantity and its direction is shown by the arrows on field lines.

$$\text{gravitational field strength} = \frac{\text{gravitational force}}{\text{mass}}$$

$$g = \frac{F}{m}$$

This equation is given in the IB *Physics data booklet*.

In general we know, from Newton's second law of motion, that $a = F/m$, so that gravitational field strength ($g = F/m$) in $\mathrm{N\,kg^{-1}}$ is numerically equal to the acceleration due to gravity in $\mathrm{m\,s^{-2}}$.

Imagine you were on an unknown planet and wanted to find experimentally the gravitational field strength. This can be done easily by hanging a mass of 1 kg on a forcemeter. The reading will be the strength of the gravitational field (in $\mathrm{N\,kg^{-1}}$) and the direction of the field will be the same as the direction of the string: 'downwards' towards the centre of the planet.

Determining gravitational field strength

6.1.3 Determine the gravitation field due to one or more point masses.
6.1.4 Derive an expression for the gravitational field strength at the surface of a planet, assuming that all its mass is concentrated at its centre.

Since the gravitational force $F = Gm_1m_2/r^2$, the gravitational field strength g around a point mass m can be found in terms of G by substituting for F in the equation $g = F/m$. If m_2 is a test mass placed at a distance of r from a mass m ($= m_1$), then:

$$g = \frac{Gmm_2/r^2}{m_2}$$

$$g = \frac{Gm}{r^2}$$

This equation, although derived for a point mass, can also be used to determine the field on the surface, or outside of, a spherical mass, such as a planet or moon. To determine the field on the surface of a planet, we replace r with r_P, the radius of the planet. We can use an average density to calculate the mass, but we must then assume that the mass is concentrated at the centre of the sphere. Like gravitational force, gravitational field strength follows an inverse square law with distance. This is sketched in Figure 6.6.

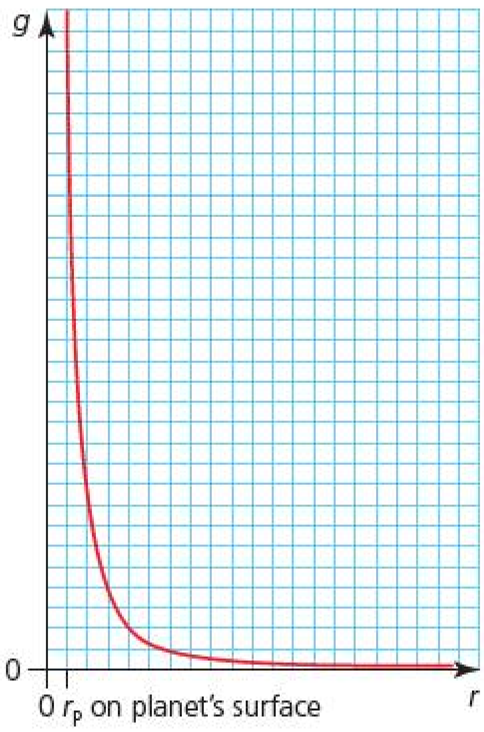


Figure 6.6 A planet's gravitational field strength, g, decreases with distance from the centre of the planet, r, showing an inverse square relationship

To investigate how the gravitational field strength depends on a planet's radius we need to use these facts:

- mass is equal to density, ρ, multiplied by volume, V
- the volume of a sphere equals $\frac{4}{3}\pi r^3$

So, we can write:

$$m = \rho\frac{4}{3}\pi r_P{}^3$$

The density of a planet is not uniform, so the value used here is an average. Putting this equation for m back into the equation for g we get:

$$g = G\rho\frac{4}{3}\pi\frac{r_P^3}{r_P^2}$$

$$g = \frac{4}{3}G\pi\rho r_P$$

This equation predicts that the gravitational field strength at the surface of a planet is proportional to its radius.

From the equation we would expect bigger planets to have stronger fields, but that is only true if they have equal average densities. (The Earth is the densest planet in our solar system, with an average density of 5510 kg m^{-3}. Venus and Mercury have similar densities to Earth, but the density of Mars is significantly lower. The outer planets are gaseous and have lower densities. Saturn has the lowest density, at 687 kg m^{-3}.)

Worked example

2 a Calculate the gravitational field strength on the surface of the Moon. The mass of the Moon is 7.35×10^{22} kg and its radius is 1740 km.

b Calculate the gravitational field strength at a point on the Earth's surface due to the Moon (not the Earth) assuming that the distance between the centre of the Moon and the Earth's surface is 3.8×10^8 m.

c Calculate the gravitational field strength on the surface of the planet Venus (radius = 6050 km, average density = 5.2×10^3 kg m^{-3}).

a $g = \frac{Gm}{r^2}$

$$g = \frac{(6.67 \times 10^{-11})(7.35 \times 10^{22})}{(1.74 \times 10^6)^2}$$

$g = 1.62$ N kg^{-1} (This is about $\frac{1}{6}$ of the Earth's gravitational field strength.)

b $g = \frac{Gm}{r^2}$

$$g = \frac{(6.67 \times 10^{-11})(7.35 \times 10^{22})}{(3.8 \times 10^8)^2}$$

$g = 3.4 \times 10^{-5}$ N kg^{-1}

c $g = \frac{4}{3}G\pi\rho r_P$

$$g = \frac{4}{3} \times (6.67 \times 10^{-11}) \times \pi \times (5.2 \times 10^3) \times (6.05 \times 10^6)$$

$g = 8.8$ N kg^{-1}

6.1.5 Solve problems involving gravitational forces and fields.

5 The radius of the planet Mercury is 2440 km. Its mass is 3.3×10^{23} kg. Calculate the gravitational field strength on the surface of Mercury.

6 a What is the gravitational field strength at a height of 300 km above the Earth's surface? (The radius of the Earth is 6.4×10^6 m.) Many satellites orbit at about this height.
b What percentage is this of the accepted value for the gravitational field strength on the Earth's surface?

7 Calculate the gravitational field strength on the surface of a planet which has a radius of 8560 km and an average density of 4320 kg m^{-3}.

8 What would the gravitational field strength be on a planet that had twice the radius of Earth and the half the density?

9 Use a spreadsheet to calculate data and draw a graph showing how the Earth's gravitational field strength varies from its surface up to a height of 50 000 km.

10 a Find out the name and details of the largest moon in our solar system and the planet around which it orbits.
b Calculate the gravitational field strength on its surface.

Additional Perspectives

Weighing the Earth

At the time Newton proposed his law of universal gravitation it was not possible to determine an accurate value for the gravitational constant, G. The only gravitational forces that could be measured were those of the weights of given masses on the Earth's surface. The radius of the Earth was also known, but that still left two unknowns in the equation $F = Gm_1m_2/r^2$; the gravitational constant and the mass of the Earth. If either of these could be found, then the other could be calculated using Newton's law of gravitation. That is why the determination of an accurate value for G was known as 'weighing the Earth'.

Figure 6.7 A modern version Cavendish's apparatus

Certainly it was possible in the 17th century to get an approximate value for the mass of the Earth from its volume and estimated average density (using $m = \rho V$). But density estimates would have been little more than educated guesses. We know now that the Earth's crust has a much lower average density (about $3000\,kg\,m^{-3}$) than most of the rest of the Earth. However, it was possible to use an estimate of the Earth's mass to calculate an approximate value for the gravitation constant. The first accurate measurement was made more than 100 years later by Cavendish in an experiment which is famous for its precision and accuracy.

To calculate a value for G without needing to know the mass of the Earth (or the Moon, or another planet) required the direct measurement of the force between two known masses. Cavendish used lead spheres (see Figure 6.7) because of their high density ($11.3\,g\,cm^{-3}$). The forces involved are very difficult to measure because they are so small, but also because similar sized forces can arise from various environmental factors. (In fact, Cavendish's main aim was to get a value for the density of the Earth rather than to measure G.)

Questions

1 a Calculate the gravitational forces between two identical 4.0 kg lead spheres with their centres 10 cm apart.
  b What is the distance between the surfaces of the spheres?
  c Estimate the weight of a grain of salt and compare your answer to the gravitational force calculated in **a**.

2 In an early attempt to estimate the gravitation constant and calculate a value for the mass of the Earth, pendulums were suspended near mountains (see an exaggerated representation in Figure 6.8).
  a What measurements should have been taken?
  b Suggest why such experiments were unlikely to be very accurate.
  c Find out what you can about Nevil Maskelyne and a Scottish mountain.

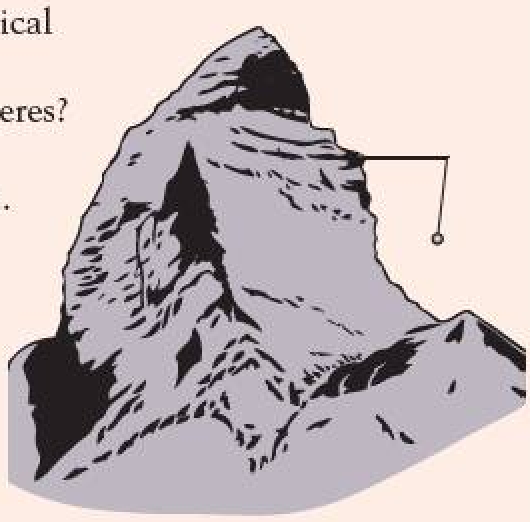


Figure 6.8 A pendulum and a mountain attract each other

Calculating combined gravitational field strengths of planets and/or moons

6.1.3 Determine the gravitational field due to one or more point masses.

It is possible that a mass may be in two or more separate gravitational fields. For example, we are in the fields of both the Earth and the Moon. However, using the values calculated in Worked example 2 on page 197, on the Earth's surface the two fields are in the ratio $9.81/(3.4 \times 10^{-5})$, or about 300 000 : 1. In other words, on the Earth's surface the Moon's gravitational field is almost negligible compared to the Earth's field. But, if a spacecraft is travelling from the Earth to the Moon, the gravitational field due to the Earth will get weaker as the Moon's field gets stronger. There will be a point at which the two fields will be equal in strength but opposite in direction (shown as P in Figure 6.9).

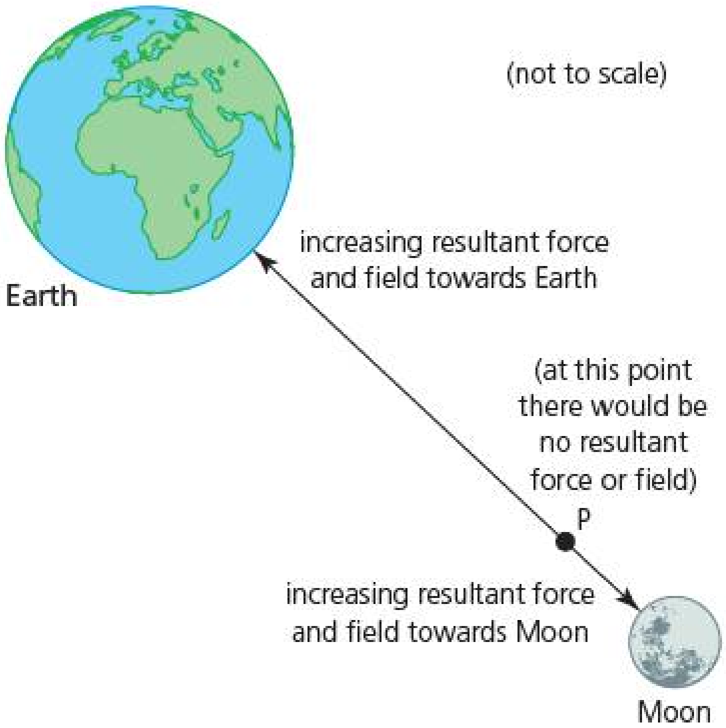


Figure 6.9 Opposing fields cancel at a precise point, P, between the Earth and the Moon

At P, the total gravitational field strength is zero and there will be no

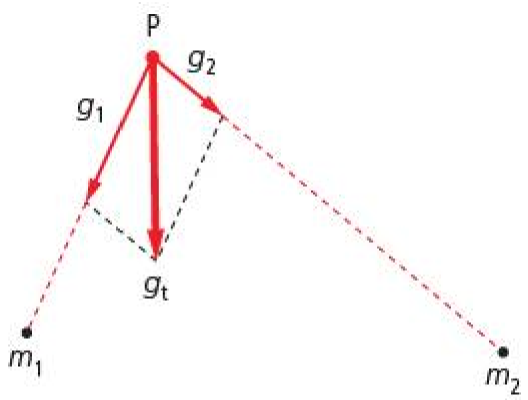


Figure 6.10 Combining the fields due to point masses m_1 and m_2

resultant force on the spacecraft because the pulls of the Moon and the Earth are equal and opposite. As the spacecraft travels from the Earth to P there is a resultant force pulling it back to Earth but this is reducing in size. After the spacecraft passes P there will be an increasing resultant force pulling the spacecraft towards the Moon.

In general, if two or more masses are creating gravitational fields at a certain point, then the total field, g_t, is determined by adding the individual fields, remembering that they are *vector quantities*. Figure 6.10 shows how to do this graphically.

However, if the location being considered is somewhere on the line joining the two masses, then the vector addition of the two fields is straightforward, as shown in Worked example 3.

Worked example

3 In Figure 6.11, P is a point midway between the centres of the planets A and B. At P the gravitational field strength due to A is $4.0\,\text{N kg}^{-1}$ and that due to B is $3.0\,\text{N kg}^{-1}$.

a What is the total gravitational field strength at P?

b What is the combined gravitational field strength at a point Q, which is the same distance as P from A?

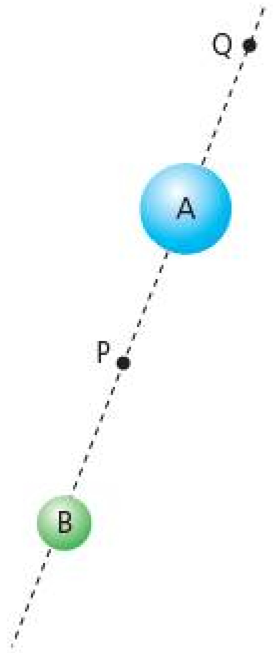


Figure 6.11

a Taking the field towards the bottom of the diagram to be positive,

$$-4.0 + 3.0 = -1.0$$

The gravitational field strength is $1.0\,\text{N kg}^{-1}$ towards A.

b The size of the field due to A is the same at Q as it was at P, although it is in the opposite direction. The strength of the field due to B at Q is 3^2 times less than at P because it is three times further away, but it is in the same direction.

$$4.0 + \left(\frac{3.0}{9}\right) = 4.3\,\text{N kg}^{-1} \text{ towards A and B}$$

6.1.5 Solve problems involving gravitational forces and fields.

11 Figure 6.12 is drawn to scale and represents two spherical masses, A and B, of equal density.

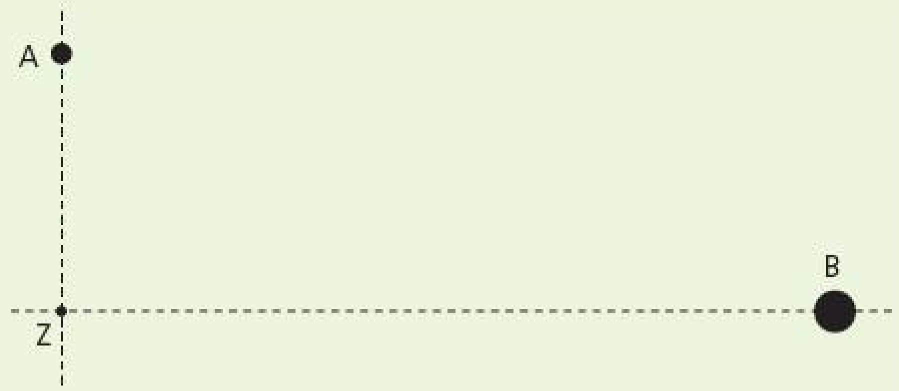


Figure 6.12

a What is the ratio of the masses?
b If the size of the gravitational field strength due to mass A at point Z is $0.064\,\text{N kg}^{-1}$, what is the size of the gravitational field strength due to B?
c Calculate the size and direction of the resultant gravitational field strength.

12 a Research the data that will allow you to set up a spreadsheet to calculate the gravitational field strengths due to the Earth and the Moon at points along a straight line joining their surfaces.
b Combine the fields to determine the resultant field and draw a graph of the results.
c Where does the resultant gravitational field equal zero?

13 The gravitational fields of the Sun and the Moon cause the tides on the world's oceans. The highest tides occur when the resultant field is greatest (at times of a 'new moon'). Draw a sketch to show the relative positions of the Earth, Sun and Moon when the resultant field on the Earth's surface is
a greatest
b weakest (at the time of a 'full moon').

6.2 Electric force and field

We have discussed the fundamental force of gravity that exists between all *masses*. In this section we will discuss the fundamental electric force which exists between all *charges*. Mass and charge are two of the basic properties of the particles composing all matter, and the variation of gravitational forces between masses and the variation of electric forces between charges are mathematically very similar.

Charge

6.2.1 **State** that there are two types of electric charge.

The force of gravity acts between all masses, always *attracting* them together. This force exists even between the tiny particles which make up all matter, but the gravitational force between particles is insignificantly small. However, there is another significant fundamental force that acts between some particles, and that force can be either **attractive** or **repulsive**.

We can explain this by saying that some particles are **charged**, and that there are two types of charge: **positive** and **negative**.

If a particle has a positive charge it will attract a particle that has a negative charge, whereas similarly charged particles (for example, two positive charges) will repel each other. This is illustrated in Figure 6.13.

The sub-atomic particle, the **proton**, has a positive charge of $+1.6 \times 10^{-19}$ C, and **electrons** have a negative charge of -1.6×10^{-19} C. Because they have opposite charges, we know that a proton and an electron attract each other. **Neutrons** do not have any charge and are said to be **neutral** (uncharged). Therefore, there is no electric force between neutrons and protons or neutrons and electrons.

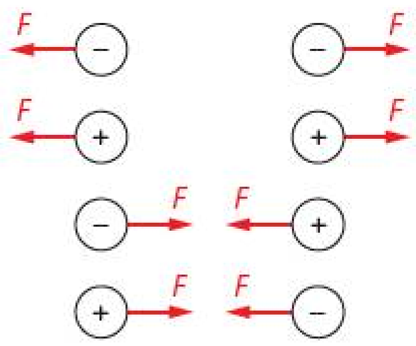


Figure 6.13 Forces between similar and opposite charges

Combining charges

6.2.2 **State** and **apply** the law of conservation of charge.

When more and more charged particles combine, the overall charge is found by simply adding up the total of the individual charges, taking their signs into account. For example, a certain chlorine atom containing 17 protons $[17 \times (+1.6 \times 10^{-19}\text{C})]$, 18 neutrons (0 C) and 17 electrons $[17 \times (-1.6 \times 10^{-19}\text{C})]$ has no overall charge – it is neutral. If an electron were to be removed, however, it would become a positively charged atom, called an **ion**, of charge $+1.6 \times 10^{-19}$ C.

The charge on electrons and protons ($\pm 1.6 \times 10^{-19}$ C) is the basic quantity of charge. It is not possible to have a smaller quantity of charge than this in a free particle. All larger quantities of charge consist of various multiples of this fundamental charge. For example, it is possible for the charge on something to be 1.6×10^{-19} C, 3.2×10^{-19} C, 4.8×10^{-19} C, 6.4×10^{-19} C, and so on, but intermediate values (for example, 2.7×10^{-19} C) do not exist (for this reason charge is said to be **quantized**).

One coulomb is a large amount of charge for an isolated 'charged' object, and more usually we deal with microcoulombs ($1\,\mu\text{C} = 10^{-6}$ C) and nanocoulombs ($1\,\text{nC} = 10^{-9}$ C).

The **law of conservation of charge** states that the total charge in any isolated system remains constant.

Under certain circumstances it may be possible to create or destroy individual charged particles, but only if the total charge of the system involved remains unchanged. Conservation laws like this are very useful tools for physicists. For example, if there are two isolated and identical spheres and one is charged with 4.6×10^{-10} C and the other is neutral, when they touch the charge will be shared so that they both have 2.3×10^{-10} C. The total charge remains the same (assuming that the spheres remain isolated). If the spheres are not identical, the charge will still be shared, but not equally.

Conductors and insulators

6.2.3 Describe and **explain** the difference in the electrical properties of conductors and insulators.

An electrical **conductor** is a material through which charges can flow (making an electric current). An electrical **insulator** is a material through which charges cannot flow.

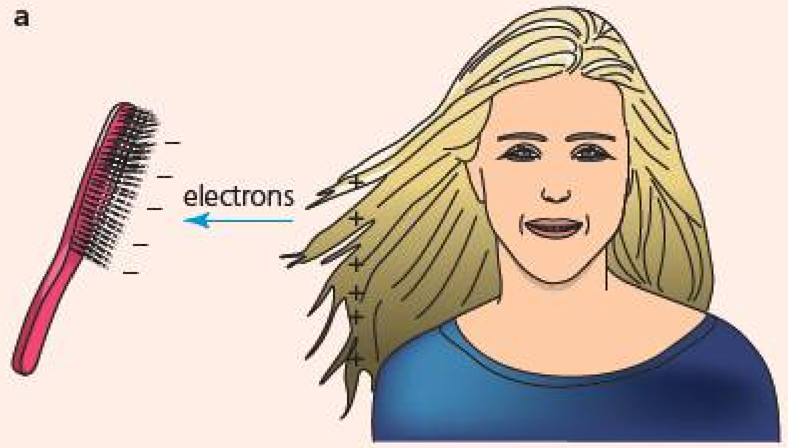
A material which is a good conductor will have a low **resistivity**, whereas insulators have high resistivities (see Chapter 5).

For a solid to be a good conductor it needs to contain a large number of electrons which can move around freely. Metals are good conductors because they contain large numbers of **free electrons** which are not bound to any particular atom. Such electrons are also commonly described as being **delocalized.** Insulators have relatively few free electrons. Semi-conductors, such as silicon and germanium, have properties between these extremes and, importantly, their ability to conduct is easily altered by adding small amounts of impurities, or by changing external properties such as temperature and light level. As any solid gets hotter the increasing vibrations of the atoms, ions or molecules get in the way of any electrons that are moving through the material, tending to increase its resistance. But in insulators and semi-conductors, the extra thermal energy frees some electrons and the material may become a better conductor (having lower resistance).

Additional Perspectives

Electrostatic effects and uses of electrostatics

Everyday objects contain an enormous number of charges, with approximately equal numbers of positive and negative charges. When we talk about a (previously neutral) object becoming charged, it will be because a very large number of electrons have been added or removed from it. One common way for this to happen is due to friction, such as when brushing dry hair with a plastic brush, as shown in Figure 6.14. If electrons are transferred in the process, one object (such as the hair brush) becomes negatively charged, while at the same time removing electrons from the other object (the hair), leaving it with a positive charge. The two objects then attract each other. In this example, individual hairs with similar charge are repelled from each other. Protons, unlike electrons, are located in the nuclei of atoms and cannot be separated and moved from their positions, so they are not involved in producing electrostatic effects.



Figures 6.14 a and b When you brush your hair, hairs move apart because of the repulsion between similar charges

Electrostatic effects tend to be noticeable under dry conditions and on insulators, rather than with conductors. This is because electrons move quickly through a metal conductor to **discharge** a charged object. If we want to be sure that there is no charge on an object, we make a good contact between the object and the ground. This is called **earthing.**

Large-scale electrostatic effects can be unwanted and even dangerous. For example, cars and planes can become charged as they move through the air or along the ground, and this can be a problem when they stop for refuelling – any sparks from a charged vehicle might cause an explosion of the fuel and air. This risk can easily be prevented by making sure that the vehicle and the fuel supply are well earthed.

Electrostatic effects can be useful. For example, electrostatic paint spraying is an efficient way of painting three-dimensional objects (see Figure 6.15) and many types of printers and photocopiers use charge to make sure that the ink goes to the right places. Electrostatics also provides one important method of reducing air pollution – electrostatic precipitators can remove dust and small particles from the waste gases and smoke released from factories and power stations. A simplified version is shown in Figure 6.16.

As dust and other small particles pass through the mesh, they become negatively charged. Afterwards, as these charged particles pass near earthed metal plates, they are attracted to the plates and they stay there until they are removed. Gases pass up through the chimney to the atmosphere.

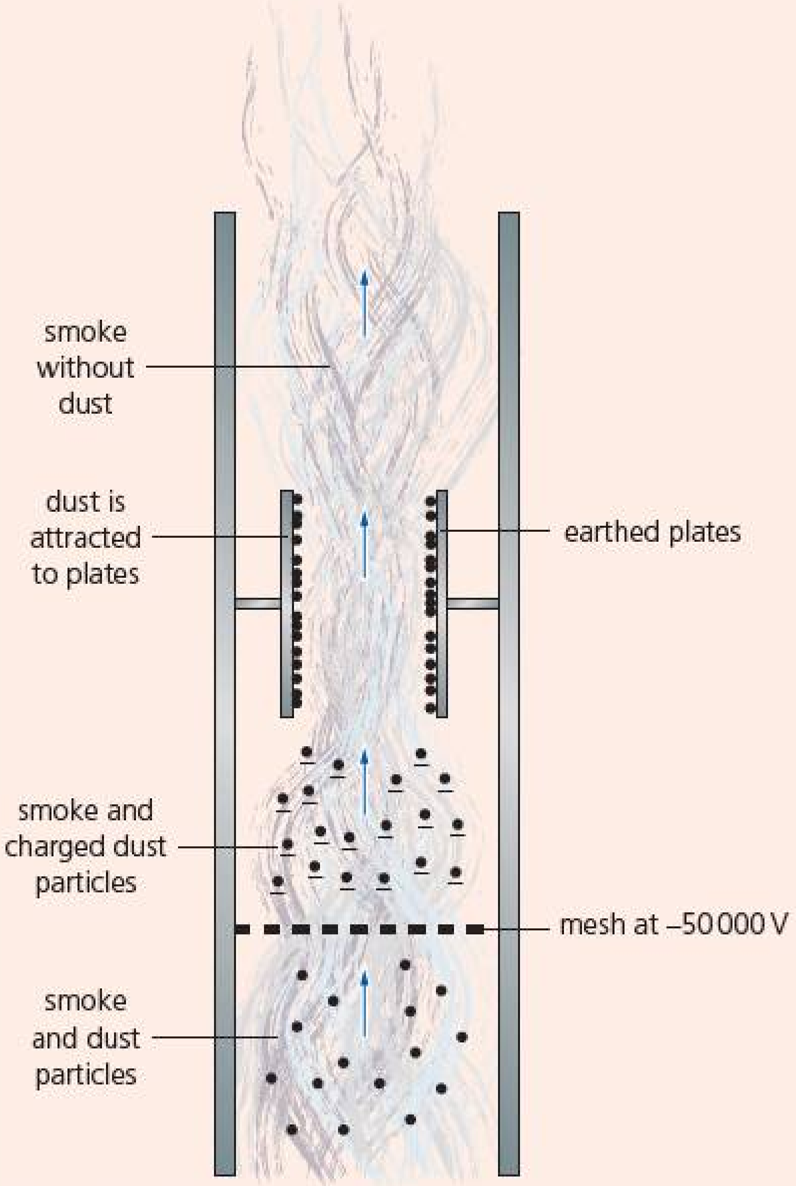


Figure 6.16 Electrostatic precipitators help reduce pollution

Figure 6.15 Electrostatic paint spraying ensures an even coating

Questions

1 Electrostatic precipitators are especially useful where there is a lot of unwanted dust and other particles, or where it is important to have a dust-free environment.
   a Electrostatic precipitators can also be used as part of air purification systems in offices and homes. Research into the advantages and disadvantages of electrostatic precipitators used in this way.
   b Give some other examples of places where electrostatic precipitators might be useful.

2 Prepare a report giving details of how electrostatic effects are used in photocopiers, computer printers, or paint spraying, or any one other application of your choice.

Coulomb's law

6.2.4 State Coulomb's law.

The greater the distance between two charged objects (or particles), the smaller the electrical force between them. This relationship follows an inverse square law in exactly the same way as the gravitational force between two masses. Figure 6.17 shows how the relationship between force and separation might be tested for two charged spheres. Two conducting spheres are charged and an electric top-pan balance is used to register the change in force as the spheres are moved closer together or further apart. This is not an easy experiment to do accurately because the forces are small and the charge on the spheres may not remain constant.

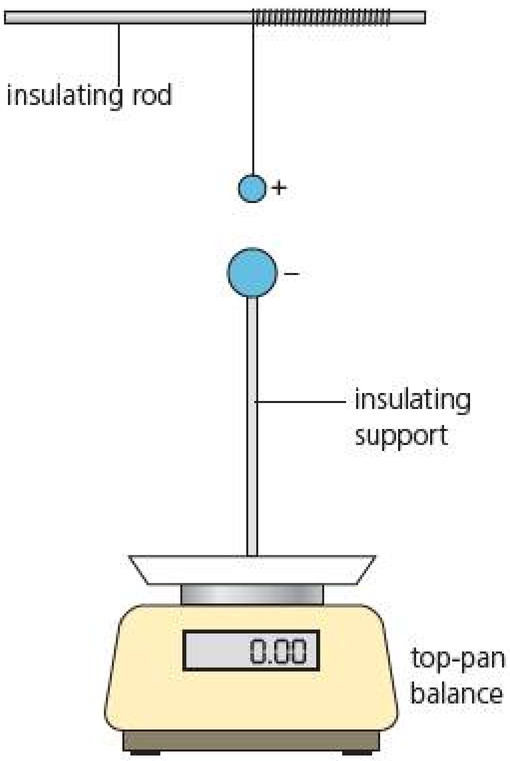


Figure 6.17 Testing Coulomb's law using a top-pan balance to measure the force between known charges

Figure 6.18 shows how the electric force vectors decrease in size as the separation of charges, r, changes. The example shown is for the repulsion of similar charges, q_1 and q_2.

By comparison with Newton's universal law of gravitation, we can write:

$$F = \text{constant} \times \frac{q_1 q_2}{r^2}$$

where q_1 and q_2 are the two charges and r is the distance between them.

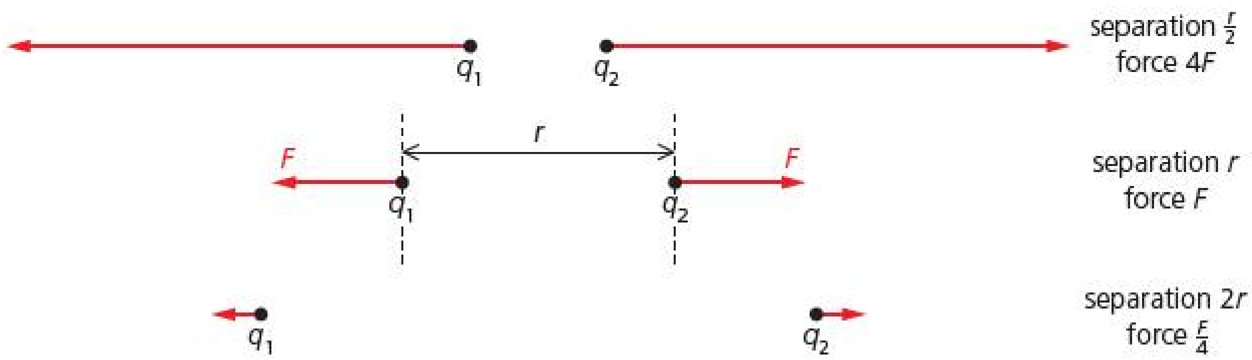


Figure 6.18 Repulsive force varying with distance between similar charges (vector arrows are not drawn to scale)

As with Newton's law of gravitation, this relationship applies only for charges concentrated at a point. However, charged spherical objects behave as if all their charge were concentrated at a point (the centre).

Putting in the symbol k for the Coulomb constant, we get

$$F = \frac{kq_1 q_2}{r^2}$$ This is called **Coulomb's law.**

The law was first published by Charles Augustin de Coulomb in 1783 (Figure 6.19). This equation is given in the IB *Physics data booklet*, as is the value of k ($= 8.99 \times 10^9 \text{N m}^2 \text{C}^{-2}$).

The Coulomb constant is sometimes expanded to the form:

$$k = \frac{1}{4\pi\varepsilon_0}$$

Figure 6.19 French physicist Charles Augustin de Coulomb (1736–1806)

where ε_0 is called the electrical **permittivity** of 'free space' (vacuum) and has a value of $8.85 \times 10^{-12} \text{C}^2 \text{N}^{-1} \text{m}^{-2}$.

The expanded form of Coulomb's law $\left(F = \frac{q_1 q_2}{4\pi\varepsilon_0 r^2}\right)$ and the value of ε_0 are also given in the IB *Physics data booklet*.

Electric forces and fields can pass through a vacuum and the permittivity of free space is a fundamental constant which represents the ability of a vacuum to transfer an electric force and field.

The permittivities of other substances are all greater than ε_0, although dry air has similar electrical properties to free space.

Similarities and differences between electric and gravitational forces

Newton's and Coulomb's laws both apply to the forces around points (masses and charges), spreading radially into the space around them, so it is not surprising that the two equations have similar forms. But, of course, there are also differences between electrical and gravitational forces:

- The electrical force between two isolated point charges is much, much larger than the gravitational force between two isolated point masses. A comparison of the size of the two constants in the equations demonstrates this difference.
- As far as we know there is only one kind of mass, but there are two kinds of charge; it seems that there is no such thing as a repulsive gravitational force between masses.
- As masses get larger and larger, so do the gravitational forces involved. When objects get larger and larger there will normally still be (approximately) equal numbers of positive and negatively charged particles, so electrical forces do not tend to increase with size.
- On the microscopic scale of atoms, ions, molecules and other particles, electrical forces dominate and gravitational forces are negligible. However, on the large scale, the only significant forces between planets and stars are gravitational.
- The value of the gravitational force between two masses in a certain arrangement will always be the same. The value of the electrical force between two charges depends on the electrical permittivity of the medium in which they are located, although we often assume that the medium is air or a vacuum.

Worked examples

4 What is the electric force between a negatively charged sphere, which has a charge of 0.82 μC, placed with its centre 21 cm away from the centre of a sphere charged with +0.47 μC?

$$F = \frac{kq_1q_2}{r^2}$$

$$F = \frac{(8.99 \times 10^9)(8.2 \times 10^{-7})(4.7 \times 10^{-7})}{0.21^2}$$

$$F = 0.079\,\text{N}$$

The same force acts on both spheres, but in opposite directions.

5 Figure 6.20 shows two equally charged balls of the same mass suspended from the same point. They both hang at the same angle to the vertical. Draw diagrams to show what would happen if:

a ball 1 was replaced with a heavier ball with the same charge

b the charge on the original ball 1 was doubled.

a Ball 1 would hang at a smaller angle. Ball 2 would be unchanged.

b The force on both balls would be greater, so they would both hang at the same, greater angle to the vertical.

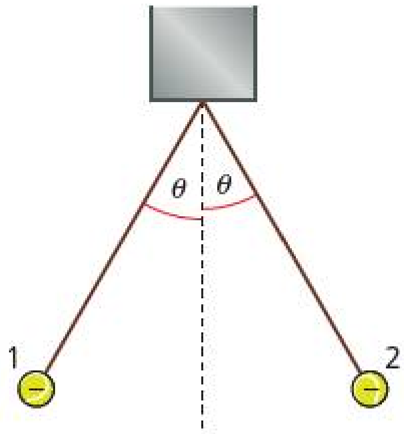


Figure 6.20

6.2.8 Solve problems involving electric charges, forces and fields.

14 How far apart would two point charges of 160 nC have to be in order that the force between them was 0.01 N?

15 In an attempt to observe and determine a value for a small electric force, a student set up an experiment as shown in Figure 6.21. A and B are two equally charged spheres. A is fixed in position, but B is free to swing away on the end of an insulating thread.

a Draw a free-body diagram of sphere B.

b Calculate the horizontal force repelling B away from A.

c Calculate the charge on the spheres.

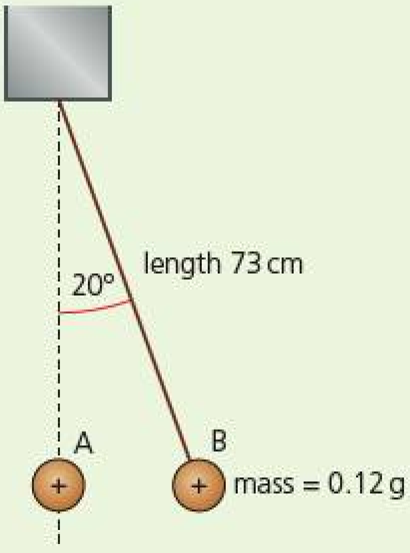


Figure 6.21

16 Calculate the resultant electric force on charge B shown in Figure 6.22.

Figure 6.22

17 Four equal charges of 0.14 μC are fixed in position at the corners of a square of sides 25 cm. Calculate the resultant electric force on any one charge.

18 **a** A proton has a charge of $+1.6 \times 10^{-19}$ C, and the charge of an electron is -1.6×10^{-19} C. Estimate the electric force between these two particles in a hydrogen atom, assuming that they are 5.3×10^{-11} m apart.
b Give an order of magnitude for the ratio of the electric force between these particles and the gravitational force (calculated in question **4**)?

19 Outline how you would use the apparatus shown in Figure 6.17 to determine the relationship between force and distance between two charged spheres.

20 Describe and explain any one electrostatic event that you have experienced in your everyday life.

21 An uncharged, lightweight, conducting sphere is hanging vertically on an insulating thread. A similar sphere which is positively charged is also hanging on an insulating thread, and is brought close to the first sphere, but without touching it. Use a diagram to explain what happens.

Electric fields

6.2.5 **Define** *electric field strength.*

An **electric field** is a region of space in which a charge experiences a force.

Electric field strength is defined as the force per unit charge that would be experienced by a small *positive* test charge placed at that point.

The similarity with the definition of gravitational field strength should be noted, but in an electric field the forces may be attractive or repulsive. By convention, a positive charge is chosen to define electric field. Electric field strength is a vector quantity, with direction from positive to negative.

Reference is also made here to a 'small test charge' because a large test charge (compared to the charge creating the original field) would have a significant electric field of its own.

Electric field strength is given the symbol E and has the unit newtons per coulomb, N C^{-1}.

$$\text{Electric field strength} = \frac{\text{electric force}}{\text{charge}}$$

$$E = \frac{F}{q}$$

This equation is given in the IB *Physics data booklet*.

Electric field around one or more point charges

6.2.6 **Determine** the electric field strength due to one or more point charges.

Electric field strength around a single point charge

We can use the equation for electric force, $F = kq_1q_2/r^2$ to find the electric field strength, E, around a point charge, q. This can be found by substitution into the equation $E = F/q$. If q_2 is a test charge placed at a distance of r from a charge q $(= q_1)$:

$$E = \frac{kq q_2/r^2}{q_2}$$

$$E = \frac{kq}{r^2}$$

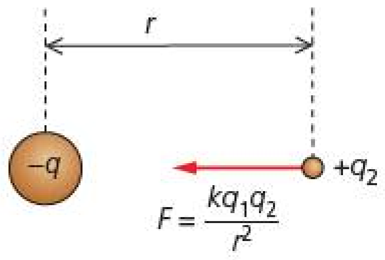


Figure 6.23 Force on a test charge placed in an electric field

Figure 6.23 shows the example of a test charge, $+q_2$, placed in the field of a negative charge $-q$.

The equation can also be used to calculate the electric field strength on the surface, or outside of, a charged sphere.

Worked example

6 What force will be experienced by a charge of +6.3 µC placed at a point where the electric field strength is 410 N C^{-1}?

$$E = \frac{F}{q} \quad \text{so} \quad F = Eq$$

$$F = 410 \times (6.3 \times 10^{-6})$$

$$F = 2.6 \times 10^{-3}\ \text{N}$$

7 A sphere has a radius of 60 cm and a charge of 1.4 µC.

a Calculate the magnitude of the field on the surface of the sphere.

b At what distance from the surface of the sphere will the field reduce to 20 000 N C^{-1}?

a $$E = \frac{kq}{r^2}$$

$$E = \frac{(8.99 \times 10^9)(1.4 \times 10^{-6})}{0.6^2}$$

$$E = 3.5 \times 10^4\ \text{N C}^{-1}$$

b $$E = \frac{kq}{r^2}$$

$$20\,000 = \frac{(8.99 \times 10^9)(1.4 \times 10^{-6})}{r^2}$$

r = 79 cm from the centre, or 19 cm from the surface of the sphere.

6.2.8 Solve problems involving electric charges, forces and fields.

22 **a** What is the magnitude of an electric field at a point where a test charge of 47 nC experienced a force of 6.7×10^{-5} N?

b If a charge of 0.28 µC was placed at the same point, what force would be exerted on it?

23 **a** What is magnitude of the electric field at a point which is 2.0 m away from the centre of a sphere carrying a charge of $+7.3 \times 10^{-5}$ C?

b What would be the force on a charge of +0.56 µC placed at that point?

24 **a** What charge on a sphere will produce an electric field of strength 5.3×10^5 N C^{-1} at a distance of 76 cm from its centre?

b At what distance from the centre will the field be ten times greater?

25 **a** What is the electric field at a distance of 5.3×10^{-11} m away from a single proton?

b Use $F = Eq$ to calculate the force on an electron at this distance. (Your answer should be the same as calculated in question **18**).

c What force is experienced by the proton?

d What affect does the force have on the electron?

Electric field strength along a straight line passing through charges

At any point on a straight line which passes through two point charges, we can calculate the magnitude of the combined field simply by adding the individual fields, taking their directions into account. The resultant is always directed along the line between the point charges. Consider Figure 6.24a: it shows two positive charges, A and B, and a test charge $+q$ in various positions. The direction of the field is the direction of the force experienced by the positive test charge. Along the line to the left of A, both fields are acting in the same leftwards direction. The field due to A will probably be stronger, unless B is a much greater charge than A. Similarly, to the right of B, the fields are both acting in the same direction. To the left of A and to the right of B the combined field will be stronger than either individual field.

Between A and B the two fields act in opposite directions and there must be a position where the resultant electric field is zero. If the two charges are equal in magnitude as well as sign, this position will be the midpoint between them. If one charge is larger than the other, then the zero electric field will occur closer to the smaller charge.

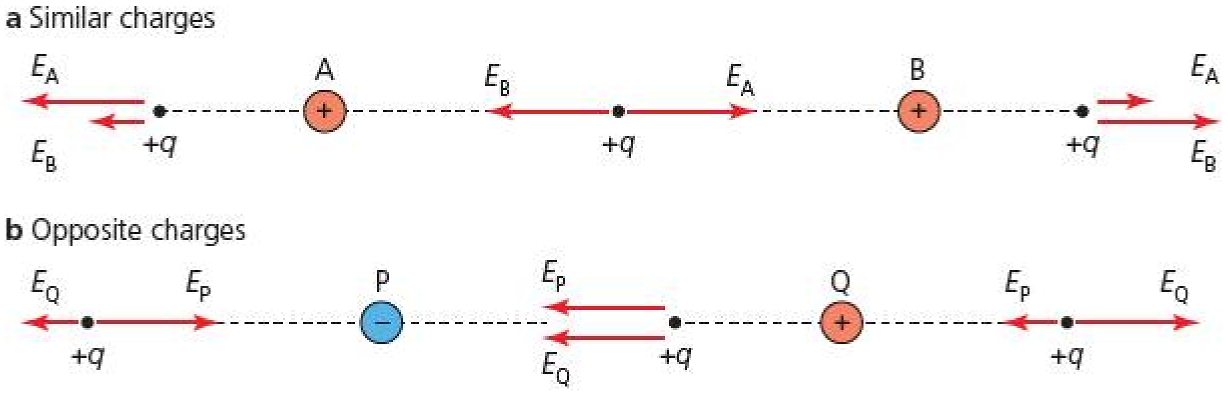


Figure 6.24 The arrows represent the electric fields acting along a line between the two point charges A and B, and P and Q

In Figure 6.24b the two charges have opposite signs and, at any point on the line between them, the two fields will combine to give a stronger electric field. To the left of P and to the right of Q the individual fields act in opposite directions and the combined electric field will be reduced. If the two charges have equal magnitudes, the combined field will never reduce to zero, but the field may be reduced to zero if one charge is greater than the other.

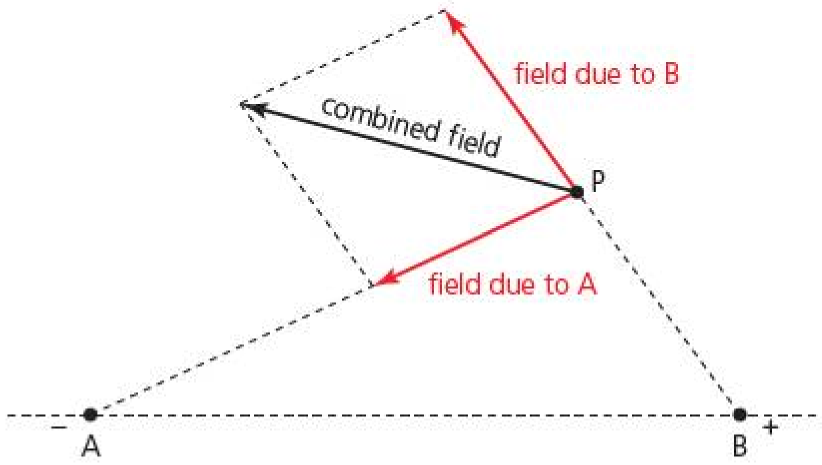


Figure 6.25 How to find the electric field strength around two or more charges using a parallelogram

Electric field strength anywhere around two or more charges

We may want to determine the combined field at a position which is not on a straight line through the point charges. In Figure 6.25, the individual radial electric fields at position P due to the two point charges, A and B, are shown by red arrows. The combined field, shown in black, can be found from the parallelogram, as shown.

6.2.8 Solve problems involving electric charges and fields.

26 Two identical positively charged spheres, each with charge +2.3 μC are placed with their centres 40 cm apart. What are the magnitudes of the electric field strengths
- **a** midway between them
- **b** 10 cm from the centre of one of the spheres along a line passing through their centres (two possible answers)?

27 Two charges of −3.8 μC and +1.6 μC are placed with their centres 20 cm apart as in Figure 6.26. Calculate the magnitude and direction of the electric field at point P.

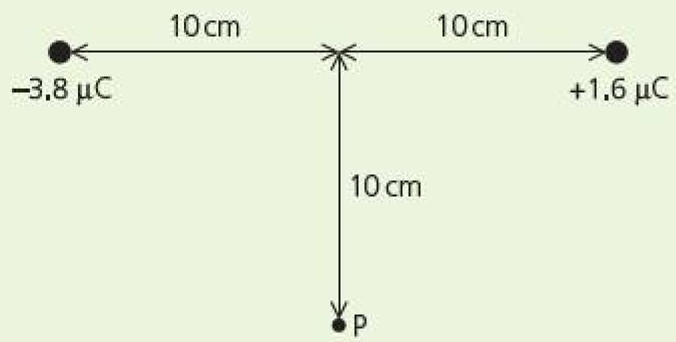


Figure 6.26

Electric field patterns

6.2.7 **Draw** the electric field patterns for different charge configurations.

Electric field lines show the direction of force on a positive charge; that is, they point from positive to negative. Field lines can never touch or cross each other.

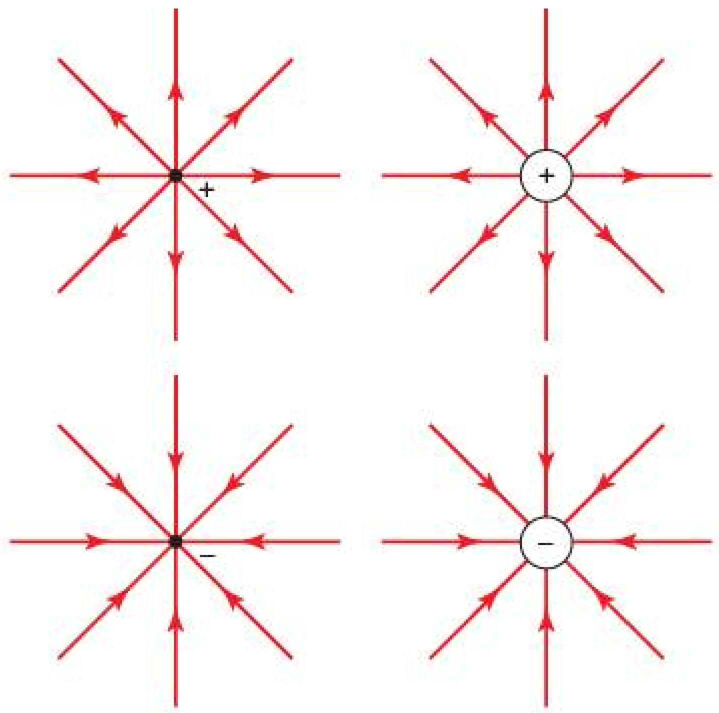


Figure 6.27 Simple radial electric fields around point charges and charged spheres

Radial fields around point charges and spheres

The simplest electric fields are the **radial fields** around single point charges or charged spheres, as shown in Figure 6.27. In a radial field, the lines spread out in straight lines from a point. As the distance from the charge increases the field gets weaker, which is shown by the fact that the field lines get further apart.

Uniform fields

Sometimes it is necessary to create a **uniform** electric field and this can be done by connecting a potential difference across parallel metallic plates. In Figure 6.28a, the cell will attract electrons off the top plate and repel them onto the bottom plate. The charges spread out evenly and create a uniform electric field between the plates, as shown from the side in Figure 6.28b. Field lines are always perpendicular to the surface of charged conductors. If they were not, then there would be a component of the electric field making the charges move along the surface.

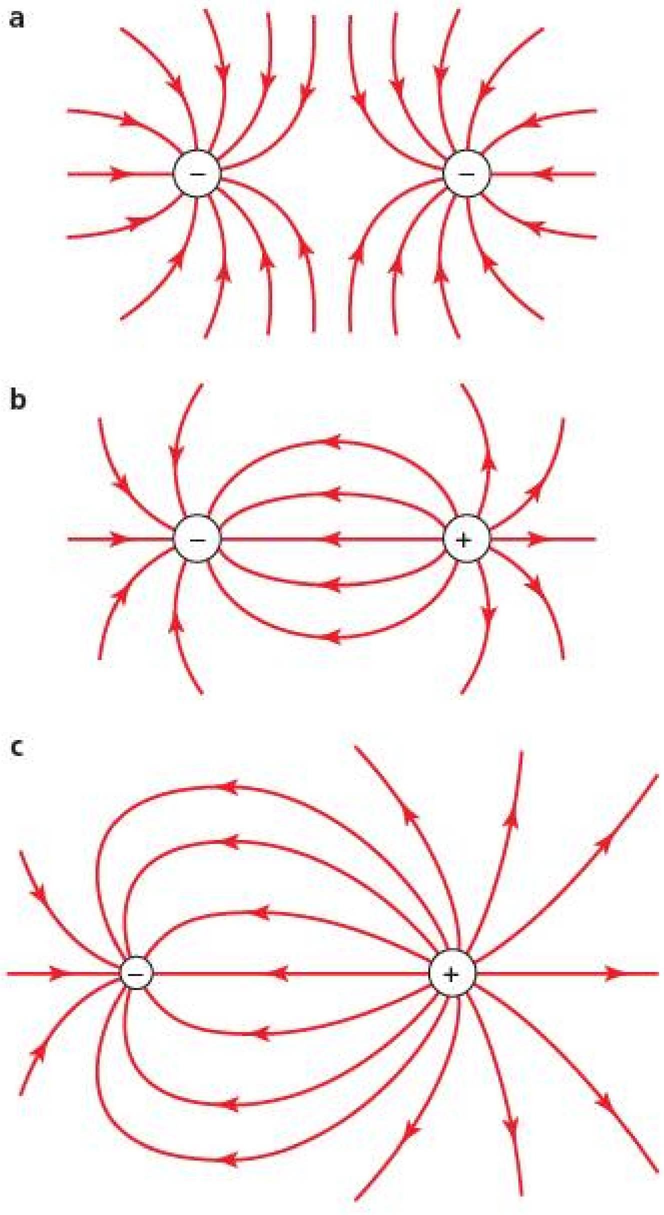


Figure 6.29 Field patterns around two charges

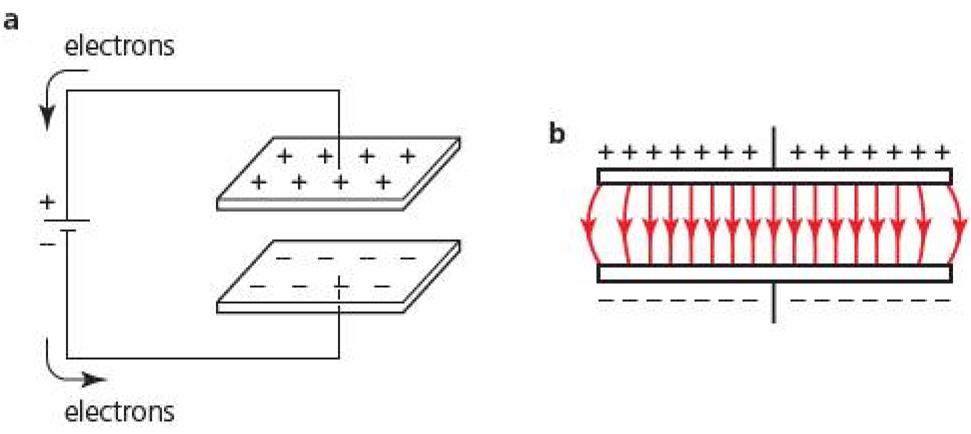


Figure 6.28 Uniform field between parallel metal plates

The electric field is uniform in the central region of the plates but falls to half of that value at the edge of the plates.

This is called the **edge effect** and the reduced field strength is shown by the greater separation of the field lines at the ends of the plates.

Combined fields

When two charges are placed close together the combined field patterns are as shown in Figure 6.29. In Figure 6.29a the field is due to two equal negative charges. Two positive charges would produce the same shaped field, but the directions would be reversed. Note the point in the middle where the electric field is zero because the individual fields from the two charges are equal in size but opposite in direction. Figure 6.29b shows the field due to opposite, but equal charges. In Figure 6.29c the field is not symmetrical because one charge is larger than the other.

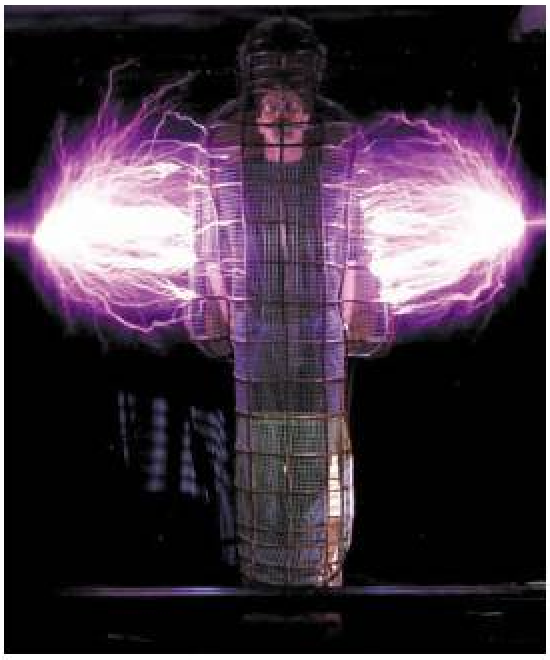


Figure 6.30 A Faraday cage, showing sparks on the outside but with someone safe inside

Electric field inside a charged conductor is zero

Note that in Figure 6.27 there are no fields *inside* the charged spheres. The fact that like charges repel each other means that any charges which are free to move on a conductor which surrounds a space will move to the outside surface, so that there will be no electric field in the space inside the conductor. A conducting surface or 'cage' surrounding people or equipment protects (or shields) them from the effects of electric fields.

Additional Perspectives

Thunder and lightning

Lightning, like that shown in Figure 6.31, will strike from a cloud to the ground (or another cloud) when the electric field between them becomes large enough to make the air conduct.

Figure 6.31 A lightning flash

Clouds contain droplets of water and sometimes ice particles, all of various sizes. These are continuously moving around and colliding with each other. During this process, charge (electrons) can be transferred between water droplets and ice particles. This can result in a large scale charge separation, with the top of a cloud becoming positively charged and the bottom of the same cloud becoming negatively charged as shown in Figure 6.32. This is just a simplification of what actually happens because the exact details are still not fully understood.

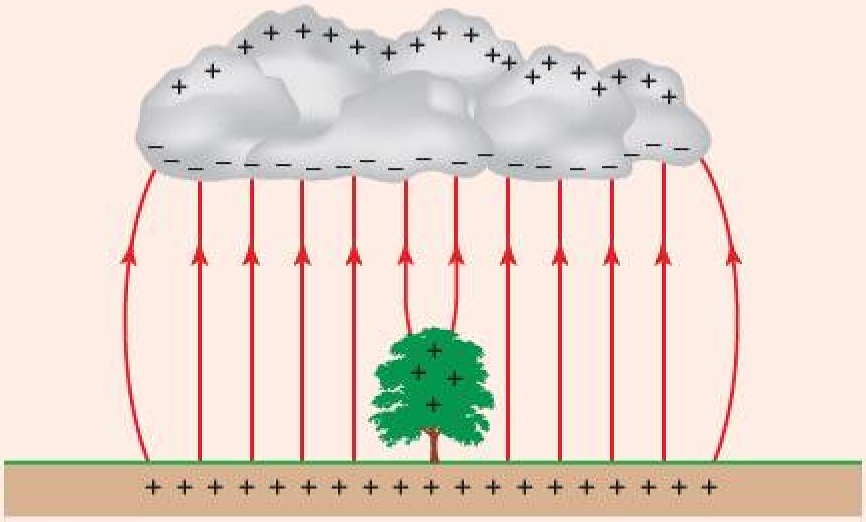


Figure 6.32 Electric field under a cloud

The large negative charge at the bottom of the cloud will repel electrons on the Earth's surface and objects on it, leaving them positively charged. A very large electric field is created between the ground and the bottom of the cloud. When the charge separation and the resulting field become large enough, a conducting path is created and lightning will strike along the path of least resistance. This is a complicated process, with charge flowing from both the Earth and the cloud. Because of the very large current involved in lightning, this is a very powerful process and a lot of energy is transferred to radiation (including light) and internal energy. The extremely rapid heating of the air by thousands of degrees causes rapid expansion and great pressure changes, resulting in the emission of sound (thunder).

Electric fields are greater near pointed objects, so anything that 'sticks out' from its surroundings generally has a greater risk of being struck by lightning. But, the electric field inside a conducting surface is always zero, so people inside cars and houses, for example, should be safe from direct lightning strikes.

Apart from the large electric fields created by charged clouds, there is a permanent, but weak, electric field of about 150 NC^{-1} above the surface of the Earth, directed downwards. This is because of a layer of positive particles surrounding the Earth (the ionosphere), which is caused by radiation from the Sun.

Questions

1 What safety advice would you give to someone caught outside in a thunder storm?
2 Research the damage that can be done to buildings that are struck by lightning. How can such damage be prevented?
3 Research the stories of people who have survived lightning strikes.

Electric field strength in a uniform field

When the charge $+q$ shown in Figure 6.33 moves from one plate to the other the energy transferred will be Vq. (From the definition of potential difference, V = energy/charge.)

The energy transferred can also be calculated from force, F, multiplied by distance, d. Hence we can write:

$$Vq = Fd \quad \text{or} \quad \frac{F}{q} = \frac{V}{d}$$

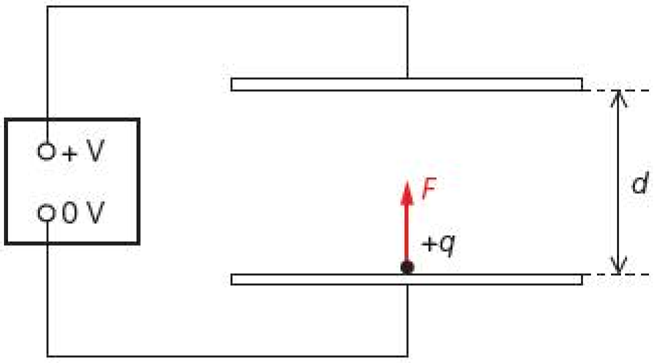


Figure 6.33 Energy transferred in moving a charge

Since electric field is defined as $E = F/q$, it should be clear that we can also calculate the strength of the uniform field between two charged parallel plates from the following equation:

$$E = \frac{V}{d}$$

Expressed in this way, the units used for electric field strength are $V\,m^{-1}$ rather than the equivalent NC^{-1}. Determining the uniform field strength between two parallel plates can then be done easily using a voltmeter and a ruler.

More generally, when a potential difference is applied along a conductor in an electrical circuit, a uniform electric field is created and the forces on the free electrons cause them to move along the conductor towards the positive terminal of the source and away from the negative terminal.

Worked example

8 A beam of electrons with a velocity of $3.7 \times 10^7\,m\,s^{-1}$ is fired across the space between parallel plates as shown in Figure 6.34.

a What is the magnitude and direction of the electric field between the plates?
b What is the magnitude and direction of the electric force on the electrons?
c What is the acceleration of the electrons?
d How long does it take the electrons to move across the space between the plates?
e What is the vertical deflection of the electron beam as it moves across the plates? (Mass of an electron = 9.1×10^{-31} kg; charge on an electron = 1.6×10^{-19} C)

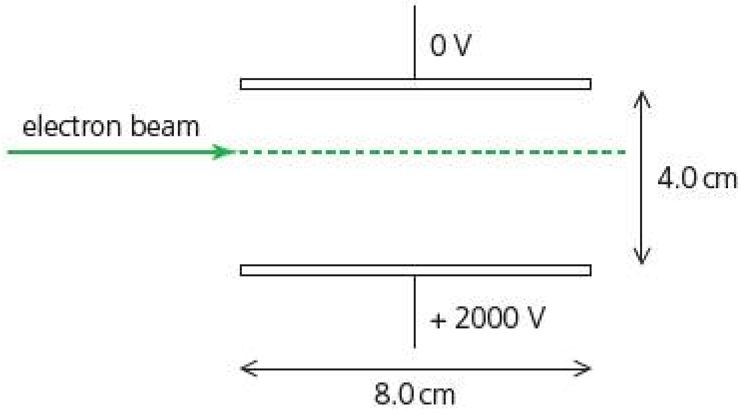


Figure 6.34

a $E = \frac{V}{d}$

$E = \frac{2000}{0.040}$

$E = 5.0 \times 10^4\,\text{V m}^{-1}$ vertically upwards

b $E = \frac{F}{q}$ so $F = Eq$

$F = (5.0 \times 10^4) \times (1.6 \times 10^{-19})$

$F = 8.0 \times 10^{-15}\,\text{N}$ downwards

c $F = ma$ so $a = \frac{F}{m}$

$a = \frac{8.0 \times 10^{-15}}{9.1 \times 10^{-31}}$

$a = 8.8 \times 10^{15}\,\text{m s}^{-2}$ downwards

d $v = \frac{s}{t}$ (The component of velocity in the horizontal direction is constant.)

$t = \frac{s}{v}$

$t = \frac{0.080}{3.7 \times 10^7}$

$t = 2.2 \times 10^{-9}\,\text{s}$

e The vertical and horizontal motions of the electrons can be considered separately. For the vertical motion:

$s = ut + \frac{1}{2}at^2$

$s = 0 + \frac{1}{2}[(8.8 \times 10^{15}) \times (2.2 \times 10^{-9})^2]$

$s = 0.021\,\text{m}$

6.2.8 **Solve** problems involving electric charges, forces and fields.

28 **a** Calculate the electric field strength between the plates shown in Figure 6.35 if the potential difference between the plates is 800 V.

b Sketch a graph of how the field strength would vary with separation of the plates as it was slowly reduced to 4 cm.

c Calculate the force on the small sphere *Q*, which has a charge of -4.8×10^{-9} C, when the separation of the plates is as shown in the diagram.

d If the sphere has a mass of 0.12 g, what acceleration will the electric field give it?

e In which direction would it accelerate under the effect of the electric force alone?

f Assuming that the sphere also falls freely under the effect of gravity, sketch the motion of the sphere.

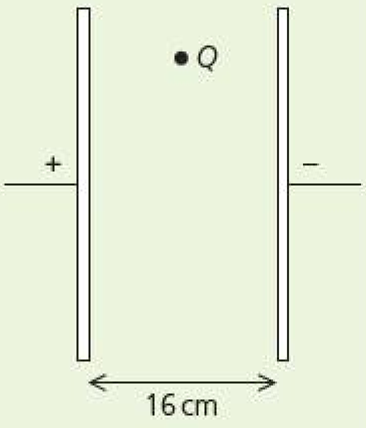


Figure 6.35

29 An electric field of about $3 \times 10^6\,\text{V m}^{-1}$ is needed between two contacts in dry air before as spark can occur.

a What potential difference is needed to produce a spark if the gap between the contacts is 5 mm?

b Suggest why the voltage needed will be less if the air is not dry.

c Why are electrostatic effects in everyday life much more common when the humidity is low?

6.3 Magnetic force and field

Gravitational fields exist around masses and electric fields exist around charges, but what causes a magnetic field? The answer 'magnets' may be true, but the kind of **'permanent' magnets** with which we are all familiar (for example, that hold notes onto a refrigerator door) are only one small and fairly unimportant example of magnetic effects. *All* electric currents produce magnetic fields and for this reason, it is difficult to separate the topics of electricity and magnetism. **Electromagnetic effects** dominate our lives and are essential for the operation of such things as power stations, trains, cars, planes, televisions, hairdryers and computers.

■ **Additional Perspectives**

The Earth's magnetic field

Magnetic materials have been known for thousands of years, long before they were actually understood. It was recognized by early civilizations that a piece of magnetic material that is free to move will always twist until it is pointing approximately north–south. The end which points to the north was called the north (-seeking) pole and the other end was called the south (-seeking) pole. Magnets used in this way are called **compasses** and, for many centuries, they have been a great help to people for finding their way around.

The simplest bar magnets have one pole at each end and are called **dipole** magnets. It is not possible to have a magnet with only one pole and if a dipole magnet is cut in half the result will be two smaller and weaker dipole magnets. (Magnets can be designed to have complex shapes and multiple poles.) When dipole magnets are brought close to each other, it quickly becomes obvious that opposite poles attract each other and similar poles repel each other. If at least one of a pair of such magnets is free to move, they will align with each other. This then helps us explain the action of a compass: the Earth itself behaves like a very large bar magnet and the small, freely moving magnet of the compass is just twisting to line up with the Earth's magnetic field.

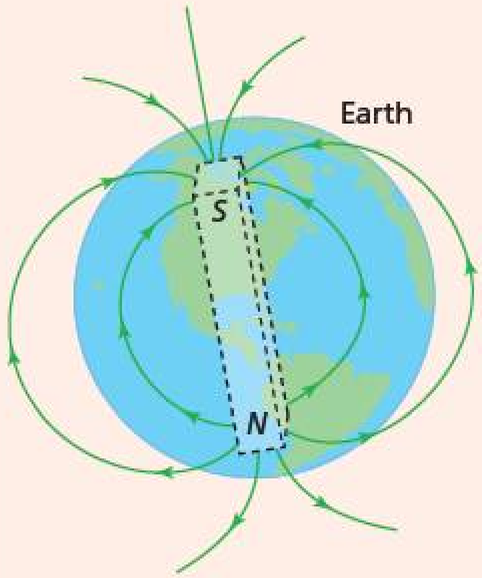


Figure 6.36 Earth's magnetic field

The terms *north pole* and *south pole* are still used for describing the ends of magnets, although this can very easily cause confusion with the Earth's geographic poles. By definition, the end we call the north pole of a magnet is attracted to the geographic north of the Earth. This means that the Earth's magnetic field has a *magnetic* south pole at the geographic north and a *magnetic* north pole at the geographic south (see Figure 6.36).

The Earth's geographic poles are located where the axis of rotation reaches the surface, but the poles of the Earth's magnetic field are not in the same place, nor on the surface, and they slowly rotate. At the moment the geographic and magnetic poles at the north of the Earth are about 800 km apart. If this is not understood by a traveller, it could cause some problems with navigation over large distances using a magnetic compass, especially when close to the poles.

The Earth's magnetic field is caused by electric currents in its liquid outer core and is believed to reverse its polarity (north becomes south and south becomes north) on average about every 300 000 years, for reasons that are still not fully understood. The last such reversal was about 800 000 years ago.

Questions

1 The Earth's magnetic field extends a long way above the Earth's surface. Research into the Earth's magnetosphere and how it protects the Earth from the 'solar wind'.

2 It is often reported that some birds, fish, sea mammals and even bacteria can navigate by their natural ability to detect (in some way) the Earth's magnetic field. Use the Internet to find out if this is scientific fact, theory or myth.

Magnetic fields

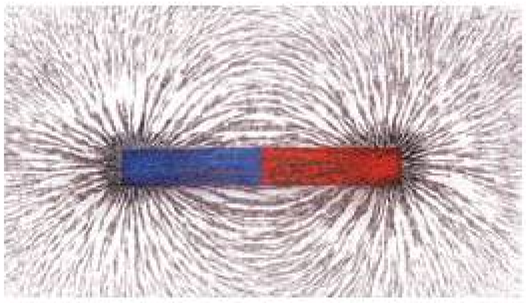


Figure 6.37 Iron filings used to demonstrate the magnetic field around a bar magnet

6.3.1 **State** that moving charges give rise to magnetic fields.

We have seen that a charge (which may be moving or stationary) produces an *electric field* around it. **Magnetic fields** are produced around *moving* charges. Since electric currents are a flow of moving charge, *all* electric currents have magnetic fields around them.

The field around a permanent magnet (see Figure 6.37) has its origin in the movement of electrons in certain atoms (like iron, which is described as *ferrous*). In most non-ferrous materials these small scale magnetic effects cancel each other out.

Magnetic field patterns

6.3.2 Draw magnetic field patterns due to currents.

Magnetic fields can be represented on paper or screen using **magnetic field lines** (in a similar way to gravitational and electric field lines). By convention the direction of a magnetic field is the same as the direction in which a compass points: from *magnetic* north to *magnetic* south.

Field due to a current in a straight wire

Figure 6.38a shows the field produced by a steady current flowing in a long straight wire. In the diagram the current is flowing perpendicularly into the page. The field is circular around the wire and gets weaker as the distance from the wire increases, shown by the increasing distance between the field lines. (The field strength is inversely proportional to the distance from the wire.) The direction of the field lines is shown and can be predicted using the 'right-hand grip rule': if the thumb points in the conventional direction of current flow, then the fingers indicate the direction of the magnetic field.

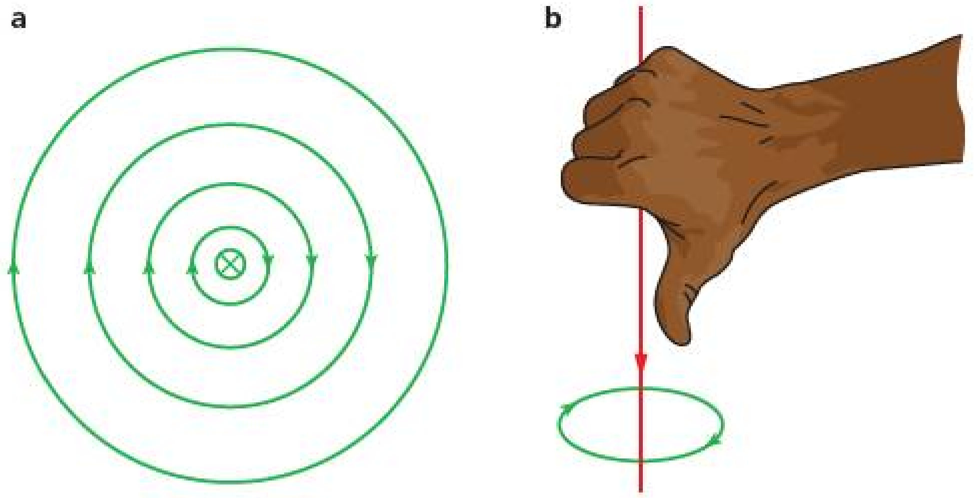


Figure 6.38 a The magnetic field due to a current flowing in a long straight wire into the plane of the paper; **b** using the right-hand grip rule to predict the direction of the field

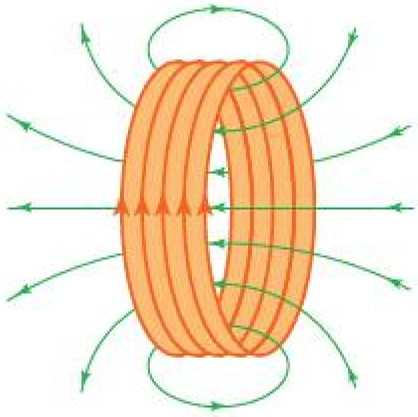


Figure 6.39 Magnetic field due to the current in a flat circular coil

Field due to a current in a flat circular coil

To increase the strength of the magnetic field around a current in a wire the obvious thing to do is to increase the current. When this is done, the field increases proportionally to the current. However, if the current is too high, the wire will get hot and melt. Winding the wire into a **coil** with as many turns as possible produces a strong magnetic field when a reasonably sized current is passed through it, see Figure 6.39.

Different materials have different magnetic properties. Magnetic **permeability** is a measure of a medium's ability to transfer a magnetic field and it can be compared to electric **permittivity** for electric fields. The magnetic permeability of free space is given the symbol μ_0 and its value is given in the IB *Physics data booklet*.

Other materials have greater permeabilities. If a coil of wire is wound around a material with high permeability, the field produced can be very much stronger than it would be without the core. In this way very strong, but adjustable, electromagnets can be made. The magnetic material used as the core must gain and lose its magnetic properties quickly and, for this reason, 'soft' iron is used as the core of most electromagnetic devices, such as transformers and motors. Figure 6.40 shows an electromagnet in use.

Figure 6.40 Electromagnet

Field due to a current in a solenoid

When a coil of insulated wire is wrapped regularly so that the turns do not overlap and it is significantly longer than it is wide, it is called a **solenoid**. Solenoids are useful for the strong magnetic fields produced inside them. Figure 6.41 represents a solenoid showing the parallel lines of a uniform magnetic field inside. The overall field pattern is similar in shape to that produced by a simple permanent bar magnet. One end of the solenoid acts like a north pole and the other like a south pole. Reversing the direction of the current changes the south pole to a north pole, and the north pole to a south pole. This is called reversing the **polarity** of the magnetic field. The more turns of wire in a given length, the stronger the magnetic field, for the same current.

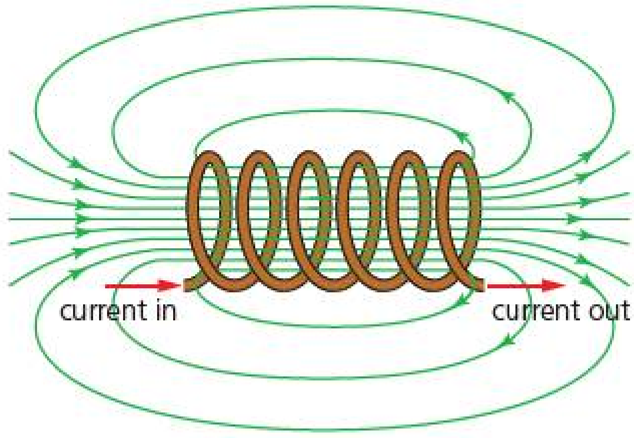


Figure 6.41 Magnetic field due to the current in a solenoid

Forces on currents moving across magnetic fields

When a current is passing through a low-mass conductor in a magnetic field, as in Figure 6.42, the conductor may be seen to move. In this example, the aluminium starts to move upwards. (A strong uniform magnetic field can be produced in the space between the opposite poles of a u-shaped permanent magnet.)

In order to describe and explain this, and many other electromagnetic effects, we need to represent and understand these situations in three dimensions. Figure 6.43a shows a situation similar to that in Figure 6.42 – a wire carrying an electric current across a magnetic field. The current is perpendicular to the magnetic field from the permanent magnets. In Figure 6.43b the same situation is drawn in two dimensional cross-section, with the wire represented by the point P and the magnetic fields from the magnets (shown in green) and from the current (shown in blue) included.

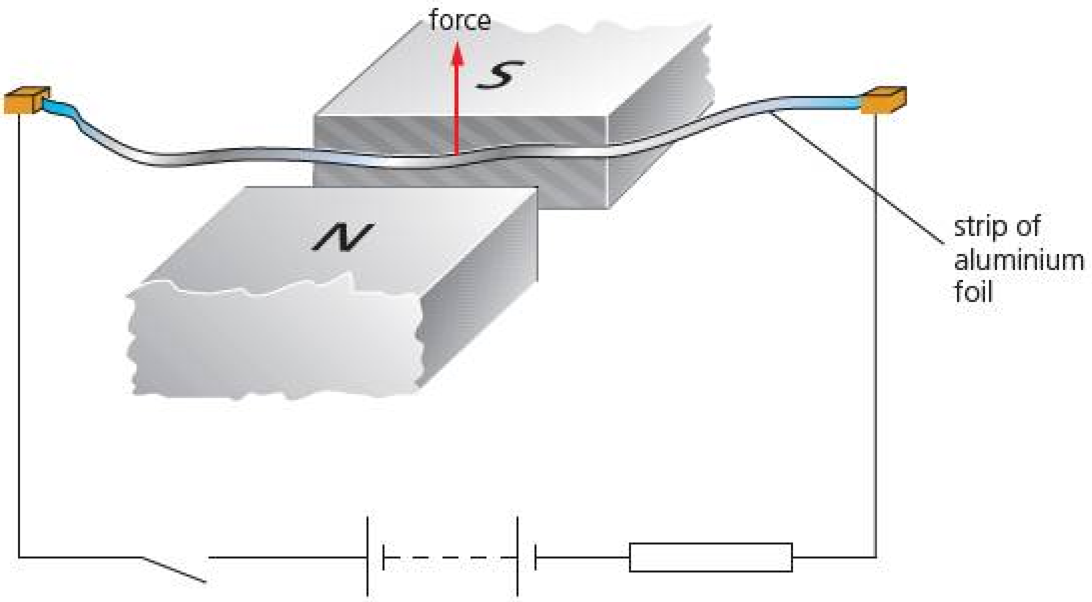


Figure 6.42 Demonstrating the force on a current flowing in a piece of aluminium

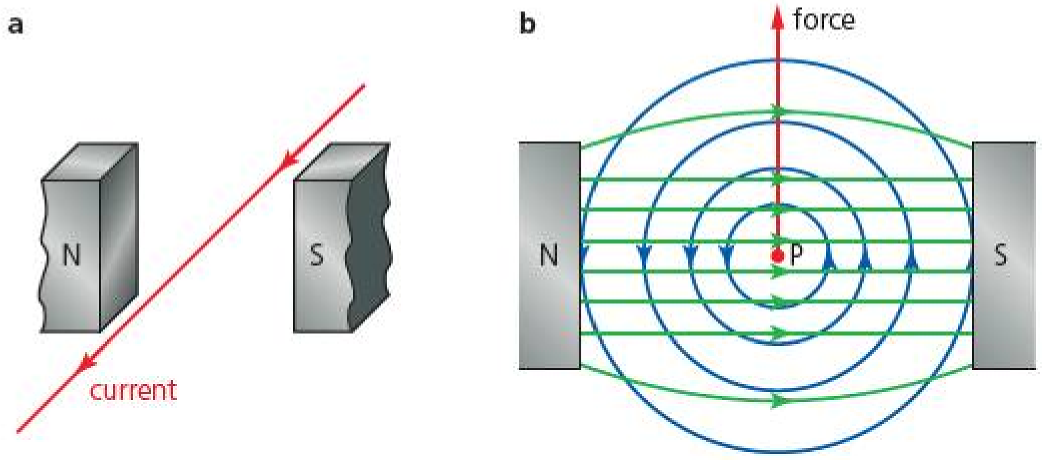


Figure 6.43 Representing the current, the field and the force in three dimensions

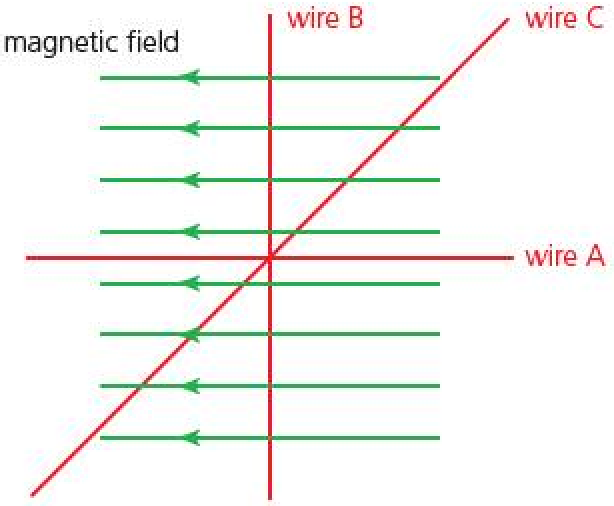


Figure 6.44 How force varies with the angle of the current to the magnetic field: there will be no force on wire A and the greatest force will be on wire B. Wire C will experience a force, but the force per unit length of wire will be less than for wire B.

The two fields are in the same plane, so it is easy to consider the combined field that they produce. Above the wire the fields act in opposite directions and they combine to produce a weaker field. Below the wire the fields combine to give a stronger field. This difference in magnetic field strength on either side of the wire produces an upwards force on the wire, which can make the wire move (if it is not fixed in position). This important and useful phenomenon is often called the **motor effect** and it provides the basic principle behind the operation of electric motors.

In order for there to be force on a current-carrying conductor in a magnetic field, the current must be flowing *across* the field. If the current is flowing in the same direction as the magnetic field lines from the permanent magnets, there will be no interaction between the fields and, therefore, no force produced. Figure 6.43 shows the situation in which the force is maximized because the field and current are perpendicular, but as the angle between field and current reduces, so too does the force. Consider Figure 6.44, which shows three different wires carrying the same current in different directions in the same uniform magnetic field.

Direction of the force on a current-carrying conductor in a magnetic field

6.3.3 **Determine** the direction of the force on a current-carrying conductor in a magnetic field.

The direction of the force is always perpendicular to both the direction of current and the direction of the magnetic field. In Figure 6.44, the forces on B and C will be in or out of the plane of the paper, depending on the direction of the current. We can use the 'left-hand rule' to predict the direction of force if the current and field are perpendicular to each other. This is shown in Figure 6.45. Remember that magnetic fields always point from magnetic north to magnetic south and the (conventional) direction of current is always from positive to negative.

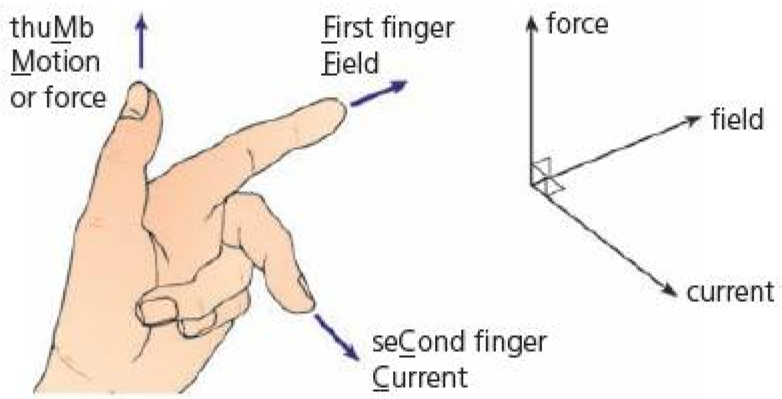


Figure 6.45 Fleming's left-hand rule predicts the direction of the force

Magnetic field strength

6.3.5 **Define** the *magnitude* and the *direction* of a magnetic field.

We know that gravitational field strength, g, equals gravitational force divided by mass, and electric field strength, E, equals electric force divided by charge, so it would be consistent to suggest that magnetic field strength equals magnetic force divided by moving charge (current). However, the size of the magnetic force depends not only on the current, I, but also on the length of the conductor in the field, L, and its direction relative to the field, θ (see Figure 6.46).

We define **magnetic field strength** (given the symbol B) as follows:

$$B = \frac{F}{IL\sin\theta}$$

Magnetic field strength is also commonly called **magnetic flux density**. The direction of the field is in the plane perpendicular to the plane containing the current and the force. The units for magnetic field strength are newtons per amp metre, $N\,A^{-1}\,m^{-1}$. Unlike gravitational and electric

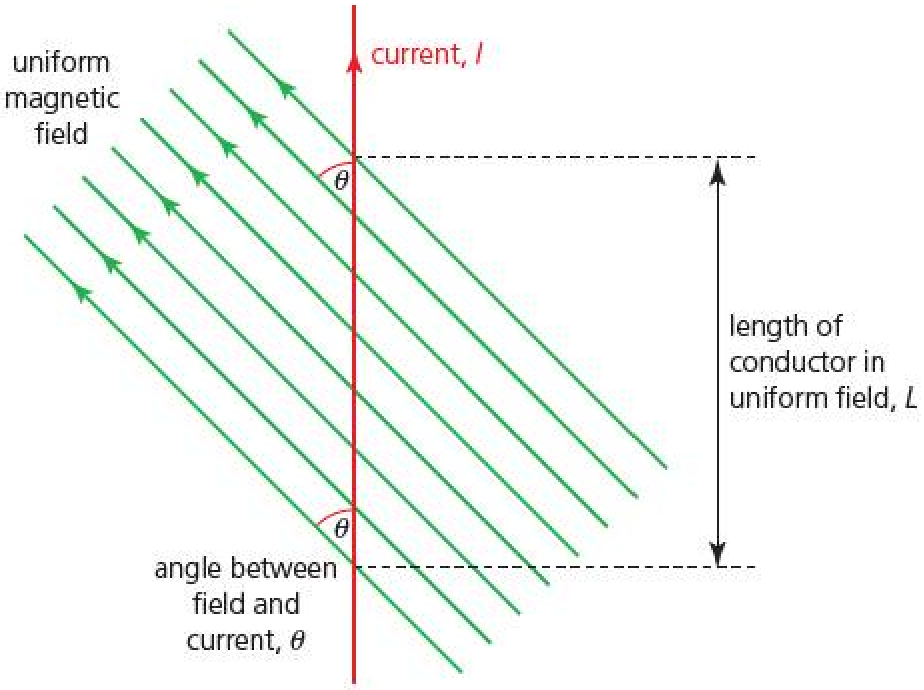


Figure 6.46 Current I flowing at an angle θ across a magnetic field

fields, these units for magnetic field strength are also given another name: $1\,\mathrm{N\,A^{-1}\,m^{-1}}$ is called a **tesla**, or **1 T**. (This unit is named after the eccentric physicist Nikola Tesla, who was born in Croatia with Serbian parents. He did a lot of important work on electromagnetism in Europe and, later, in the United States.)

The equation on page 215 is more usually written in the form:

$$F = BIL\sin\theta$$

This equation, which is given in the IB *Physics data booklet*, shows that the force on a current-carrying conductor in a magnetic field equals the size of the current, multiplied by the length of the conductor, multiplied by the component of the field perpendicular to the current.

Worked example

9 In Figure 6.47, a measured current is flowing across a small, uniform magnetic field.

a In which direction is the force on the wire?

b In which direction is the force on the balance?

c When the current is flowing, the balance indicates that there is an extra mass of 4.20×10^{-2} g on the balance. What extra downwards force is this?

d If the current is 1.64 A and the length of the field 8.13 cm, what is the strength of the magnetic field?

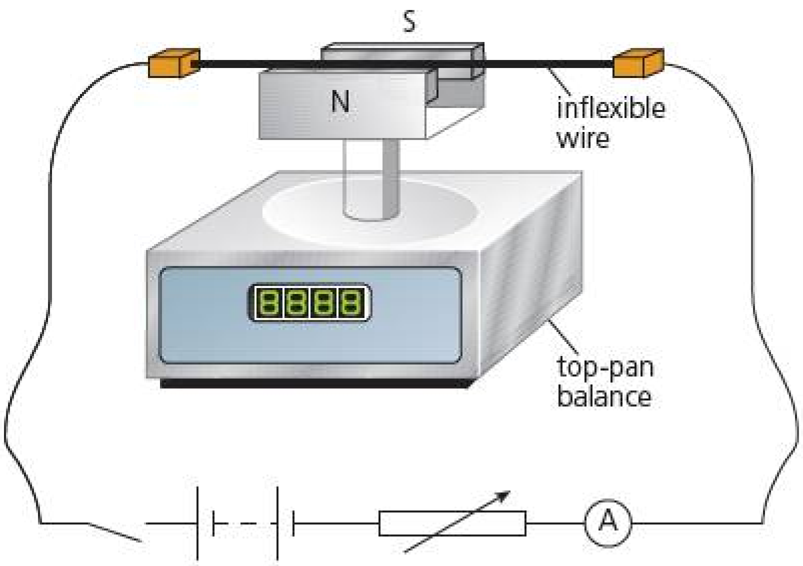


Figure 6.47 Investigating magnetic field strength

a Using the left-hand rule, the force is upwards.

b Using Newton's third law, the force is downwards.

c $W = mg = (4.2 \times 10^{-2} \times 10^{-3}) \times 9.81$

$W = 4.12 \times 10^{-4}\,\mathrm{N}$

d Using $F = BIL\sin\theta$, with $\sin\theta = 1$, gives

$4.12 \times 10^{-4} = B \times 1.64 \times 0.0813$

$B = 3.09 \times 10^{-3}\,\mathrm{T}$

6.3.6 Solve problems involving magnetic forces, fields and currents.

30 At point P in Figure 6.48, a large current is flowing vertically upwards, out of the plane of the paper.

a In what direction is the magnetic field (due to this current) at points Q and R?

b How does the strength of the field at Q compare to the strength at R?

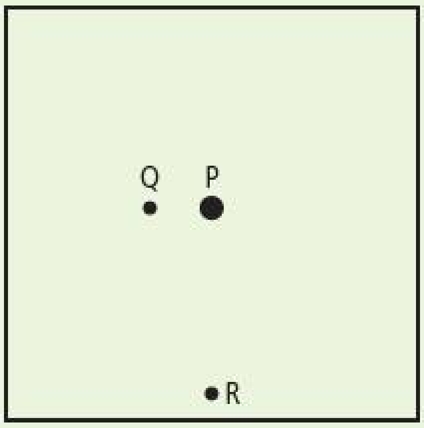


Figure 6.48

31 Figure 6.49 shows a compass placed near to a solenoid which is carrying a small electric current I.
- **a** Make a copy of the diagram showing where the compass might point if the current in the solenoid is doubled.
- **b** Where would the same compass point if the direction of the current was reversed?
- **c** Draw the compass in a position where it could point to the geographic south.

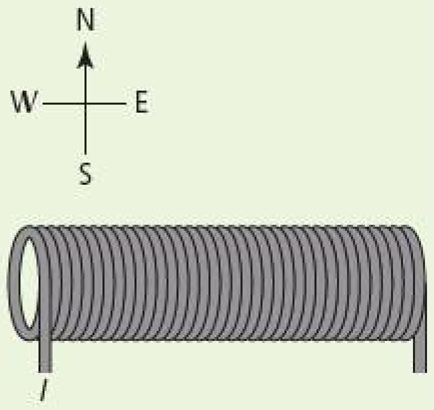


Figure 6.49

32 Calculate the magnetic force per metre on a wire carrying a current of 1.2 A through a magnetic field of 7.2 mT if the angle between the wire and the field is:
a 30° **b** 60° **c** 90° **d** 0°.

33 **a** The Earth's magnetic field strength at a particular location has a horizontal component of 24 μT. What is the maximum force per metre that a horizontal cable carrying a current of 100 A could experience?
b In which direction would the current need to be flowing for this force to be vertically upwards?

34 A current is flowing in a horizontal wire perpendicularly across a magnetic field of strength 0.36 T. It experiences a force of 0.18 N, also horizontally.
a Draw a diagram to show the relative directions of the force, field and current.
b If the length of wire in the field is 16 cm, calculate the magnitude of the current.

35 **a** A current of 3.8 A in a long wire experiences a force of 5.7×10^{-3} N when it flows through a small magnetic field of strength 25 mT. If the length of wire in the field is 10 cm, what is the angle between the field and the current?
b If the direction of the wire is changed so that it is perpendicular to the field, what would be the new force on the current?

36 Use the Internet to learn about the suspension system used in Maglev trains.

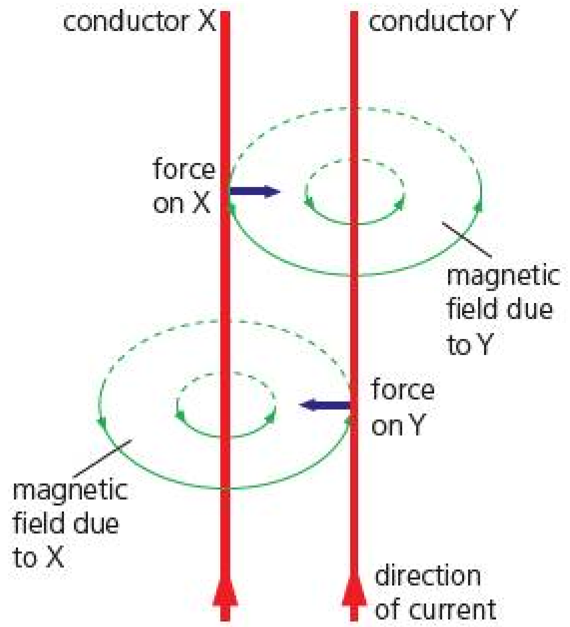


Figure 6.50 Forces between parallel currents

Forces between currents in parallel wires

Consider the two parallel wires shown in Figure 6.50. If both wires are carrying a current, then each wire is in the magnetic field of the other wire. Both wires will experience a force and, using the left-hand rule, the forces will be attractive between the wires if the currents are in the same direction. The forces are equal and opposite (remember Newton's third law). If the currents are in opposite directions the wires will repel.

This arrangement is used to define the SI unit of current, the ampere, as mentioned in Chapter 5. (One ampere, 1 A, is the current flowing in two infinitely long straight parallel wires which produces a force of $2 \times 10^{-7}\,\mathrm{N\,m^{-1}}$ between the wires if they are 1 m apart in a vacuum.)

Forces on individual charges moving across magnetic fields

It should be clear that the forces we've just discussed are not acting on the wires themselves, but on the currents in the wires. In fact, the force on a current is just the sum of the forces on the individual moving charges. Any moving charge crossing a magnetic field will experience a force, whether it is in a wire or moving freely through a vacuum.

Consider a charge, q, moving across a magnetic field, B, at an angle, θ, and with a velocity, v, as shown in Figure 6.51. In time, t, q moves from X to Y, a distance L.

We have seen that $F = BIL \sin\theta$. But in this case, $v = L/t$ and $I = q/t$, so the equation can be written as:

$$F = B(q/t)(vt) \sin\theta$$

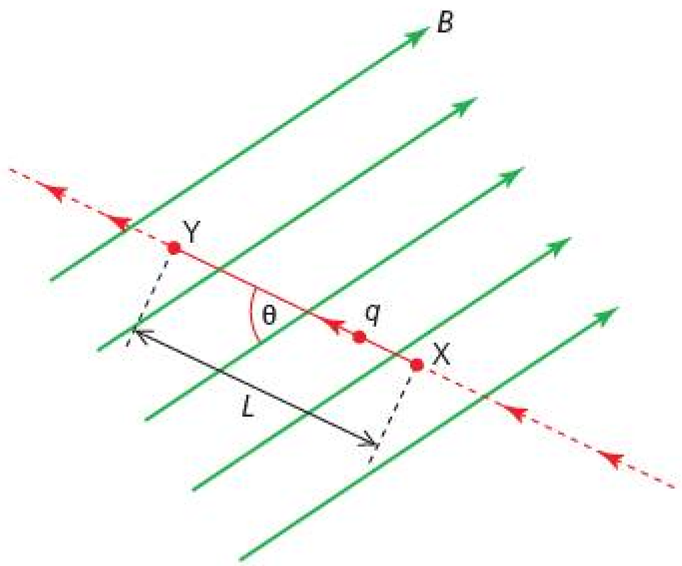


Figure 6.51 An individual charge moving across a magnetic field

Cancelling t, we get:

$$F = qvB\sin\theta$$

This equation is given in the IB *Physics data booklet*.

With this equation we can calculate the magnetic force acting on any charged particle moving across a uniform magnetic field. Note the important and interesting fact that the force is greater if the particle moves faster. There is no force on a charge that is not moving ($v = 0$). It is not normally possible to track the progress of individual charged particles, but this equation is very useful when investigating the properties of charged particles which can be formed into particle **beams**. This includes beams of electrons, protons, ions, and certain kinds of nuclear radiations (alpha and beta particles). The deflection of particle beams as they pass through magnetic fields (and/or electric fields) is a very useful way of determining the charge and/or mass of the individual particles, as well as their speeds and energies.

Direction of magnetic forces on charged particles moving across magnetic fields

6.3.4 **Determine** the direction of the force on a charge moving in a magnetic field.

Any charged particle moving perpendicularly across a magnetic field will experience a force that is always perpendicular to its instantaneous velocity. The direction of the force can be predicted by the left-hand rule, remembering that the conventional direction for current is that of the movement of positive charges. (This means that the direction of current for moving electrons will be opposite to their velocity.)

A force perpendicular to motion is the necessary condition for circular motion and the particle will move along the arc of a circle.

Figure 6.52 shows examples of charges moving perpendicularly across strong magnetic fields. In each example, the speeds and kinetic energies of the particles do not change as a result of the magnetic force. Note how the direction of a magnetic field perpendicular to the paper can be represented:

- a ⊙ shows that the field is pointing out of the paper
- a × shows that the field is pointing into the paper.

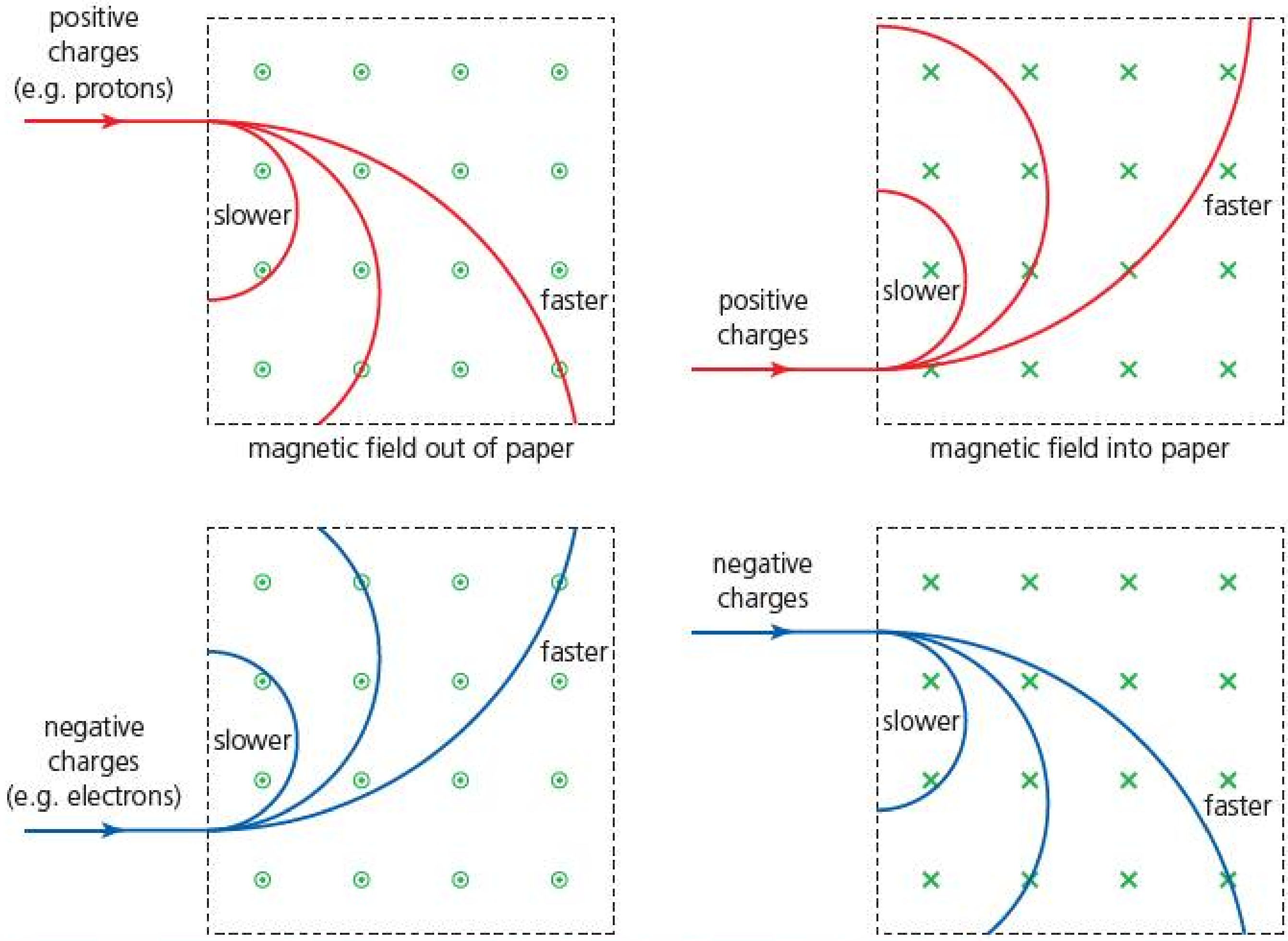


Figure 6.52 Circular paths of charges moving perpendicular to magnetic fields

Figure 6.52 also shows that the faster a charge is moving, the greater is the radius of its circular path, even though it is experiencing a greater magnetic centripetal force. This can be explained by equating the equation for magnetic force to the equation for centripetal force for a charge of mass, m, moving in a circular path of radius r (see Chapter 2).

$$F = qvB\sin\theta = \frac{mv^2}{r}$$

Remembering that, for a movement perpendicular to the field, $\sin\theta = 1$, we get

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

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

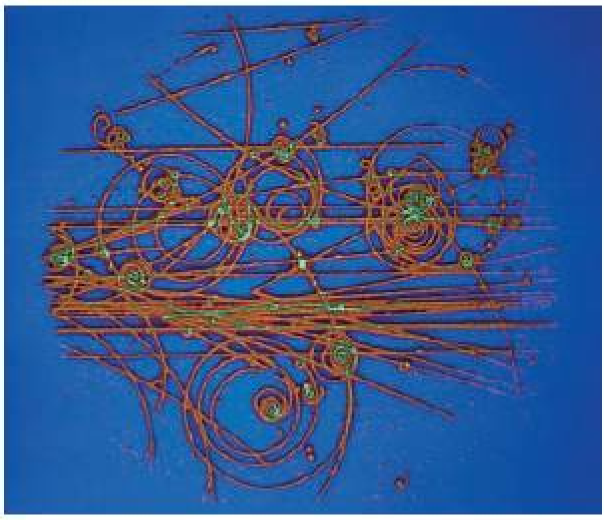


Figure 6.53 Curved paths of individual particles in a nuclear physics bubble chamber

This shows us that for a given mass, charge and magnetic field, the radius is proportional to the velocity of the charge. A charge that is losing kinetic energy and velocity (because of collisions with other particles) will spiral inwards as its radius decreases.

We can also see that if the velocity and charge of particles are kept constant, then their radius in a given magnetic field will depend only on their mass. Information about the mass and speed of sub-atomic particles can be determined by analysing their paths as they move across known magnetic fields, see Figure 6.53.

If the motion of the charged particle is not perpendicular or parallel to the field, then it will move in a spiral-like path (helix).

Worked examples

10 What is the magnetic force acting on a proton (charge = $+1.6 \times 10^{-19}$ C) moving at an angle of 32° across a magnetic field of 5.3×10^{-3} T at a speed of $3.4 \times 10^5\,\text{m s}^{-1}$?

$F = qvB\sin\theta$

$F = (1.6 \times 10^{-19}) \times (3.4 \times 10^5) \times (5.3 \times 10^{-3}) \sin 32°$

$F = 1.5 \times 10^{-16}\,\text{N}$

11 An electron of mass 9.1×10^{-31} kg and charge -1.6×10^{-19} C is moving at a speed of $1.6 \times 10^7\,\text{m s}^{-1}$ perpendicularly to a magnetic field of 1.4×10^{-4} T. Calculate the radius of its path.

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

$r = \frac{(9.1 \times 10^{-31}) \times (1.6 \times 10^7)}{(1.6 \times 10^{-19}) \times (1.4 \times 10^{-4})}$

$r = 0.65\,\text{m}$

6.3.6 Solve problems involving magnetic forces, fields and currents.

37 a Explain how it is possible for a charged particle to move through a magnetic field without experiencing a force.

b Explain whether it is possible for the same particle to move through electric and gravitational fields without experiencing forces.

38 An alpha particle has a charge of $+3.2 \times 10^{-19}$ C and a mass of 6.7×10^{-27} kg. It moves across a magnetic field of strength 0.28 T with a speed of $1.4 \times 10^7\,\text{m s}^{-1}$ at an angle of 33°.

a What is the force on the particle?

b Calculate its centripetal acceleration.

c Describe the path of the alpha particle across the field.

39 **a** What magnetic field strength is needed to provide a force of 1.0×10^{-12} N on each particle in a beam of singly charged ions ($q = 1.6 \times 10^{-19}$ C) moving perpendicularly across the field with a speed of 5.0×10^{6} m s^{-1}?

b The ions move in the arcs of circular paths of radius 2.78 m. What is their mass?

c If the beam was replaced with doubly charged ions of the same mass, but moving with half the speed, what would be the radius of their path?

40 **a** Electrons are accelerated into a beam by a voltage of 7450 V. What is the kinetic energy of the electrons in: **i** electronvolts, **ii** joules?

b Calculate the final speed of the electrons.

c What strength magnetic field is needed to make these electrons move in a circle of radius 14.8 cm?

d If the accelerating voltage is halved, what would the radius of the electrons' path be in the same field?

41 A charge of $+4.8 \times 10^{-19}$ C moving perpendicularly across a magnetic field of 1.9×10^{-2} T experiences a force of 9.5×10^{-14} N.

a What was the speed of the particle?

b What electric field would be needed to produce the same force on this charge?

c Draw the path of the particle moving across the fields in a direction such that these two forces could be equal and opposite to each other (so that the resultant force on the particle was zero).

42 Find out how the Aurora Borealis is formed (see Figure 6.54).

Figure 6.54 The Aurora Borealis

Comparison of gravitational, electric and magnetic forces and fields

Table 6.2 A comparison of gravitational, electric and magnetic forces

	Gravitational	Electric	Magnetic
What does the force act on?	All masses	All charges	All *moving* charges (currents)
What is direction of force?	Directly between masses or charges		Perpendicular to both the magnetic field and the motion
	Always attractive	Attractive between opposite charges; repulsive between similar charges	
How can force be calculated?	Using an inverse square law between two points		From the definition of magnetic field strength
	$F = \frac{Gm_1m_2}{r^2}$	$F = \frac{kq_1q_2}{r^2}$	$F = qvB\sin\theta$ on a single charge; $F = BIL\sin\theta$ on a current
How is field strength defined?	$g = \frac{F}{m}$	$E = \frac{F}{q}$	$B = \frac{F}{IL\sin\theta}$ $\left(= \frac{F}{qv\sin\theta}\right)$
How can the field around a point be calculated?	$g = \frac{Gm}{r^2}$	$E = \frac{kQ}{r^2}$	–
What is the path of a particle moving perpendicularly across the field?	Parabolic	Parabolic	Circular

SUMMARY OF KNOWLEDGE

6.1 Gravitational force and field

- An attractive gravitational force exists between all masses. The magnitude of the force between two isolated point masses is proportional to the product of the masses and inversely proportional to the square of the distance between them.
- This is expressed by Newton's universal law of gravitation: $F = Gm_1m_2/r^2$, where G is the universal gravitation constant.
- This law can also be used with large spherical masses, like planets, which have gravitational fields that behave as if all their masses were concentrated at their centres. r is then the separation of the centres of the two masses concerned.
- Gravitational field strength is defined as the gravitational force acting per unit (test) mass: $g = F/m$. The unit of g is N kg^{-1}, but m s^{-2} is an alternative. Gravitational field strength is a vector quantity.
- The gravitational force on a mass near to a planet is called its weight. The same force acts in the opposite direction on the planet. By equating weight (mg) to Gm_1m_2/r^2 we can show that the gravitational field strength at a distance r from the centre of a planet, $g = Gm/r^2$, where m is the mass of the planet. This equation can be adapted to show the relationship between the gravitational field strength on the surface of a planet, g, density, ρ and radius, r: $g = \frac{4}{3}G\pi\rho r$.
- Sometimes an object may be in two or more significant gravitational fields, for example between a planet and a moon. The combined field, in magnitude and direction, can be calculated using vector addition of the individual fields.

6.2 Electric force and field

- Electric forces exist between all charged particles.
- There is a close mathematical analogy between gravitational and electric forces and fields, but there are two types of charge (called positive and negative) and only one kind of mass. Like charges attract, opposite charges repel, but gravitational forces are always attractive.
- Electrons are negatively charged, protons are positively charged.
- The size of electric forces between (charged) particles is enormously larger than the gravitational forces between the same particles. But with increasingly large numbers of particles (of both types of charge) the electric forces tend to cancel out, while the gravitational forces get larger and larger.
- The law of conservation of charge states that the total charge in an isolated system is constant.
- A material through which charges can flow is called a conductor. If charge cannot flow through a material it is called an insulator. The difference between conductors and insulators is explained in terms of the number of free (delocalized) electrons that are able to move through the material.
- Electrostatic effects are also explained by the movement of electrons.
- The magnitude of the force between two isolated point charges is proportional to the product of the charges and inversely proportional to the square of the distance between them.
- This is expressed by Coulomb's law: $F = kq_1q_2/r^2$, where k is the Coulomb constant. The similarity between Coulomb's law and Newton's law should be noted.
- The permittivity of a medium is a measure of its ability to transfer an electric force and field. The permittivity of free space, ε_0, is an important fundamental constant in physics. The coulomb constant $= \dfrac{1}{4\pi\varepsilon_0}$, so that Coulomb's law can be written as: $F = \dfrac{1}{4\pi\varepsilon_0}\dfrac{q_1q_2}{r^2}$
- This law can also be used with charged spheres, which have electric fields that behave as if all their charge is concentrated at the centre. r is then the separation of the centres of the two spheres.
- Electric field strength is defined as the force per unit charge experienced by a small positive test charge: $E = F/q$. Its unit is N C^{-1} or V m^{-1}.
- The field around a point charge can be calculated from $E = kq/r^2$

- Electric field is a vector quantity, so its direction should usually be given. Vector addition is needed to find the total electric field due to more than one point charge.
- Electric field lines show the direction of force on a positive charge. Radial field patterns around isolated point charges and charged spheres, as well as simple combinations of charges, can be drawn and interpreted.
- Charged parallel plates produce uniform electric fields in the space between them, although the field strength reduces to half at the edges. (The electric field strength between parallel plates can be calculated from $E = V/d$; the unit is $\mathrm{V\,m^{-1}}$.)

6.3 Magnetic force and field

- Magnetic fields are produced when charges move. A flow of charge is a current, so all currents are surrounded by magnetic fields.
- Magnetic fields are represented on paper by magnetic field lines, which point from magnetic north to magnetic south (in the same direction as a compass).
- The magnetic field patterns around currents in long straight wires, coils and solenoids are particularly important.
- The strength of a magnetic field in a coil or solenoid can be increased by using a larger current, more turns (in a given length), or a core of high magnetic permeability (like soft iron).
- Permeability is a measure of a medium's ability to transfer a magnetic field. The permeability of free space is an important fundamental constant, which is given the symbol μ_0.
- When a current flows across a magnetic field, the field due to the current and the external field combine to produce a force on the current. The force is perpendicular to the plane that contains the field and the current. The 'left-hand rule' can be used to predict its direction.
- The size of the force depends on the angle between the current and the field. It can be predicted from $F = BIL\sin\theta$
- B is the magnetic field strength (unit: tesla, T), which is defined by the same equation: $B = \dfrac{F}{IL\sin\theta}$
- The equation for force on a current in a magnetic field can be rewritten in terms of the movement of the individual charges: $F = qvB\sin\theta$. This equation and the left-hand rule can be used with individual charges in a particle beam moving across a vacuum.
- The force acting perpendicularly to the motion of a charge moving across a magnetic field will make it move in circular path. By equating the equation for the magnetic force to the equation for the centripetal force, we can derive an expression for the radius of such a path: $r = \dfrac{mv}{qB}$.

Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 Which of the following correctly describes the constant, G?

A It is a vector quantity.
B It has a larger magnitude on the Moon than the Earth.
C It has a smaller magnitude on the Moon than the Earth.
D It is a fundamental constant.

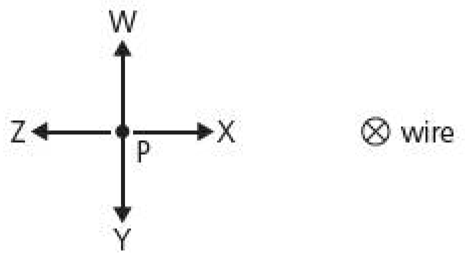
Q2 Which of the following correctly describes the process of conduction?

A There are no electrons in insulating materials.
B A current in a solid conductor is carried by positive charges moving towards the negative terminal of the battery.
C Metals conduct well because they have plenty of free electrons.
D A current in a solid conductor consists of electrons moving in the direction of the electric field.

Q3 There is a place between the Earth and the Moon where the gravitational field strength is zero. This is because:

A The Earth's gravitational field is stronger than the Moon's gravitational field.
B Gravitational fields only exist close to the surfaces of planets and moons.
C The fields from the Earth and the Moon act in opposite directions.
D The Moon does not have a gravitational field.

Q4 In the diagram, a long current-carrying wire is normal to the plane of the paper. The current in the wire ⊗ is directed into the plane of the paper.



Which of the arrows gives the direction of the magnetic field at point P?

A W **B** X **C** Y **D** Z

Standard Level Paper 1, May 09 TZ2, Q21

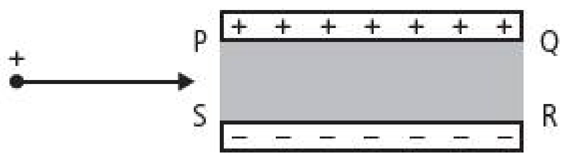
Q5 There will be a magnetic force acting on a charged particle in a magnetic field when

A the charge moves in the same direction as the field.
B the charge moves perpendicularly to the field.
C the charge moves in the opposite direction to the field.
D the charge does not move with respect to the field.

Q6 A falling mass, m, accelerates towards a planet that has no atmosphere with an acceleration a. Which of the following is the gravitational field strength of the planet?

A ma **B** $\frac{a}{m}$ **C** a **D** $\frac{m}{a}$

Q7 A positively charged particles enters the space between two charged conducting plates, with a constant velocity directed parallel to the plates, as shown.



The top plate is positively charged and the bottom plate is negatively charged. There is a magnetic field in the shaded region PQRS. The particle continues to move in a horizontal straight line between the plates. Which of the following correctly describes the magnetic field direction?

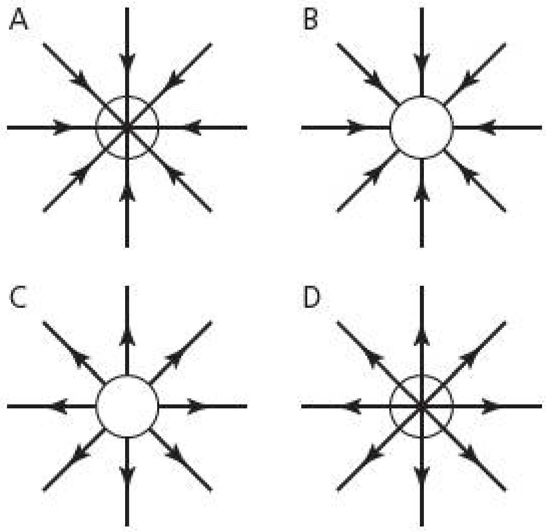
A Into the plane of paper
B Out of the plane of paper
C Up
D Down

Standard Level Paper 1, Nov 09, Q21

Q8 When a charged particle of mass m moves across a magnetic field with a speed v it experiences a magnetic force F. What force will be experienced by a particle of twice the mass but having the same charge, if it is moving across the same field in the same direction with a speed $2v$?

A F
B $\frac{F}{2}$
C $2F$
D $4F$

Q9 Which of the following four diagrams is the best representation of the electric field in and around a conducting sphere which has been charged positively?



Q10 The diagram shows two charges of opposite signs. The charges have magnitude Q and $2Q$.

At which point is the direction of the resultant electric field to the left?

A A
B B
C C
D D

Q11 Newton's universal law of gravitation concerns the force between two masses. These masses are

A planets.
B stars.
C spherical masses.
D point masses.

Paper 2 IB questions and IB style questions

Q1 This question is about gravitational fields.

a Define *gravitational field strength*. [2]

b The gravitational field strength at the surface of Jupiter is $25\,\text{N}\,\text{kg}^{-1}$ and the radius of Jupiter is $7.1 \times 10^7\,\text{m}$.

i Derive an expression for the gravitational field strength at the surface of a planet in terms of its mass M, its radius R and the gravitational constant G. [2]

ii Use your expression in **b i** above to estimate the mass of Jupiter. [2]

Standard Level Paper 2, Specimen Paper 09, QA3

Q2 This question is about gravitational and electric fields.

a The equation for the magnitude of the gravitational field strength due to a point mass may be written as below.

$$Y = \frac{KX}{s^2}$$

The equation for the magnitude of the electric field strength can also be written in the same form.

Copy and complete the table to identify the symbols used in the equation. [4]

Symbol	Gravitational field quantity	Electrical field quantity
Y		
K		
X		
s		

b The magnitude of the electrostatic force between the proton and the electron in a hydrogen atom is F_E. The magnitude of the gravitational force between them is F_G.

Determine the ratio $\frac{F_E}{F_C}$. [3]

Standard Level Paper 2, Nov 09, QA3

7 Atomic and nuclear physics

STARTING POINTS

- Matter is made from tiny particles called atoms.
- Atoms are composed of three sub-atomic particles: protons (positively charged), neutrons (neutral) and electrons (negatively charged).
- Atoms have a very small positively charged nucleus surrounded by negatively charged electrons.
- Opposite charges attract; similar charges repel.
- Coulomb's law states that the electrostatic force between two point electric charges is directly proportional to the sizes of the charges, but inversely proportional to the square of the distance between the two charges.
- The Coulomb attraction between electrons and protons allows the formation of stable atoms.
- Charged particles may be deflected by electric and magnetic fields.
- The effects of gravity may be ignored for atoms and sub-atomic particles.

Introduction

Nuclear reactions power stars, generating energy and forming the chemical elements found in nature. Nuclear physics is the study of how protons and neutrons interact to form nuclei; this helps us to understand how the heavy elements are formed in the violent explosions of stars.

Applications of nuclear physics are of great value to the global economy, from energy generation (via nuclear reactors) and defence (nuclear weapons) to industry and medicine. Thousands of lives are saved every year through the use of radioisotopes in cancer treatments and in medical imaging techniques such as positron emission tomography (PET).

7.1 The atom

Atomic structure

The nuclear model of the atom

7.1.1 **Describe** a model of the atom that features a small nucleus surrounded by electrons.

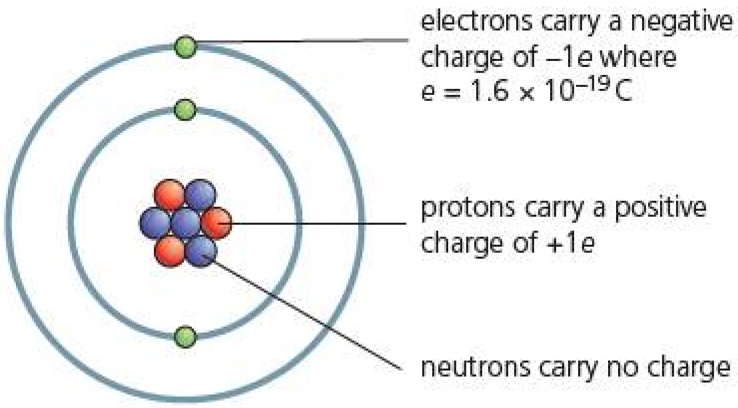


Figure 7.1 Simple nuclear model of a lithium atom

The simple **nuclear model** (Figure 7.1) was developed during the 20th century and describes the atom as consisting of a very small and dense central **nucleus** surrounded by electrons arranged into energy levels (known as 'shells' in chemistry).

The atom is composed of **sub-atomic** particles known as **protons, electrons** and **neutrons**. The protons and neutrons are collectively termed **nucleons** and contain almost all of the mass of the atom. The protons are positively charged and the neutrons are electrically neutral. The electrons are negatively charged, but have very little mass in comparison to protons and neutrons. Atoms are electrically neutral due to the presence of equal numbers of protons and electrons. The vast majority of an atom is empty space – a vacuum. The properties of protons, neutrons and electrons are summarized in Table 7.1.

Table 7.1 Properties of the sub-atomic particles

Name of particle	Approximate relative mass	Relative charge
Proton	1	+1
Neutron	1	0
Electron	$\frac{1}{1840}$	–1

One of the functions of the neutrons is to overcome the repulsion between the protons and bind them together in the nucleus. The nuclei of some atoms are unstable and release radiation in the form of small particles and/or gamma rays. This is termed ionizing radiation since it causes atoms to lose electrons and form ions. Atoms that release ionizing radiation are described as being radioactive (see Section 7.2).

In this simple model of the atom, the electrons rotate around the nucleus in circular orbits at fixed distances from the nucleus. The electrons are kept in these orbits by electrostatic attraction between the positive protons and the negative electrons. (This simple electrostatic model has been replaced by a quantum mechanical model based on the wave behaviour of electrons – see Chapter 13.)

If an atom loses one or more electrons it will form a charged particle called an **ion**. For example, if a sodium atom loses one of its electrons it becomes a positively charged sodium ion.

Na	→	Na^+	+	e^-
sodium atom		sodium ion		electron

If an atom gains an electron, it becomes a negatively charged ion. For example, a chlorine atom can gain an electron during a chemical reaction.

Cl	+	e^-	→	Cl^-
chlorine atom		electron		chloride ion

1 Calculate the total positive charge in coulombs and mass in kilograms of an iron atom which has 26 protons, 26 electrons and 30 neutrons. The values for the rest masses of the proton, electron, the neutron and the elementary charge are in the IB *Physics data booklet*.

2 The molar mass of the metallic element sodium is 23 g mol^{-1} and its density is 970 kg m^{-3}. Calculate the volume (in m^3) of a single sodium atom ($N_A = 6.02 \times 10^{23}$ mol^{-1}).

The Geiger–Marsden experiment

7.1.2 **Outline** the evidence that supports a nuclear model of the atom.

One method of finding out about atomic structure is to aim a stream of fast-moving tiny particles at much larger atoms and see how they break up or how the projectiles bounce off them.

In 1909, Ernest Rutherford and two of his research students, Geiger and Marsden, working at the University of Manchester, directed a narrow beam of **alpha particles** (helium nuclei consisting of two protons and two neutrons) from a radioactive source at a very thin piece of gold foil. A zinc sulfide detector was moved around the foil to determine the directions in which alpha particles travelled after striking the foil (Figure 7.2).

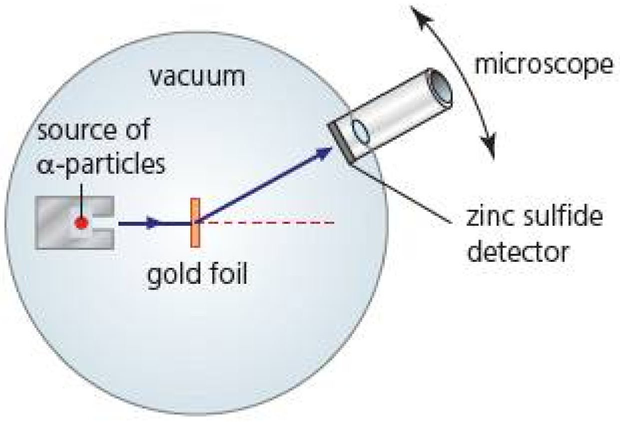


Figure 7.2 The alpha particle scattering experiment

Rutherford published his findings in 1911. He reported that:

- most of the alpha particles passed through the foil with very little or no deviation from their original path
- a small number of particles (about 1 in 1800) were deviated through an angle of more than about 10°
- an extremely small number of particles (about 1 in 10 000) were deflected through an angle greater than 90°.

From these observations, Rutherford drew the following conclusions:

- Most of the mass of an atom is concentrated in a very small volume at the centre of the atom. Most alpha articles would therefore pass through the foil undeviated (no turning) since most of the atom was empty space (a vacuum).
- The centre (or nucleus) of an atom must be positively charged in order to repel the alpha particles (helium nuclei), which are also positively charged. Alpha particles that pass close to the nucleus will experience a strong electrostatic repulsive force, causing them to deviate (turn).
- Only alpha particles that pass very close to the nucleus, almost striking it head-on, will experience electrostatic repulsion large enough to cause them to deviate through angles greater than 90°. The fact that so few particles did so confirms that the nucleus is very small, and that most of the atom is empty space (vacuum).

Figure 7.3 shows some of the possible trajectories (paths) of the alpha particles. Rutherford used the nuclear model of the atom and classical equations based on electrostatics and Coulomb's inverse square law (covered in Chapter 6) to describe the force between charged particles. He used these forces to calculate the fraction of alpha particles expected to be deviated through various angles. The calculations agreed very closely with the results from the experiment, confirming Rutherford's nuclear model of the atom.

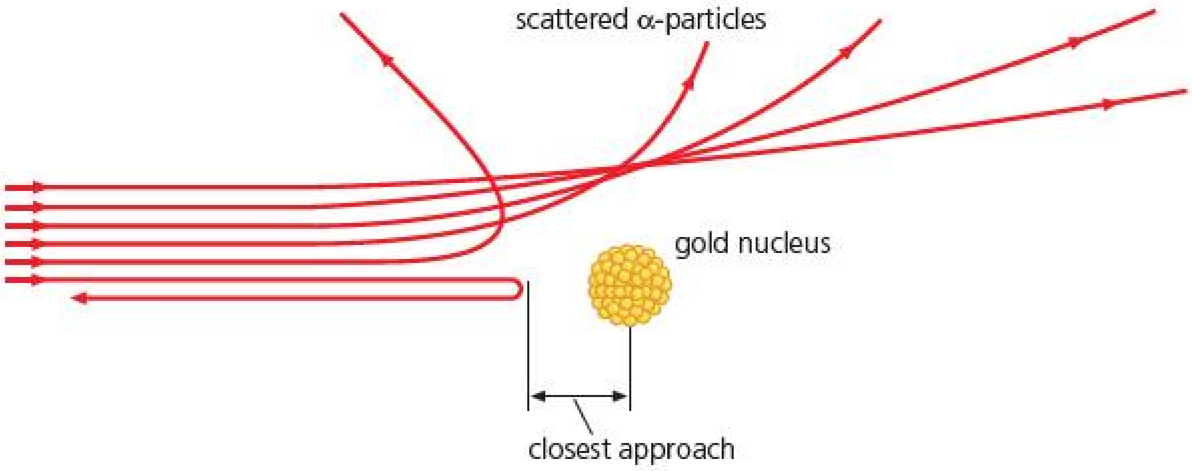


Figure 7.3 Alpha particle trajectories (all hyperbolas) from the gold foil experiment

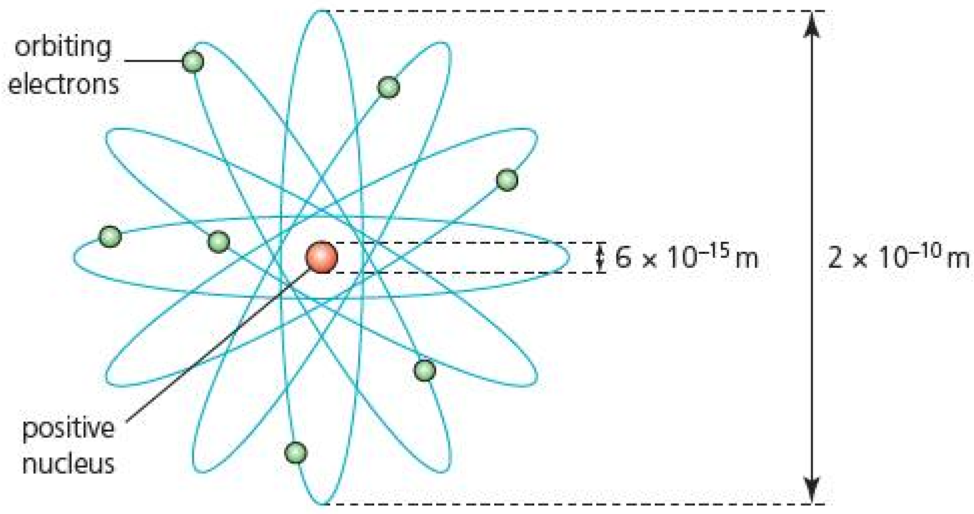


Figure 7.4 The structure and dimensions of a nitrogen atom, with seven electrons orbiting the positive nucleus. Note that the diameter of the atom is more than 30 000 times larger than that of the nucleus

From his results, Rutherford calculated that the diameter of the nucleus is in the order of 10^{-15} m, and the diameter of the whole atom is in the order of 10^{-10} m. Figure 7.4 shows the features of the nuclear model of a nitrogen atom with accurate dimensions.

Additional Perspectives

High-energy scattering

Rutherford's experiment was repeated some years later by Chadwick and Beiler using accelerated alpha particles of considerably higher energy than those emitted from radioactive sources. At these energies the Rutherford scattering formula began to break down, that is, the distribution of scattered alpha particles no longer fitted the prediction. This meant that one of the assumptions used to derive the equation was valid at low energy and not at high energy. At high energies the alpha particles get much closer to the target nucleus, suggesting that a new short-range force operates over these short distances and changes the interaction. This was the first evidence for the **strong nuclear force** (see page 235), a short-range interaction that binds neutrons and protons together in the nucleus. However, it has no long-range effects (i.e. greater than about 10^{-15} m), and does not act on electrons.

Question

1 Find out about the work of the Japanese scientist Yukawa and his description of the strong nuclear force as an exchange force.

TOK Link: Plum pudding model

The first scientist to propose a detailed model of the atom was Joseph John ('J.J.') Thomson in 1904. He proposed that an atom consists of a uniform sphere of positive charge in which negative electricity (electrons) is randomly distributed. This model was known as the 'plum pudding' model (Figure 7.5) and assumed that the mass of the atom was spread evenly over the entire atom. Thomson's model could explain the electrical neutrality of the atom, but could not be reconciled with Rutherford's scattering of alpha particles. Rutherford's nuclear model was a paradigm shift in thinking about atomic structure and Thomson's model was abandoned.

If Thomson's model was correct, the maximum deflecting force on the alpha particle as it passes through an atom would be small, since the positive and negative charges and mass are uniformly distributed in the sphere of the atom. The highly energized alpha particles would pass through mostly undeflected (Figure 7.5).

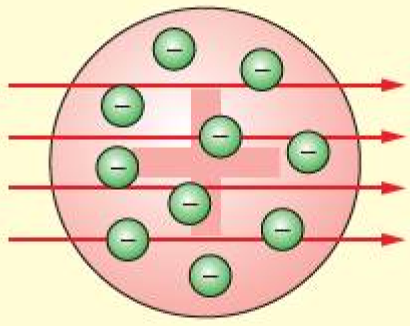


Figure 7.5 Thomson's 'plum pudding' model of atomic structure showing undeflected alpha particles

Question

1 Explain why the results of the gold foil experiment do not support Thomson's 'plum pudding' model of the atom.

3 Calculate the electric field at the surface of a nitrogen nucleus (of seven units of positive charge) and radius 6.0×10^{-15} m. Compare this with the value of the electric field of the same charge that is now spread over a sphere of radius 1.0×10^{-10} m. Use the expression $E = k\frac{q}{r^2}$ from Chapter 6.

4 An alpha particle nearly collides with a gold nucleus and returns along the same path. Sketch a graph showing how electric potential energy and kinetic energy possessed by the alpha particle vary with the distance of the alpha particle from the gold nucleus.

5 a What would have happened if neutrons had been used in Rutherford's experiment? Explain your answer.

b What would have happened if aluminium had been used instead of gold in Rutherford's experiment? Explain your answer.

c Why was a *thin* gold foil used in Rutherford's experiment?

Refining the nuclear model of the atom

7.1.3 **Outline** one limitation of the simple model of the nuclear atom.
7.1.4 **Outline** evidence for the existence of atomic energy levels.

Although Rutherford's model explained much about the behaviour of atoms, it had its limitations. Rutherford's simple nuclear model is not compatible with Maxwell's theory of electromagnetism. According to Maxwell's theory an accelerating charged particle radiates energy in the form of electromagnetic radiation. In the Rutherford model, the electrons are

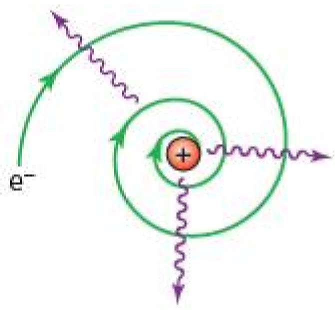


Figure 7.6 An electron that is accelerating in a simple classical model of the atom radiates energy and as it loses energy, it spirals into the nucleus

revolving around the nucleus at a constant speed. However, because they are constantly changing direction (Chapter 2), the electrons are experiencing centripetal acceleration.

Since electrons are charged, Maxwell's theory predicts that they should emit electromagnetic radiation all the time. Their speed should decrease and the electron should move closer and closer to the nucleus in a spiral path until the atom collapses (Figure 7.6). However, atoms do not emit electromagnetic radiation continuously, nor do they collapse – most atoms are stable.

In addition, the Rutherford model does not make any predictions about the arrangement of electrons around the nucleus or their energies. Finally, Rutherford's model does not explain why the nuclei of *some* atoms are unstable and, hence, radioactive (see Section 7.2).

The Bohr model

A new model was needed. The Danish physicist Niels Bohr came up with a model that could resolve the difficulties with the Rutherford model. In Bohr's model, electrons in an atom can only have certain specific potential energies. These potential energies are the electron **energy levels** of the atom.

Figure 7.7 shows how these energy levels can be represented for the hydrogen atom. The energy levels are shown as a series of lines against a vertical scale of potential energy. All atoms have a similar set of energy levels that converge (get closer together); the energy levels of an atom are like a vertical ladder whose rungs get closer and closer together. The electron in the hydrogen atom can have any one of these potential energy values, but it cannot have an energy between them. The electron's energy in the atom is said to be **quantized** (see Chapter 13).

The energy level labelled zero corresponds to a free electron which is not under the electrostatic influence of the nucleus and is free to move away. The atom is then said to have been ionized. The electrons that are bound to the nucleus are said to have negative energy. To free the electrons, external work must be done against the electrostatic force, which is 'holding' the electrons. This work is equal to the ionization energy.

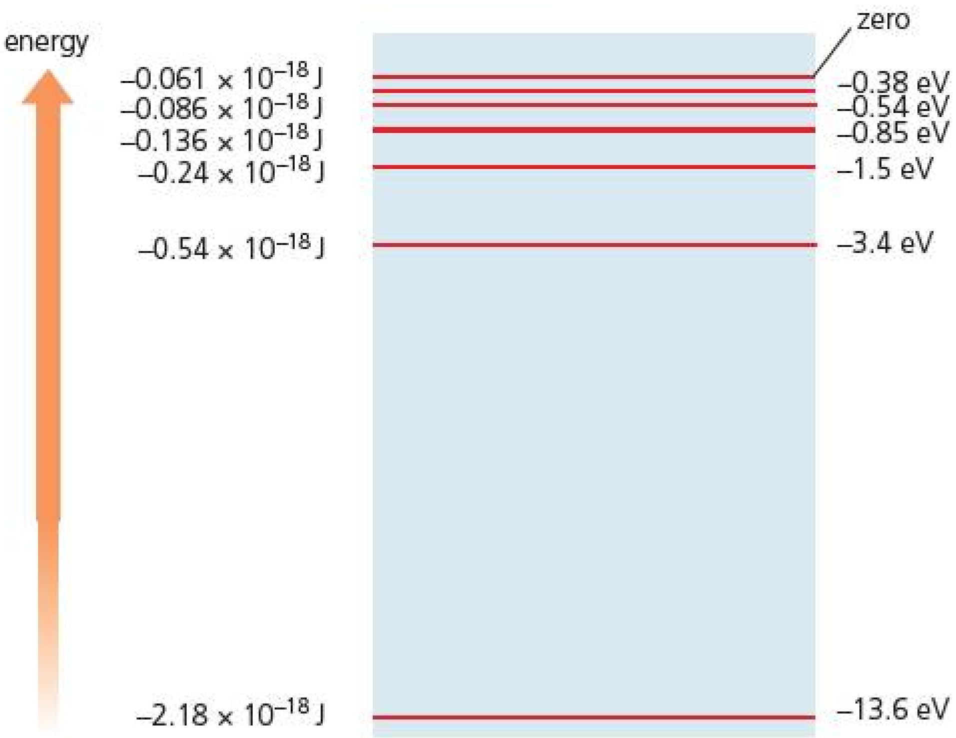


Figure 7.7 Electron energy levels for the hydrogen atom

Electrons in atoms usually occupy the lowest energy levels available. This minimizes the potential energy of the atom, and the atom and its electron(s) are said to be in the **ground state**. Figure 7.8a represents a hydrogen atom in the ground state, as its single electron is in the lowest energy state.

If the electron absorbs energy it may be promoted to a higher energy level. This movement of an electron between energy levels is called an **electron transition**. For it to occur, the energy absorbed must be *exactly equal* to the difference in energy between the two energy levels. When the electron has moved to a higher energy level, the atom and its electron are described as being in an **excited state** (Figure 7.8b).

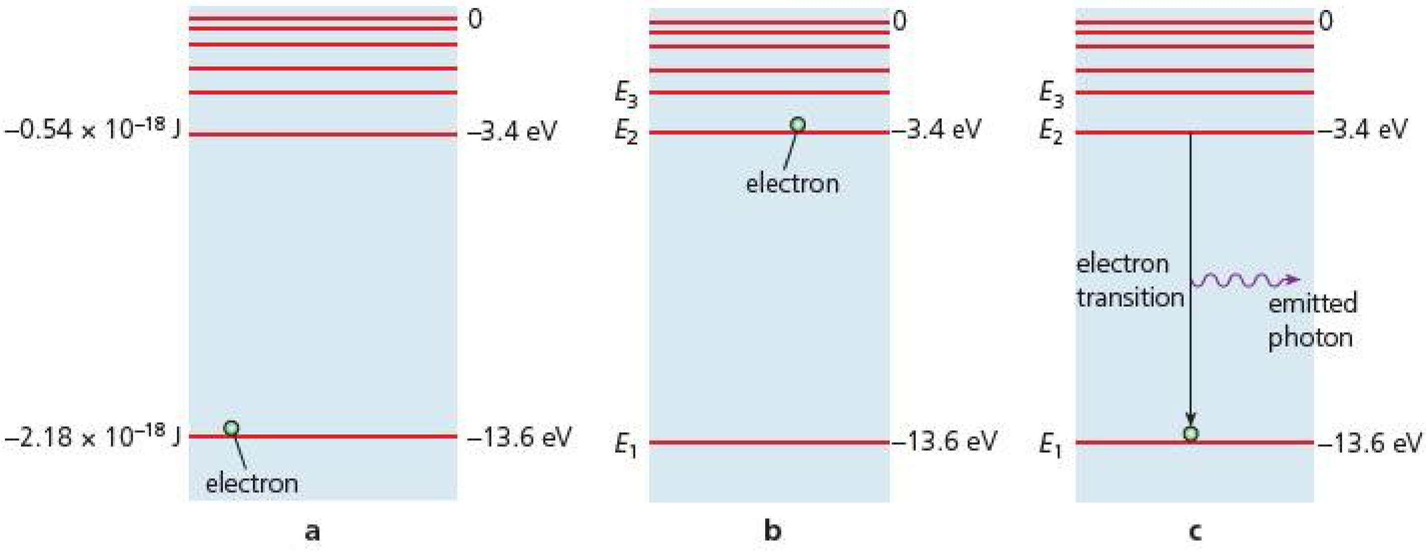


Figure 7.8 A hydrogen atom and its electron: **a** in the ground state, **b** in a short-lived excited state, **c** returning to the ground state

An excited atom is unstable and, after very short period of time, the excited electron will 'fall down' to a lower energy level. For this to occur, the electron must lose its energy. It does this by emitting a photon of electromagnetic radiation (often visible light) with energy equal to the difference between the higher and lower energy level. This is shown in Figure 7.8c.

The energy, hf, of the photon is given by the following relationship:

$$hf = E_2 - E_1$$

where E_2 represents the energy of the higher energy level, E_1 represents the energy of the lower energy level and h is the **Planck constant** (6.63×10^{-34} J s). This is an example of light behaving as 'particles'; a stream of photons emitted as a result of electron transitions. (Evidence for the particle theory of light is derived from a study of the photoelectric effect – see Chapter 13.)

The equation above can be used to find the frequency of the light emitted. For example, when the excited electron in Figure 7.8b returns to the ground state from the second energy level, the energy difference ΔE $(= E_2 - E_1)$ is 1.64×10^{-18} J (equivalent to 10.2 eV). Therefore 1.64×10^{-18} J of energy are carried away by the photon.

The frequency of the electromagnetic radiation is calculated from:

$$f = \frac{\Delta E}{h} = \frac{1.64 \times 10^{-18}}{6.63 \times 10^{-34}} = 2.47 \times 10^{15}\ \text{Hz}$$

Using the wave equation (Chapter 4), the relationship between the speed of light, c, the frequency, f, and the wavelength, λ, of the emitted electromagnetic radiation (light) is given by:

$$\lambda = \frac{c}{f} = \frac{hc}{E}$$

The larger the energy of the electron transition, the higher the frequency (and the shorter the wavelength) of the emitted radiation.

We have seen that a downwards transition, which corresponds to a decrease in the total energy of the atom, results in the emission of a photon. But the atom and its electron can be raised to an excited state by the *absorption* of a photon, provided the photon has the correct amount of energy, corresponding to the exact difference in energy between the excited and ground states. Therefore, a downward transition corresponds to photon emission, and an upward transition to photon absorption.

Types of spectra

Evidence for Bohr's model comes from the study of spectra. When white light from the Sun passes through a prism, the light is dispersed into its component colours. The band of different merging colours (a 'rainbow') is an almost **continuous spectrum** (Chapter 4). Since this spectrum has been produced by the emission of light from the Sun, it is referred to as an **emission spectrum**.

Figure 7.9 A discharge tube containing atoms of neon gas (at low pressure)

Emission spectra from single elements can be examined using a **discharge tube** (Figure 7.9), which is a glass tube containing a small amount of the gaseous element, maintained at very low pressure. When a high potential difference (voltage) is applied across the two electrodes in the tube, energy is transferred to the gas atoms and light is emitted. Examination of the light with a **spectroscope** shows that the emitted spectrum is not continuous, but consists of a number of bright lines. (A simple spectroscope is just an eyepiece and a slit in a tube placed behind a prism.)

The spectrum produced by the gaseous element in the discharge tube is known as a **line spectrum** (Figure 7.10). It consists of a number of discrete or separate colours, each colour being seen as the image of the slit in front of the source. The wavelengths corresponding to the lines of the emission spectrum are characteristic of the gaseous element present in the discharge tube. Each element has it own unique emission spectrum.

If white light is passed through a sample of gaseous atoms or molecules of an element maintained at low pressure and the spectrum analysed (using film or a detector), it is found that light of certain wavelengths is missing. In their place are a series of sharp dark lines.

This type of spectrum is termed an **absorption spectrum** (Figure 7.11).

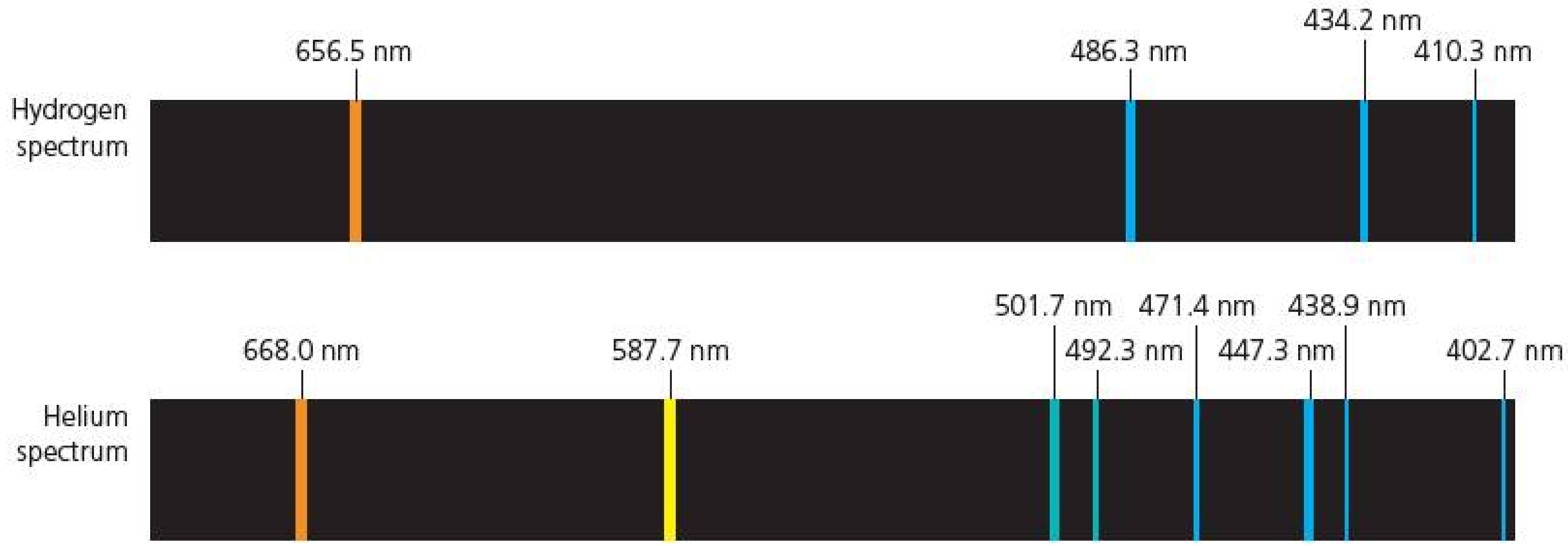


Figure 7.10 Emission spectra of hydrogen and helium atoms from a discharge tube

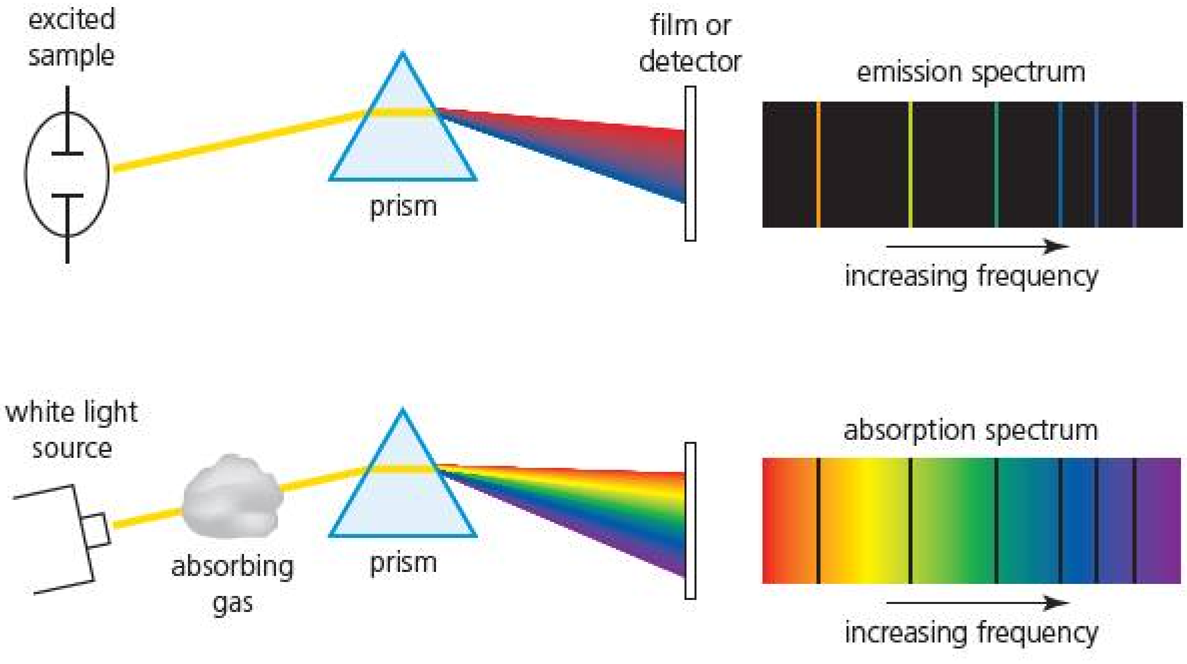


Figure 7.11 Production of emission and absorption spectra of the same element

Interpreting spectra

The existence of lines in emission and absorption spectra can be explained in terms of energy level transitions in atoms. This provides evidence to support the Bohr model of the atom.

As white light passes through a sample of gaseous atoms or molecules, some photons are absorbed by the atoms or molecules and undergo transitions to higher energy levels. The wavelengths of the light that are absorbed correspond exactly to the energies needed by electrons to make transitions from one energy level to a higher level (upward transitions). When the excited electrons return to lower energy levels, photons are emitted in all directions, not just in the original direction of the white light.

Some of the possible electron transitions that may take place when electrons in an excited atom return to lower energy levels are shown in Figure 7.12. Each of the transitions results in the emission of a photon with a particular wavelength. The electron transition from E_4 to E_1 involves the greatest amount of energy, and so results in light with the highest frequency and shortest wavelength. The transition from E_4 to E_3, which involves the least energy, gives light with the lowest frequency and longest wavelength.

All elements have different spacings of energy levels. This means that the energy differences between energy levels are unique to each element, so that each element (in the gaseous state) produces a different and characteristic emission spectrum. These spectra can be used to identify the presence of a particular element in a sample that has been vaporized. This can also be applied to stars by studying the light that they emit (Option E: Astrophysics).

The study of spectra is called **spectroscopy**, and instruments used to measure the wavelengths of spectra are spectrometers (Figure 7.13).

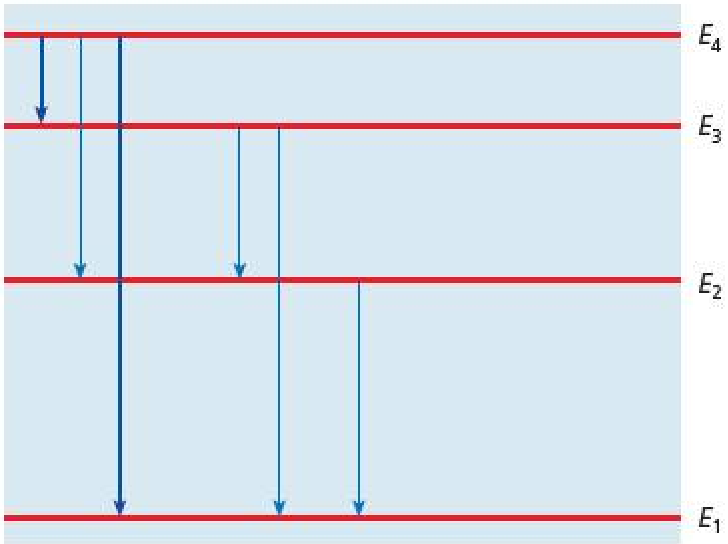


Figure 7.12 Some possible electron transitions involving four energy levels in an excited atom

Figure 7.13 A spectrometer

6 When the spectrum emitted by the Sun is observed closely using a spectrometer it is found that light of certain frequencies is missing and in their place are dark lines.
- **a** Explain how the cool outer gaseous atmosphere of the Sun is responsible for the absence of these frequencies.
- **b** Suggest how an analysis of the solar absorption spectrum could be used to determine which elements are present in the Sun's atmosphere.

7 **a** Find out about the work of Joseph von Fraunhofer and Joseph Lockyer.
- **b** Find out how the Doppler effect uses absorption spectra to measure the speeds of stars and galaxies.

8 Rydberg discovered that all the lines in the atomic emission spectrum of hydrogen could be described by the following expression:

$$f = 1.09 \times 10^7\,\text{m}^{-1}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right), \text{ where } n_1 < n_2$$

Use a spreadsheet to calculate the frequencies of the Lyman series, which are in the ultraviolet region. These are all transitions between other levels and the first energy level. Set n_1 to 1 and let n_2 vary from 2 to a very large number (infinity). Show that the series converges close to a value of 91 nm.

9 a For ultraviolet radiation of wavelength 255 nm calculate the frequency and the amount of energy absorbed by a single molecule when it absorbs a photon of this radiation.

b What is the corresponding energy (in joules) absorbed by one mole of molecules absorbing at this frequency?

10 Calculate the wavelength (in nm) of ultraviolet radiation emitted from an atom when an electron makes a transition from the second energy level (2.0×10^{-17} J) to the first energy level (5.0×10^{-18} J).

Nuclear structure

7.1.5 **Explain** the terms *nuclide, isotope* and *nucleon*.
7.1.6 **Define** nucleon number A, *proton number* Z and *neutron number* N.

The proton number *Z*

Electrons and protons have equal and opposite charges. This makes it possible for electrically neutral atoms to exist. The number of protons in the nucleus of an atom determines which element it is. So, atoms of a particular element are identified by their **proton number** (also called atomic number), which is given the symbol Z. The periodic table of the elements arranges the elements in order of increasing proton number.

The proton number is the number of positive protons in the nucleus of the atom.

Because atoms are electrically neutral, the number of protons is equal to the number of electrons orbiting the nucleus.

Nucleon number *A*

Neutrons and protons are both found in the nucleus of the atom. The term *nucleon* is used for both these particles.

The **nucleon number** is defined as the total number of protons and neutrons in the nucleus.

The nucleon number gives the mass of an atom, since the mass of an electron is negligible. The nucleon number has the symbol A.

Neutron number *N*

The **neutron number** (N) is defined as the number of neutrons in the nucleus.

The difference between the nucleon number (A) and the proton number (Z) gives the number of neutrons in the nucleus (N), i.e. $N = A - Z$.

Isotopes

If two atoms have the same number of protons, they are atoms of the same element. However, two atoms with the same proton number may have *different* numbers of neutrons. The atoms are the same element, but have different nucleon numbers. These atoms are called **isotopes**. The existence of isotopes was established by mass spectrometry (Chapter 13). Isotopes of an element have the same chemical properties.

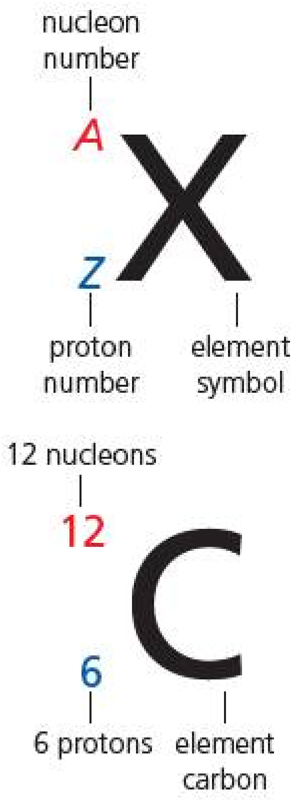


Figure 7.14 Standard notation for specifying a nuclide

Nuclides

The term **nuclide** is used to specify one particular isotope of an element. The term refers to atoms with the same nucleon number as well as proton number. There is a standard notation (Figure 7.14) used to represent nuclides by identifying their proton number Z and nucleon number A. As we have seen, since $A = Z + N$, the number of neutrons, N, is given by $A - Z$.

It is important to make clear the difference between nuclides and isotopes. When referring generally to the nuclei of different atoms we call them nuclides, especially if they are from different elements. However, when referring to atoms of the same element with different nuclei, we call them isotopes.

Some elements have many isotopes, but others have very few or even one. For hydrogen, the most common isotope is hydrogen-1, ${}^{1}_{1}H$. Its nucleus is a single proton. Hydrogen-2, ${}^{2}_{1}H$, is called deuterium; its nucleus contains one proton and one neutron. Hydrogen-3, ${}^{3}_{1}H$, with one proton and two neutrons, is called tritium. Hydrogen atoms (Figure 7.15) are involved in fusion reactions (see Section 7.3 and Chapter 8).

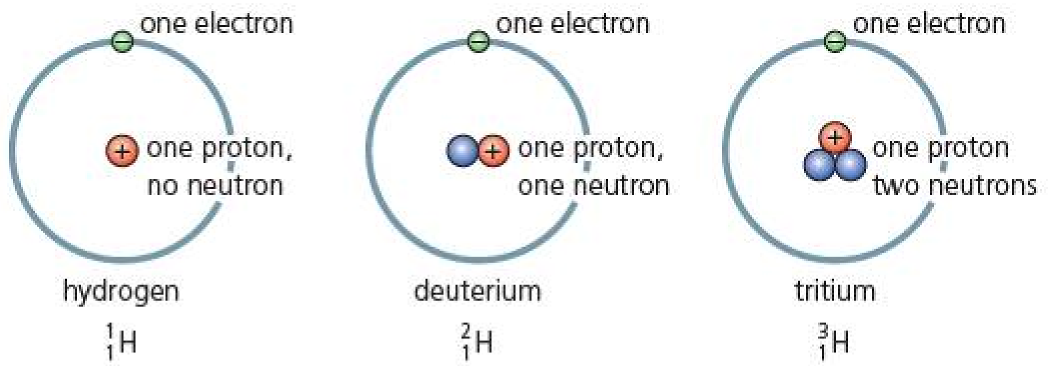


Figure 7.15 The three isotopes of hydrogen

The following nuclides are isotopes of carbon:

${}^{12}_{6}C$ (6 protons, 6 neutrons)

${}^{13}_{6}C$ (6 protons, 7 neutrons)

${}^{14}_{6}C$ (6 protons, 8 neutrons)

Samples of elements are mixtures of isotopes. Isotopes cannot be separated by chemical means. Separation can be achieved by processes that depend on the difference in masses of the isotopes, for example the diffusion rate of gaseous compounds.

The notation for describing nuclides can also be applied to the nucleons. For example, the proton can be written as ${}^{1}_{1}p$ and the neutron as ${}^{1}_{0}n$. The electron can also be represented using this notation, since the proton (atomic) number is equal to an atom's electric charge in units of $|e|$, the elementary charge (charge of one electron). In terms of this unit, the charge of the electron is -1, and so the electron can be represented as ${}^{\ 0}_{-1}e$. The mass number of the electron is effectively zero compared to the proton and neutron.

Worked examples

1 Deduce the nucleon, proton and neutron numbers for the following two isotopes of lithium: ${}^{6}_{3}Li$ and ${}^{7}_{3}Li$. Draw simple diagrams showing the structure of these atoms.

Lithium-6: 6 nucleons, 3 protons and $6 - 3 = 3$ neutrons

Lithium-7: 7 nucleons, 3 protons and $7 - 3 = 4$ neutrons

Figure 7.16

2 An oxygen atom nucleus is represented by $^{16}_{8}O$. Describe its atomic structure.

The nucleus has a proton number of 8 and a nucleon number of 16. Thus, its nucleus contains 8 protons and 16 – 8 = 8 neutrons. There are also 8 electrons.

3 A potassium atom contains 19 protons, 19 electrons and 20 neutrons. Deduce its nuclide notation.

Proton number $Z = 19$, and nucleon number $A = 19 + 20 = 39$.
$^{39}_{19}K$

4 State the number of protons and neutrons contained in an atom of the isotope plutonium-239, $^{239}_{94}Pu$.

The number of protons (proton number) $Z = 94$
The nucleon number $(A) = 239$
The neutron number $N = A - Z = 239 - 94 = 145$

11 The nuclides $^{129}_{53}I$, $^{137}_{55}Cs$ and $^{90}_{38}Sr$ were all formed during atomic weapons testing. State the number of neutrons, protons and electrons in the atoms of these nuclides.

12 What is the electric charge of the nucleus $^{4}_{2}He$?

13 The number of electrons, protons and neutrons in an ion of sulfur are equal to 18, 16 and 16, respectively. What is the correct nuclide symbol for the sulfur ion?

14 State the number of nucleons in one carbon-13 atom, $^{13}_{6}C$.

Interactions within the nucleus

7.1.7 Describe the interactions in a nucleus.

Protons are positively charged, so there are strong repulsive electrostatic forces operating between protons within a nucleus. These powerful forces are examples of Coulombic or **Coulomb interaction**.

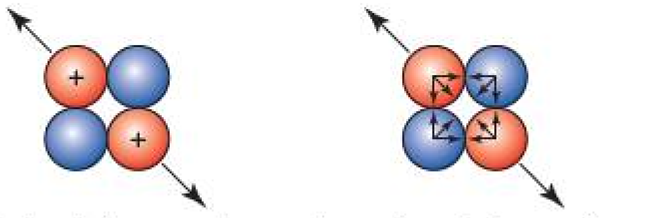


Figure 7.17 Repulsive and attractive forces operating in a helium-4 nucleus (alpha particle)

For a nucleus to be stable there must therefore be a second force which binds all the nucleons (protons and neutrons) together (Figure 7.17). This attractive, very short-range force is known as the **strong nuclear force**. (The strong nuclear force also inhibits the decay of neutrons and the release of beta radiation; see Section 7.2.)

The strong nuclear force is an attractive force and is much stronger than the electrical force if the separation between the two nucleons is very small – about 10^{-15} m. For separations larger than, this the effect of the strong nuclear force becomes negligible. In other words, the strong nuclear force is a short-range force that only extends to its immediate neighbours – it does not extend to all the nucleons within the nucleus.

The relative size of the Coulomb repulsion and the strong nuclear interaction depends, in part, on the ratio of protons to neutrons in a nucleus (see Section 7.2). The balance between these two forces controls nuclear stability.

Large nuclei with large numbers of protons and neutrons are often unstable. The long-range repulsive electrical forces from a large number of protons may be greater than the short-range strong nuclear force.

15 What is the dominant force (or interaction) acting between two protons separated by a distance of:
a 1.0×10^{-15} m
b 3.0×10^{-15} m?

■ **Additional Perspectives**

The strong nuclear force

Protons and neutrons are made of smaller particles called quarks. The strong nuclear interaction is related to the force that binds quarks together. The strong nuclear interaction only acts on particles composed of quarks. It has no long-range effects (i.e. at separations greater than about 10^{-15} m), and does not act on electrons, which are elementary particles that are not composed of quarks. There is another force acting in the nucleus, known as the weak nuclear force, which is responsible for beta decay: the emission of fast-moving electrons from the nuclei of unstable atoms.

Question

1 Find out about the role of gluons in the strong nuclear force.

7.2 Radioactive decay

Radioactivity

Radioactivity is the emission of ionizing radiation by an element due to changes in unstable atomic nuclei. The process by which radioactive atoms change into other atoms is known as **radioactive decay.**

Types of radiation

Alpha radiation

7.2.1 Describe the phenomenon of natural radioactive decay.

Alpha radiation consists of helium-4 nuclei. The nuclei have a relative mass of four and a relative charge of +2. They are represented by the symbols ${}^{4}_{2}\alpha$ or ${}^{4}_{2}\text{He}$. The emission of an alpha particle results in the loss of two protons and two neutrons from a nucleus. Since two protons are lost, the atomic number decreases by two and a new element is formed.

For example, alpha particles are emitted when a radium-226 nucleus undergoes decay, resulting in the formation of a radon-222 nucleus. This is described by the following nuclear equation (see Section 7.3):

$$ {}^{226}_{88}\text{Ra} \rightarrow {}^{222}_{86}\text{Rn} + {}^{4}_{2}\text{He} $$

The process is known as alpha decay.

Beta radiation

Beta radiation consists of streams of high-velocity electrons moving at velocities close to the speed of light. An electron is emitted from an unstable nucleus as the result of a neutron being converted into a proton and an electron in a process called beta decay.

The proton remains in the nucleus, but the electron is ejected at high speed. The moving electron is known as a beta particle and is represented by the symbols ${}^{0}_{-1}\beta$ or ${}^{0}_{-1}\text{e}$. In addition to the beta particle, an additional particle called an electron antineutrino, $\overline{\nu}_e$, is released when a neutron decays in this way.

$$ {}^{1}_{0}\text{n} \rightarrow {}^{1}_{1}\text{p} + {}^{0}_{-1}\text{e} + \overline{\nu}_e $$

This type of beta decay is called beta-negative (β^-) decay.

When beta decay occurs, the number of nucleons remains the same, but the number of protons increases by one, so a new element is formed. For example, beta particles are emitted when a strontium-90 nucleus undergoes beta decay, resulting in the formation of a yttrium-90 nucleus.

$$ {}^{90}_{38}\text{Sr} \rightarrow {}^{90}_{39}\text{Y} + {}^{0}_{-1}\text{e} + \overline{\nu}_e $$

There is another type of beta decay called beta-positive (β^+) decay, which is described in Chapter 13.

Gamma radiation

Gamma radiation (γ) is high-energy electromagnetic radiation released from the nucleus of an atom. Gamma rays generally have shorter wavelengths and higher energies than X-rays, which are produced by electron transitions *outside* the nucleus. Since gamma rays have a high energy and short wavelength, they have a long range and a high penetrating power. Gamma rays are often emitted when nuclides emit alpha or beta particles.

For example, when a thorium-234 nucleus is formed from a uranium-238 nucleus by alpha decay, the thorium nucleus contains excess energy and is said to be in an excited state (see Chapter 13). The excited thorium nucleus (shown by the symbol * in the equation) returns to the ground state by emitting gamma rays:

$${}^{234}_{90}\mathrm{Th}^* \rightarrow {}^{234}_{90}\mathrm{Th} + {}^{0}_{0}\gamma$$

This process is known as gamma decay.

Deflecting radiation in fields

7.2.2 **Describe** the properties of alpha (α) and beta (β) particles and gamma (γ) radiation.

Alpha and beta radiation both consist of a stream of charged particles. Hence they can be deflected by both magnetic and electric fields. Gamma radiation is uncharged, so it cannot be deflected.

Figure 7.18 shows the deflections of the three types of ionizing radiation in a strong magnetic field. Fleming's left-hand rule (see Chapter 6) can be applied to confirm the deflection of the alpha and beta particles into circular paths. The degree of deflection of a charged particle in a magnetic field depends on its mass-to-charge ratio.

Alpha particles have a much greater mass-to-charge ratio than beta particles and hence undergo less deflection.

Similarly, alpha and beta radiation can be deflected by electric fields, as shown in Figure 7.19. Alpha particles are attracted to the negative plate; beta particles are attracted to the positive plate. The deflection of the alpha particles is small in comparison to beta radiation, again due to the greater mass-to-charge ratio.

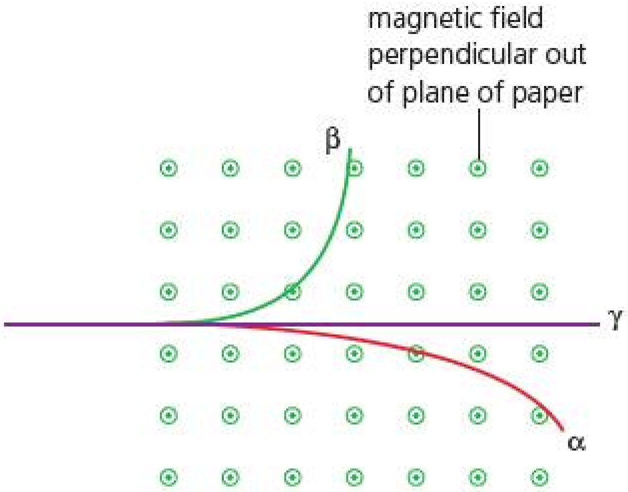


Figure 7.18 Behaviour of ionizing radiations in a magnetic field

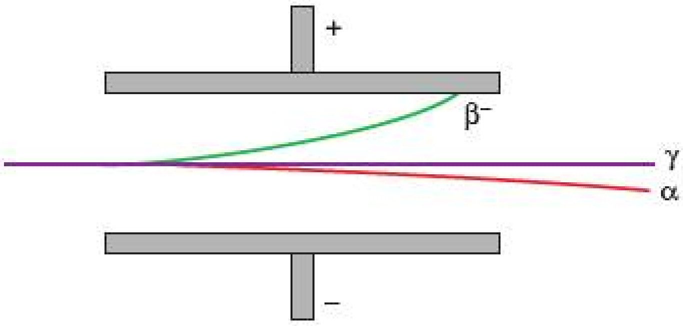


Figure 7.19 Behaviour of ionizing radiations in an electric field

Properties of nuclear radiation

7.2.3 Describe the ionizing properties of alpha (α) and beta (β) particles and gamma (γ) radiation.

As alpha particles travel through matter, they collide with nearby atoms causing them to lose one or more electrons. The ionized atom (ion) and the resulting free electron are called an ion pair (Figure 7.20).

The production of an ion pair requires the separation of unlike charges, hence this process requires energy. Alpha particles have a relatively large mass and charge, so they are efficient ionizers. They may produce as many as 10^5 ion pairs for every centimetre of air through which they travel. As a result, they lose energy relatively quickly, and have low penetrating power.

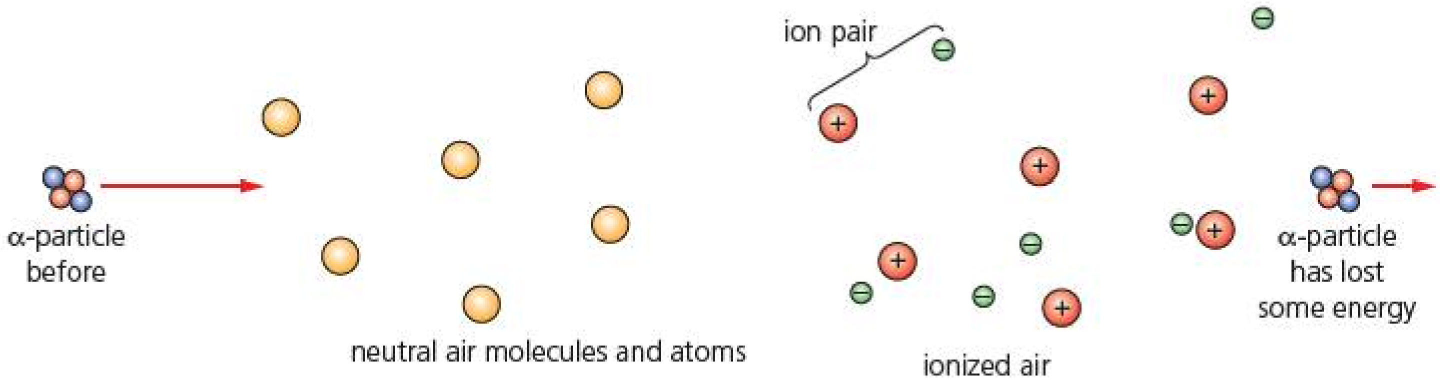


Figure 7.20 Formation of ion pairs by alpha particles from molecules and atoms in the air

Beta particles travel at much greater speeds than alpha particles. They have a smaller charge and considerably less mass than alpha particles. Consequently, they are much less efficient than alpha particles in producing ion pairs. They are, therefore, far more penetrating than alpha particles.

Since gamma rays have no charge or mass, their ionizing power is much less than that of alpha or beta particles.

16 Alpha particles lose about 5×10^{-18} J of kinetic energy in each collision they make with an atom or molecule in the air. An alpha particle travelling through air makes 10^5 ionizing collisions with molecules or atoms in the air for each centimetre of travel. Calculate the range of an alpha particle if the particle begins with an energy of 4.7×10^{-13} J.

17 Alpha particles are stopped much more easily than beta particles with the same energy, yet alpha particles are more massive than beta particles. Explain this.

18 Why does ionizing radiation travel further in air if the pressure is reduced?

Additional Perspectives

Detecting radioactivity

The methods used to detect radioactive emissions are all based on the ionizing properties of the particles or radiation.

The Geiger counter

Figure 7.21 shows a Geiger–Müller tube connected to a ratemeter or scaler. When an ionizing radiation particle enters the tube it ionizes the argon atoms and creates ion pairs: $Ar \rightarrow Ar^+ + e^-$. The argon ions and electrons (the ion pairs) are accelerated by the potential

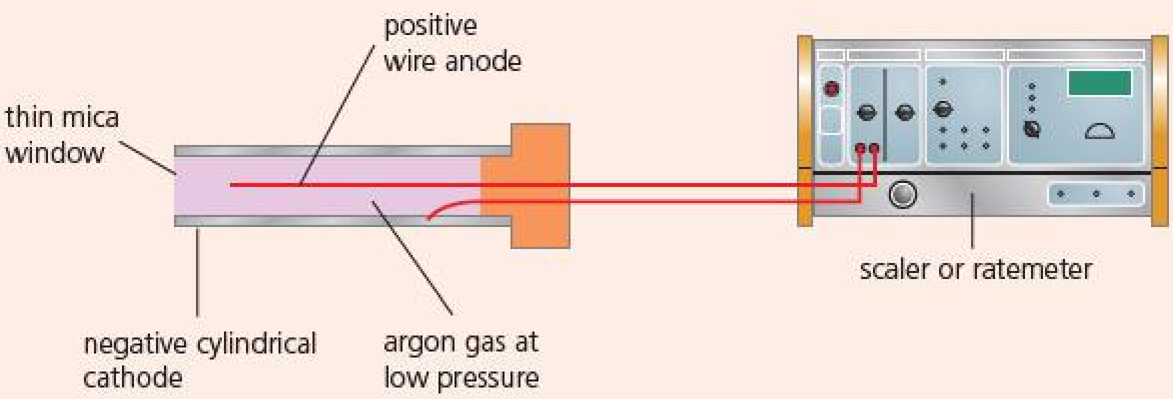


Figure 7.21 Geiger–Müller tube with a scaler connected to it

difference between the anode and cathode. The accelerated particles then cause further ionization (an avalanche effect). For each radioactive particle that enters the tube and produces one argon ion pair, a very large number of electrons and argon ions arrive at the anode and cathode. This gives a pulse of electrons (charge) which is counted (after amplification) by the scaler or ratemeter. A scaler measures the total count of pulses in the tube during the time that it is operating and a ratemeter records the number of counts per second.

The scintillation counter

Early researchers working with radioactive materials, such as Rutherford, used glass screens coated with zinc sulfide to detect radiation. When radiation strikes the surface of the zinc sulfide, it emits a tiny flash of light called a scintillation. How quickly these pulses are emitted indicates the intensity of the radiation. Originally, the flashes of light were counted manually using a microscope, but now a scintillation counter is used (Figure 7.22).

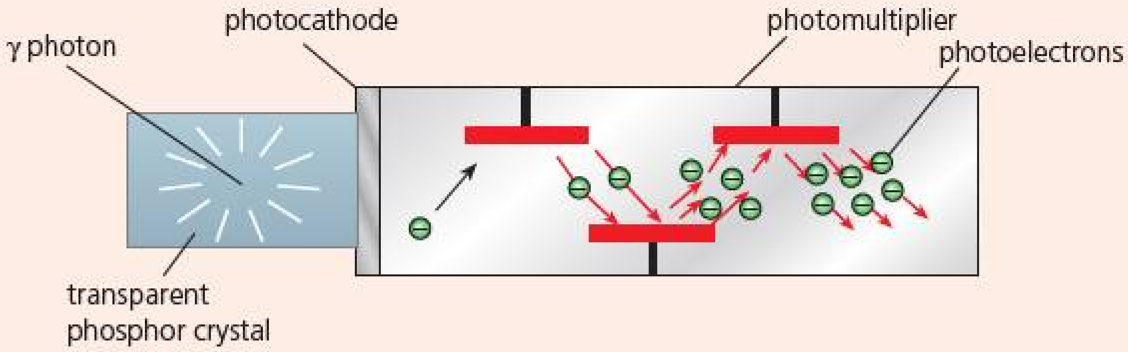


Figure 7.22 Scintillation counter

Questions

1 Find out how scintillation counters can be used in national and homeland security.
2 The scintillation counter was invented during the Manhattan Project. Find out about this research and development program.

Table 7.2 summarizes the properties of the three ionizing radiations.

Table 7.2 Summary of the properties of alpha, beta and gamma radiations

Property	Alpha (α)	Beta (β)	Gamma (γ)
Relative charge	+2	−1	0
Relative mass	4 u	$\frac{1}{1840}$ u	0
Typical penetration	5 cm air; thin paper	30 cm air; few millimetres of aluminium	Highly penetrating – partially absorbed by thick dense materials
Nature	Helium nucleus (particle)	Electron (particle)	Electromagnetic wave/photon
Typical speed	$10^7\,m\,s^{-1}$ $\left(\frac{1}{10}c\right)$	$\approx 2.7 \times 10^8\,m\,s^{-1}$ $\approx\left(\frac{9}{10}c\right)$	$3.00 \times 10^8\,m\,s^{-1}$ c
Notation	^4_2He or $^4_2\alpha$	$^0_{-1}e$ or $^0_{-1}\beta$	γ or $^0_0\gamma$
Relative approximate ionization ability	1 Heavy or very high	$\frac{1}{100}$ Light	$\frac{1}{10000}$ Very low
Relative approximate penetration	1 least (stopped by a thick piece of paper)	100 (stopped by 3 mm aluminium)	10 000 most (intensity halved by 2 cm thick lead)

Figure 7.23 shows apparatus used to study the absorption and relative ionizing ability of the three ionizing radiations. The distance between the Geiger–Müller tube and the radioactive source should not be more than 5 cm, as an air column of this thickness will absorb most of the alpha particles.

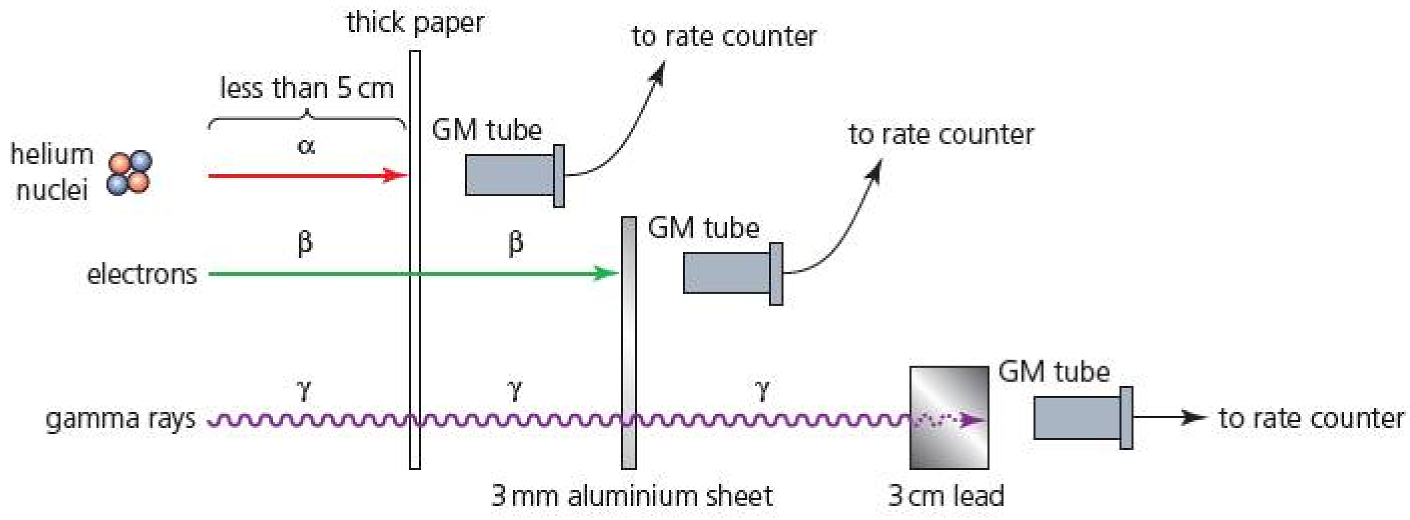


Figure 7.23 Absorption of ionizing radiations

Effects of exposure to ionizing radiation

7.2.4 Outline the biological effects of ionizing radiation.

All the effects of radiation depend on the way they transfer energy and cause ions to be formed from atoms and molecules. All radiation that can cause ionization is dangerous to humans. Ionizing radiation comes in many forms, chemical and physical, and from many sources. It includes the radiation from natural radioactive substances and radiation from artificial sources, such as X-ray machines in hospitals and nuclear reactors (see Chapter 8).

X-rays have the same properties and the same effects on humans as natural gamma radiation. Nuclear reactors produce large amounts of gamma radiation and radioactive waste materials. Nuclear reactors also produce large numbers of neutrons, which are released from the nuclei of uranium atoms. High-speed neutrons are also a form of dangerous ionizing radiation. Hence ionizing radiation includes both electromagnetic radiation of high frequency (X-rays and gamma rays) and fast-moving alpha particles, beta particles (electrons), neutrons and protons.

The natural radiation to which people are exposed to every day during their lives is called **background radiation**. In most countries the majority of the average total radiation dose received by a person comes from background radiation. The pie chart in Figure 7.24 shows where the background dose of radiation comes from in the UK.

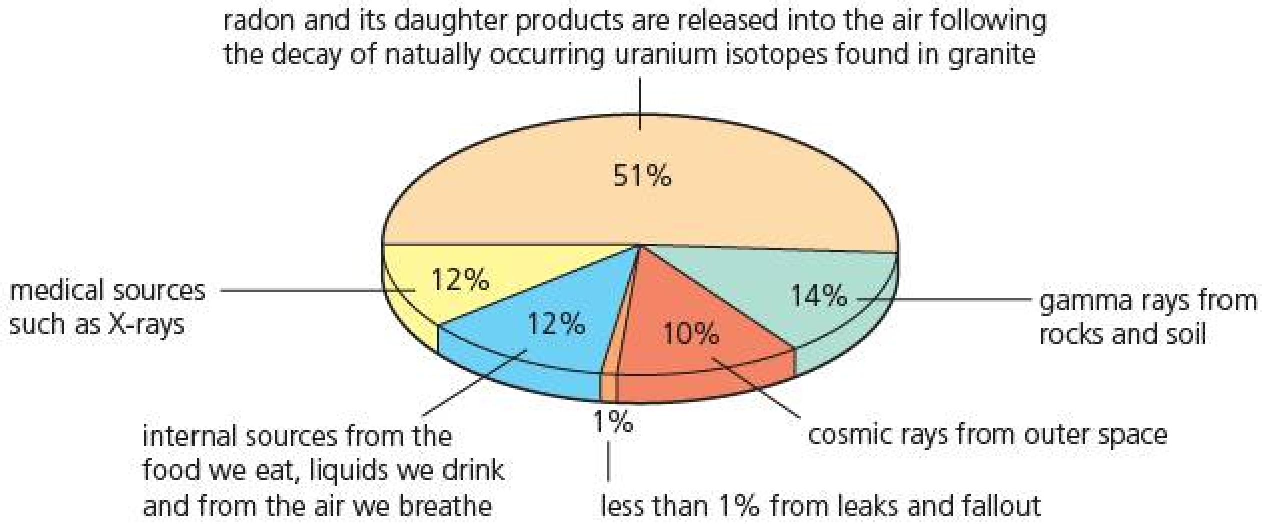


Figure 7.24 Sources of background radiation in the UK

The internal sources are radioactive nuclides, such as potassium-40, ${}^{40}_{19}K$, which are found in our bodies. The radon gas in the air comes from naturally occurring radioactive materials such as radium, thorium, or uranium in the Earth's crust. Most of the cosmic radiation from space is absorbed by the Earth's atmosphere, but some reaches the ground. However, people in high-flying aircraft and astronauts receive a much larger cosmic radiation dose. Cosmic rays include fast-moving nuclei of hydrogen and helium atoms.

The effect of radiation on humans

People are exposed to a variety of ionizing radiations which are health risks. The danger is due to the absorption of energy from the radiation by tissues of the body. This results in the formation of ions, which can kill or change living cells.

Very high doses of radiation can cause cells to stop functioning and prevent them from undergoing cell division, resulting in the death of the cell. Widespread damage of cells in different tissues can result in death. There are also possible long-term delayed effects of ionizing radiation, such as sterility and cancers, especially leukaemia (cancer of white blood cells) and inherited genetic defects (mutations) in the children of people who have been exposed to radiation.

The short-term effects of exposure to large doses of radiation include radiation burns (redness and sores on the skin) mainly caused by beta and gamma radiations. There may also be blindness and formation of cataracts in the eyes. Radiation sickness occurs when a person has a single large exposure to radiation. Nausea and vomiting are the main symptoms, together with fever, hair loss and poor wound healing.

Medical experts are uncertain about the long-term effects of exposure to low doses of radiation. Estimates of the risk of radiation damage are based on the assumption that as the dose of radiation becomes smaller, so the risk is reduced in proportion.

TOK Link: Effects of acute doses of radiation

The predicted effects on humans of a specific acute dose of ionizing radiation are statistical expectations based upon the results from large groups of people, often the survivors of the atomic bombs in Japan at the end of World War Two. For example, an acute exposure to 400 rem of radiation predicts 60% fatality after 30 days. The rem is a measure of the biological effect of radiation. An acute dose of radiation is a large dose of radiation received over a short time period.

It should be noted that the predicted death rate is based on the laws of probability. There have been cases of people receiving much higher acute doses of radiation and surviving, and others with much lower doses who have not. Cause and effect become complicated when working with probability issues. There are many other factors involved, such as age, health, medical care, or lack of medical care.

Question

1 What are the moral, social and environmental aspects to consider when a country decides to commission commercial nuclear power reactors? What 'ways of knowing' are used by people when they consider whether they oppose or support the use of nuclear power?

Precautions for humans

The main risks from alpha and beta radiation come from sources that get inside a person. Since alpha and beta particles do not penetrate very far into the body, the risks from external solid or liquid sources are fairly small. However, care must be taken to stop radioactive materials from being eaten or inhaled (breathed in) from the air. So no eating, drinking, or smoking is allowed where any radioactive materials are handled, and disposable gloves and protective clothing are worn. Masks are worn in mines where radioactive dust particles are airborne.

Gamma radiation and X-rays can be absorbed deep inside the body, and people exposed to external sources of X-rays and gamma radiation must be protected as much as possible. The dose a person receives can be limited in a number of ways: by using lead shielding, by keeping a large distance between the person and source, and by keeping exposure times at short as possible.

People who work with ionizing radiation may wear a film badge which gives a permanent record of the radiation dose received. Workers are also checked for radiation contamination by using sensitive radiation monitors before they leave their place of work. A worker handling radioactive materials may use remote-controlled tools and sit behind a shielding wall made of thick lead and concrete.

When radioactive materials are used in medicine they are chosen carefully to have the least damaging effects on the body.

19 Find out about the difference between an acute and chronic dose of radiation.

20 Find out why different tissues within the body have different sensitivities towards ionizing radiation.

Nuclear stability

7.2.5 Explain why some nuclei are stable while others are unstable.

Within the nucleus there is a delicate balance between the electric force (Coulomb repulsion that pushes the protons apart) and the strong nuclear force (which pulls the nucleons together). Neutrons help to achieve this balance, since their presence in the nucleus increases its size and so increases the distance between the protons.

The main factor that determines nuclear stability is the ratio of protons to neutrons in the nucleus. If all the stable nuclides are plotted on a graph of neutron number (N) versus proton or atomic number (Z), there is a clear pattern (Figure 7.25).

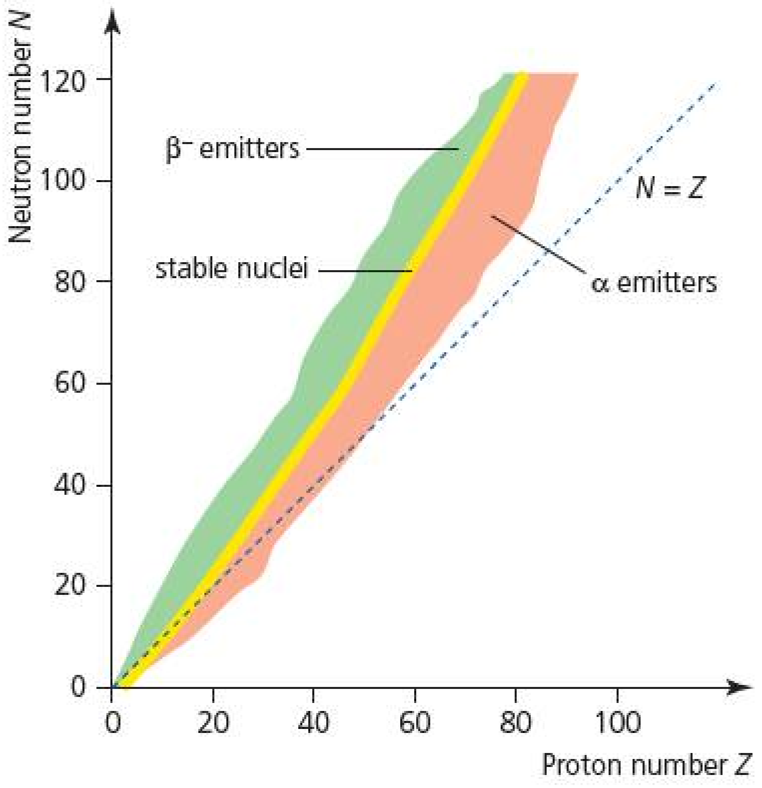


Figure 7.25 A plot of neutron number versus proton number for stable nuclides

For light nuclei to be stable, $N = Z$ (blue dotted line); for heavy nuclei to be stable, $N > Z$. The additional neutrons in heavy nuclei help overcome the Coulomb repulsion of the protons. There is a band of stability that runs up the centre of the graph.

Nuclei above the stability band have 'too many neutrons'. Heavy nuclei ($Z > 80$) are likely to decay by emitting alpha particles (these reduce N and Z by two each) and their products, which also lie above the stability band, are likely to be beta emitters. Beta-negative decay converts a neutron to a proton. This reduces N by one and increases Z by one, moving the nucleus down and to the right (Figure 7.26). This brings it closer to stability.

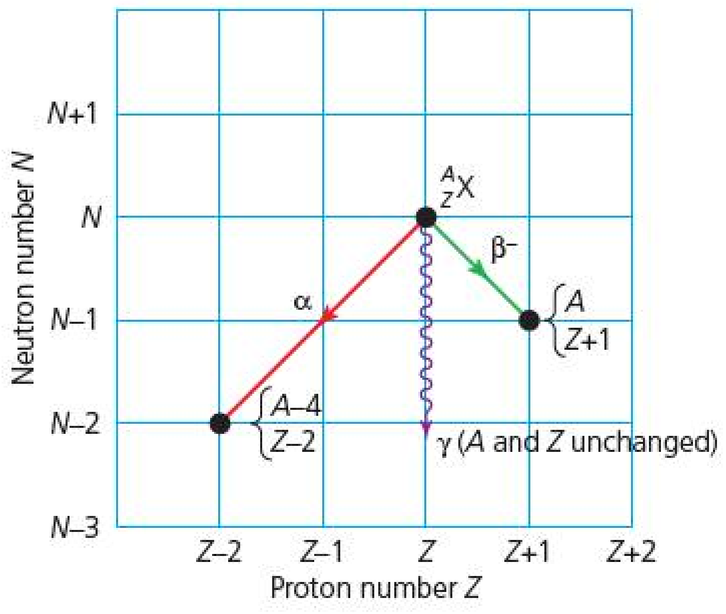


Figure 7.26 The effect of alpha and beta emission on proton number Z and neutron number N (the nucleon number $A = N + Z$)

21 Consider the N–Z plot for isotopes (Figure 7.25). Explain where alpha and beta-negative emitters are found. Explain their positions on the N–Z plot.

TOK Link: The liquid drop model

A model called the 'liquid drop' model for the nucleus provides a simple way of visualizing how the nucleus behaves, regarding it as being made of a sort of 'nuclear fluid'. The strong force between nucleons has very similar properties to the intermolecular force (hydrogen bonds) between molecules in a molecular liquid, such as water. There is a repulsive component which makes the nuclear fluid incompressible, like a real liquid, and an attractive component that keeps the nucleons in the nucleus. This model expects the nucleus to adopt a spherical shape.

However, if a slow-moving neutron enters a large nucleus it will set it oscillating – this process occurs during the fission of uranium-235 (see Section 7.3). If the nucleus forms a dumb-bell shape it may split into two parts of similar size (Figure 7.27). The two smaller nuclei have less energy than the single larger nucleus.

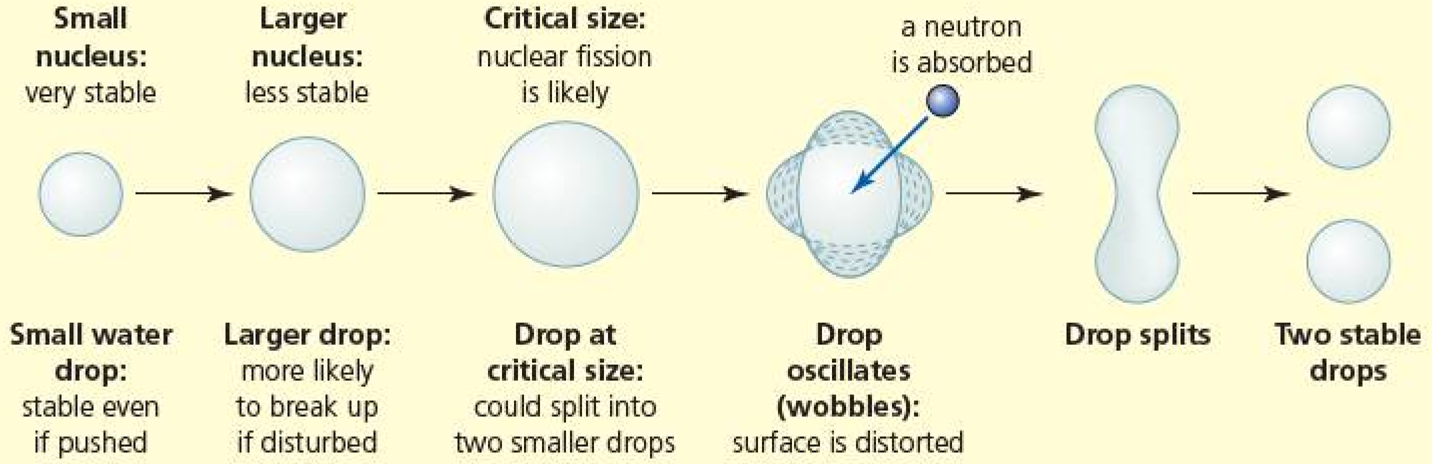


Figure 7.27 The liquid drop model of nuclear fission

Question

1 Find out about the assumptions and limitations of the liquid drop model as a description of the nucleus.

Half-life

The process of radioactive decay

7.2.6 State that radioactive decay is a random and spontaneous process and that the rate of decay decreases exponentially with time.

Radioactive decay is the *spontaneous* and *random* disintegration of nuclei.

During the process of radioactive decay, a nucleus emits a particle (alpha or beta particle) and/or a high-energy photon (gamma radiation). Spontaneous means that the process happens by itself and random means you cannot predict when a particular nucleus is going to decay (change).

Radioactive decay is spontaneous because the decay of nuclei is not affected by chemical reactions (because chemical reactions involve the electrons in the energy levels) or external factors, such as temperature and pressure. The decay of a nucleus is not affected by the presence of other nuclei. Radioactive decay is not like chemical or biological decay, for example, the source may not change in appearance.

Although the decay of individual nuclei is unpredictable, for a very large number of nuclei, it is possible to predict the approximate percentage of nuclei that will decay in a small interval of time. The accuracy of this prediction is improved when the number of radioactive nuclides is very large, as is usually the case. Each nucleus in a sample of radioactive nuclides has the same probability (chance) of decay per second.

Figure 7.28 Dice experiment simulation of radioactive decay

The process of radioactive decay can be simulated by the throwing of dice (Figure 7.28). Consider the following experiment in which a group of students throw 6000 dice. Each time a six is thrown, that die is removed. The results for the number of dice after each throw are shown in Table 7.3.

Table 7.3 Results of dice throwing experiment

Number of throws	Number of dice remaining	Number of dice removed
0	6000	
1	5020	980
2	4163	857
3	3485	678
4	2887	598
5	2420	467
6	2009	411
7	1674	335
8	1399	75

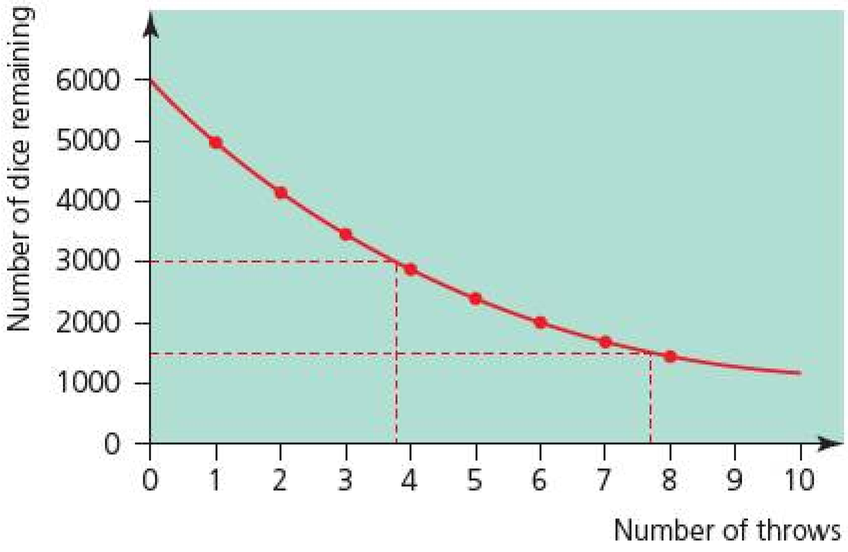


Figure 7.29 Graph of the number of dice remaining against the number of throws

Figure 7.29 is a graph of the number of dice remaining against the number of throws. This is known as a **decay curve**. The rate at which dice (with a six) are removed is not linear. Reading values from the graph shows that approximately 3.8 throws would be required to halve the number of dice. After another 3.8 throws the number of dice has halved again.

The dice experiment is a useful model for describing radioactive decay, with the 6000 dice representing radioactive nuclides. The emission of radioactivity is represented by a six. All dice scoring six are removed, because once a nucleus has undergone radioactive decay, if it is stable, it is no longer capable of undergoing decay.

The graph in Figure 7.29 is an example of a quantity undergoing exponential decay since the rate of decay (modeled by the dice) is directly proportional to the amount (number of dice remaining). The mathematics of exponential decay are described in Chapter 13.

Radioactive half-life

7.2.7 Define the term *radioactive half-life*.

Figure 7.30 is a graph of the number of undecayed radioactive nuclides against time, starting with 500×10^6 nuclides. The graph has the typical decay curve shape. It is not possible to state how long the entire sample will take to decay (its 'life') because decay is a random process, but it is possible to state the time for the number of radioactive nuclides to decrease by half. This is known as the **radioactive half-life**, and in this example its value is 3 minutes. The half-life is often abbreviated to $T_{1/2}$. It is also the time taken for the level of radiation detected to halve.

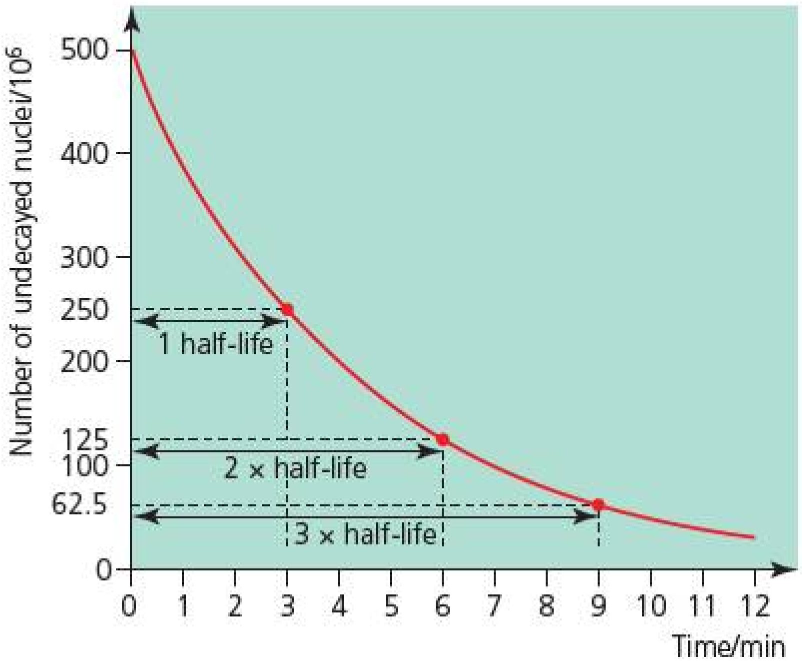


Figure 7.30 Radioactive decay curve

Half-life is defined as the time taken for half the atoms in any given sample of a radioactive substance to decay and release their radiation.

For example, radium-226 (an alpha emitter) has a half-life of 1620 years. This means that if we start with 1 gram of radium-226, then 0.5 grams of it will have undergone radioactive decay in 1620 years. After another 1620 years (one half-life), half of the radium-226 atoms remaining will have decayed, leaving 0.25 grams, and so on.

The half-lives of radioactive isotopes vary considerably. The half-lives of some unstable nuclides are shown in Table 7.4.

Table 7.4 Examples of half-lives of radioactive nuclides

Radioactive nuclide	Half-life
Uranium-238	4.5×10^9 years
Radium-226	1.6×10^3 years
Radon-222	3.8 days
Francium-221	4.8 minutes
Astatine-217	0.03 seconds

Measuring the rate of radioactive decay

7.2.8 Determine the half-life of a nuclide from a decay curve.

The Geiger–Müller tube can be used in school laboratories to determine the half-life of a radioactive substance. The source is placed close to the mica window, which lets in all types of radiation, including alpha particles. The counter has a digital display of the number of radioactive particles or waves that have entered the tube in one second. The measured *count rate* will be proportional to the number of decays occuring every second in the source – which is called the **activity**. Activity is measured in SI units of becquerels (Bq): 1 Bq = 1 decay per second.

The count rate obtained from a Geiger counter will decrease if the source is moved away from the mica window, since less radiation will enter the tube. (The isotope used must have a resasonably short half-life and it is usually extracted from a mixture of isotopes just before use.) The reading is also affected by the background count due to the background radiation. The background count must be deducted from measurements recorded with the source.

Determining half-life from a decay curve

A decay curve is produced by plotting the count rate on the *y*-axis against time on the *x*-axis. Because the activity (and count rate) depends on the number of undecayed nuclei, this is an exponential decay curve. This means that the time taken for a sample of radioactive nuclides to decrease by any particular proportion remains constant. As we have seen above, the time taken for the number of undecayed nuclei in a sample of radioactive nuclides to halve is known as the half-life, $T_{1/2}$ (Figure 7.31).

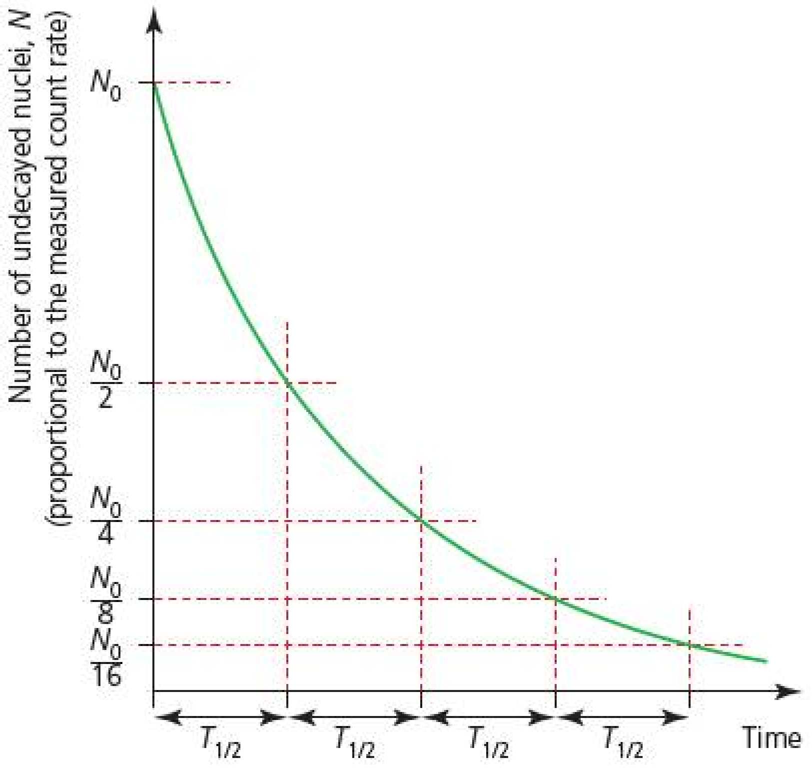


Figure 7.31 Generalized radioactive decay curve

The half-life can therefore be found from the curve by measuring the time taken for the count rate to halve. This should give the same result, whatever initial value is taken. Several values should be taken and a mean calculated.

22 The initial activity of a sample of radioactive nuclide is 8000 Bq. The half-life of the nuclide is 5 minutes. Sketch a graph to show how the activity of the sample changes over a time interval of 25 minutes.

23 The table below shows the variation with time *t* of the count rate of a sample of a radioactive nuclide X. The average background count during the experiment was 36 min^{-1}.
a Plot a graph to show the variation with time of the corrected count rate.
b Use the graph to determine the half-life of the nuclide X.

t/hour	0	1	2	3	4	5	6	7	8	9	10
Count rate/min^{-1}	854	752	688	576	544	486	448	396	362	334	284

Radioactive decay problems

7.2.9 Solve radioactive decay problems involving integral numbers of half-lives.

If you know the half-life of a nuclide, then you can use this fact to calculate the number of undecayed nuclei after a given period of time, or the activity after a period of time.

Figure 7.32 illustrates the radioactive decay of a sample of a type of americium-242 which has a half-life of 16 hours. Initially there are 40 million undecayed nuclei. The grey circles represent undecayed nuclei and the orange circles represent decayed nuclei. With each half-life, half of the remaining undecayed nuclei decay.

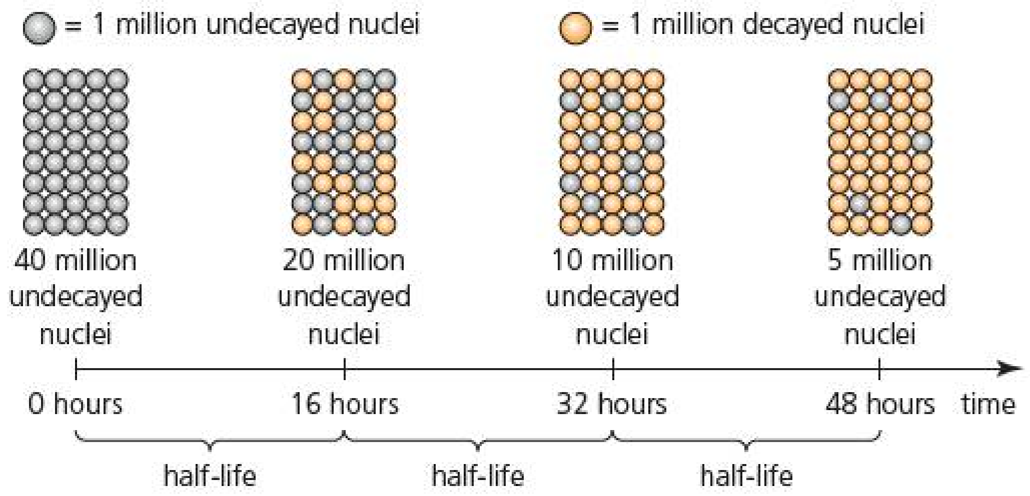


Figure 7.32 The radioactive decay of a sample of americium-242

This process of radioactive decay is summarized in numerical form in Table 7.5.

Table 7.5 The radioactive decay of a sample of americium-242

Number of undecayed nuclei	Fraction of original undecayed nuclei remaining	Number of decayed nuclei	Number of half-lives elapsed	Number of hours elapsed
40×10^6	1	0	0	0
20×10^6	$\frac{1}{2}$	20×10^6	1	16
10×10^6	$\frac{1}{4}$	30×10^6	2	32
5.0×10^6	$\frac{1}{8}$	35×10^6	3	48
2.5×10^6	$\frac{1}{16}$	37.5×10^6	4	64

Worked example

5 The half-life of francium-221 is 4.8 minutes. Calculate the fraction of francium-221 remaining undecayed after a time of 14.4 minutes.

The half-life of francium-221 is 4.8 minutes, so 14.4 minutes is $14.4 \div 4.8 = 3$ half-lives.
After one half-life, half of the sample will remain undecayed.
After two half-lives, $0.5 \times 0.5 = 0.25$ of the sample will remain undecayed.
After three half-lives, $0.5 \times 0.25 = 0.125$ of the sample will remain undecayed.
Hence after 14.4 minutes the fraction of francium-221 remaining undecayed is 0.125 or $\frac{1}{8}$.

The number of undecayed nuclei or activity can also be easily found from using the following equations, provided that the time passed is an integral (whole number) multiple of the half-life.

In terms of the number of undecayed nuclei:

$$N = N_0\left(\frac{1}{2}\right)^n$$

where N represents number of undecayed nuclei, N_0 represents the initial number of undecayed nuclei and n represents the number of half-lives.

Or in terms of the activity:

$$A = A_0\left(\frac{1}{2}\right)^n$$

where A represents the activity of the radioactive source, A_0 represents the initial activity of the source and n represents the number of half-lives.

Worked example

6 Caesium-137 has a half-life of 30 years. If the initial number of caesium nuclei is 10^4, how many caesium nuclei have decayed after 90 years?

90 years is equivalent to $\frac{90}{30} = 3$ half-lives

After 90 years, the number of caesium nuclei that remain is $10^4 \times \left(\frac{1}{2}\right)^3 = 1250$.

After 90 years, the number of caesium nuclei that have decayed is $10^4 - 1250 = 8750$.

24 A radioactive substance has a half-life of 5 days and the initial count rate is 500 min^{-1}. If the background count was found to be 20 min^{-1}, what will be the count rate after 15 days?

25 **a** The half-life of francium-221 is 4.8 minutes. Calculate the fraction of a sample of francium-221 remaining undecayed after a time of 24.0 minutes.

b The half-life radon-222 is 3.8 days. Calculate the fraction of a sample of radon-222 that has decayed after 15.2 days.

c Cobalt-60 is used in many applications where gamma radiation is required. The half-life of cobalt-60 is 5.26 years. A cobalt-60 source has an initial activity of 2.00×10^{15} bequerel. What will be its activity after 26.30 years?

d A radioactive element has a half-life of 80 minutes. How long will it take for the count rate to decrease to 250 per minute, if the initial count rate is 1000 per minute?

e The half-life of radium-226 is 1601 years. For an initial sample:

i what fraction has decayed after 4803 years?

ii what fraction remains undecayed after 6404 years?

Additional Perspectives

Carbon dating

Archaeological specimens with ages up to about 50 000 years can be dated using the isotope carbon-14. Carbon-14, which has a half-life of 5730 years, is constantly being formed in the upper atmosphere. When cosmic rays enter the atmosphere they can produce neutrons. The nuclear reaction shown below can then take place:

$${}^{14}_{7}N + {}^{1}_{0}n \rightarrow {}^{14}_{6}C + {}^{1}_{1}H$$

This type of reaction is an example of a natural transmutation (see Section 7.3) in which a nuclide absorbs another smaller particle into its nucleus and undergoes a nuclear reaction to form a new nuclide.

The carbon-14 (present in the atmosphere as carbon dioxide, $^{14}CO_2$) is then taken in by plants during photosynthesis. The plants are eaten by animals, which may then be eaten by other animals. However, once the animal or plant dies no more radioactive carbon-14 is taken in and the percentage in the remains starts to decrease due to radioactive decay (Figure 7.33).

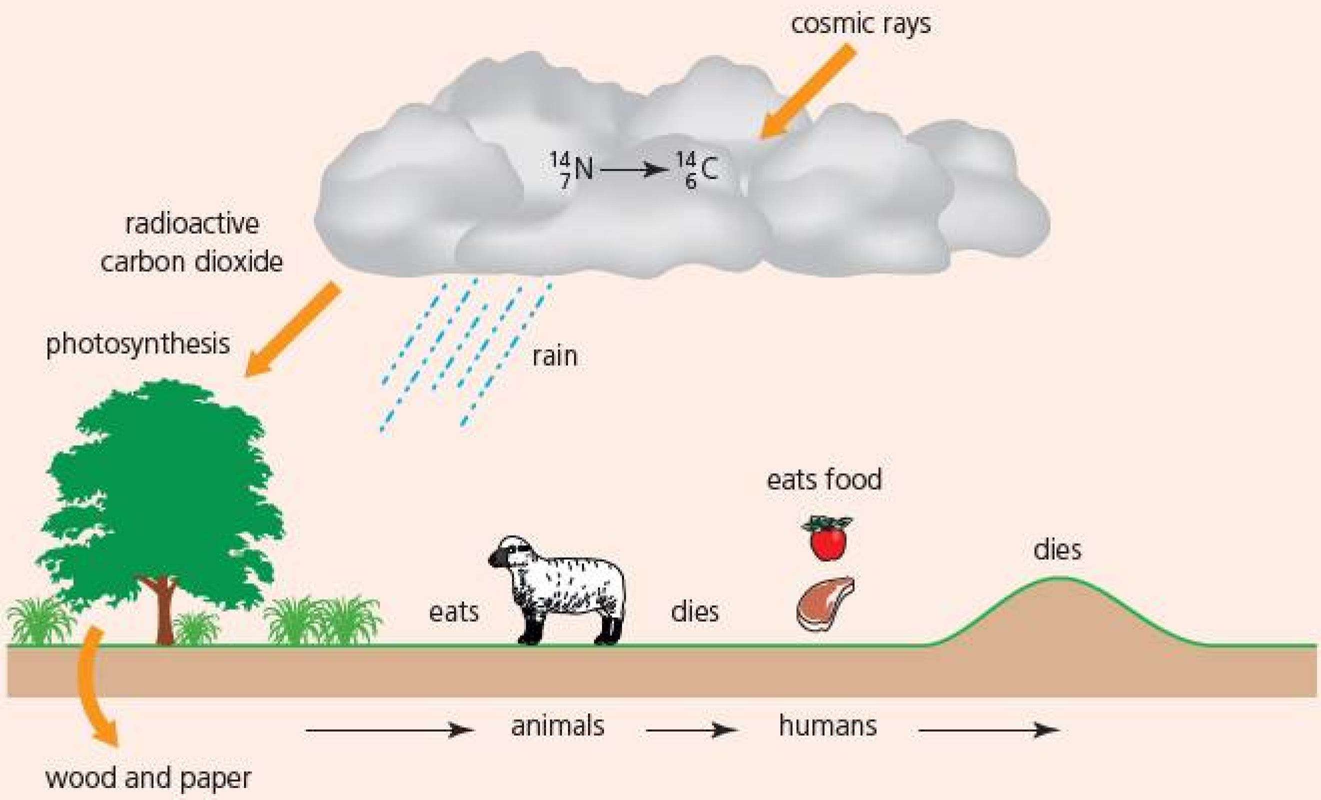


Figure 7.33 Carbon-14 is incorporated into living tissue

This means that if the ratio of carbon-14 to carbon-12 is known then the age of the specimen can be determined once the amount of carbon-14 remaining in it has been measured. The activity of carbon-14 in living materials is about 19 counts per minute for each gram of specimen. This method of dating can be used to determine the ages of animal remains and wood, paper and cloth.

Question

1 Research the use and importance of carbon dating in archaeology.

7.3 Nuclear reactions, fission and fusion

Nuclear reactions

Transmutation

7.3.1 **Describe** and **give an example** of an artificial (induced) transmutation.

Nuclear transmutation is the transforming (changing) of one element/isotope into another element/isotope. Radioactive decay of unstable atoms is an example, but stable atoms can sometimes be transformed into radioactive atoms by bombardment with high-speed particles (for example, the formation of carbon-14 in the atmosphere as discussed above).

Many *artificial* transmutations are carried out to prepare nuclides in medicine (diagnostic and therapy) and industry (imaging, tracers, process monitoring, etc.). They are made by neutron bombardment in nuclear reactors and by proton, deuteron $\left(^{2}_{1}H\right)$ or alpha particle bombardment by accelerators.

In the earliest experimental studies, high-speed alpha particles from bismuth-214 were used to carry out artificial transmutation. In 1919, Rutherford and his collaborators in the Cavendish Laboratory at Cambridge carried out the first nuclear reaction between alpha particles and nitrogen atoms, which is described by the following nuclear equation:

$$^{14}_{7}N + ^{4}_{2}He \rightarrow ^{17}_{8}O + ^{1}_{1}H$$

In this reaction, a nitrogen nucleus reacted with a high-speed helium nucleus to form an oxygen-17 atom and a proton. This transmutation reaction showed the possibility of artificially converting one element into another. Natural transmutation and artificial transmutation both produce new atoms but with one difference; artificial transmutation involves deliberately bombarding the nuclei with a high-speed particle.

For nuclear reactions to occur, the bombarding particles must have a high kinetic energy. This is required to overcome the Coulomb repulsion (assuming the particle is positive) exerted by the target nuclei. The particles therefore have to be accelerated to a high velocity.

Nuclear equations

7.3.2 Construct and complete nuclear equations.

Nuclear equations are used to describe alpha decay, beta decay and transmutation reactions. Nuclear equations must balance: the sum of the nucleon numbers and the proton numbers must be equal on both sides of the equation.(For the sake of simplicity the beta decay equations do not include the antineutrino.)

Alpha decay

During alpha decay the nucleus of a radioactive element ejects an alpha particle and forms a new atom of a different element. The generalized equation for alpha decay is:

$^{A}_{Z}X$	$\rightarrow$	$^{A-4}_{Z-2}X$	+	$^{4}_{2}He$	+	energy
parent nuclide		daughter nuclide		alpha particle		

When a nuclide decays by emitting an alpha particle its proton number, Z, decreases by two and its nucleon number, A, decreases by four. The result of alpha decay is that a new element is produced with a proton number two below its parent in the Periodic table of elements.

Examples of alpha decay:

$$^{241}_{95}Am \rightarrow ^{237}_{93}Np + ^{4}_{2}He + \text{energy}$$

$$^{238}_{92}U \rightarrow ^{234}_{90}Th + ^{4}_{2}He + \text{energy}$$

Beta decay

During beta decay the nucleus of a radioactive element ejects a beta particle and forms a new atom of a different element. The generalized equation for beta decay is:

$^{A}_{Z}X$	$\rightarrow$	$^{A}_{Z+1}X$	+	$^{0}_{-1}\beta$	+	energy
parent nuclide		daughter nuclide		beta particle		

When a nuclide decays by emitting an beta particle its proton number, Z, increases by one but the nucleon number, A, does not change. This can be explained by the decay of a neutron inside the nucleus to produce a proton, which remains in the nucleus, and a beta particle (electron), which is ejected.

Examples of beta decay:

$$^{90}_{38}Sr \rightarrow ^{90}_{39}Y + ^{0}_{-1}\beta + \text{energy}$$

$$^{131}_{53}I \rightarrow ^{131}_{54}Xe + ^{0}_{-1}\beta + \text{energy}$$

$$^{1}_{0}n \rightarrow ^{1}_{1}p + ^{0}_{-1}\beta + \text{energy}$$

Electric charge is conserved because the positive charge of the proton and the negative charge of the beta particle (electron) sum to zero, indicating there is no change in the total charge.

Gamma decay

Gamma particles are often emitted during alpha and beta decay. When a nucleus ejects an alpha or beta particle there is often some excess energy produced during the decay process, resulting in an excited nucleus. This energy is released in the form of gamma rays. After their emission, the nucleus has less energy but its nucleon number and its proton number have not changed.

The generalized nuclear equation for gamma decay is:

$^{A}_{Z}X^{*}$	→	$^{A}_{Z}X$	+	energy (γ ray photons)
parent nuclide		daughter nuclide		

Decay series

Heavy radioactive nuclides, such as radium-226 and uranium-238, cannot become stable by emitting a single radioactive particle. They undergo a radioactive **decay series**, producing either an alpha or a beta particle and gamma radiation during each step, until a stable nuclide is formed. For example, uranium-238 undergoes a decay series (Figure 7.34) to eventually form stable lead-206.

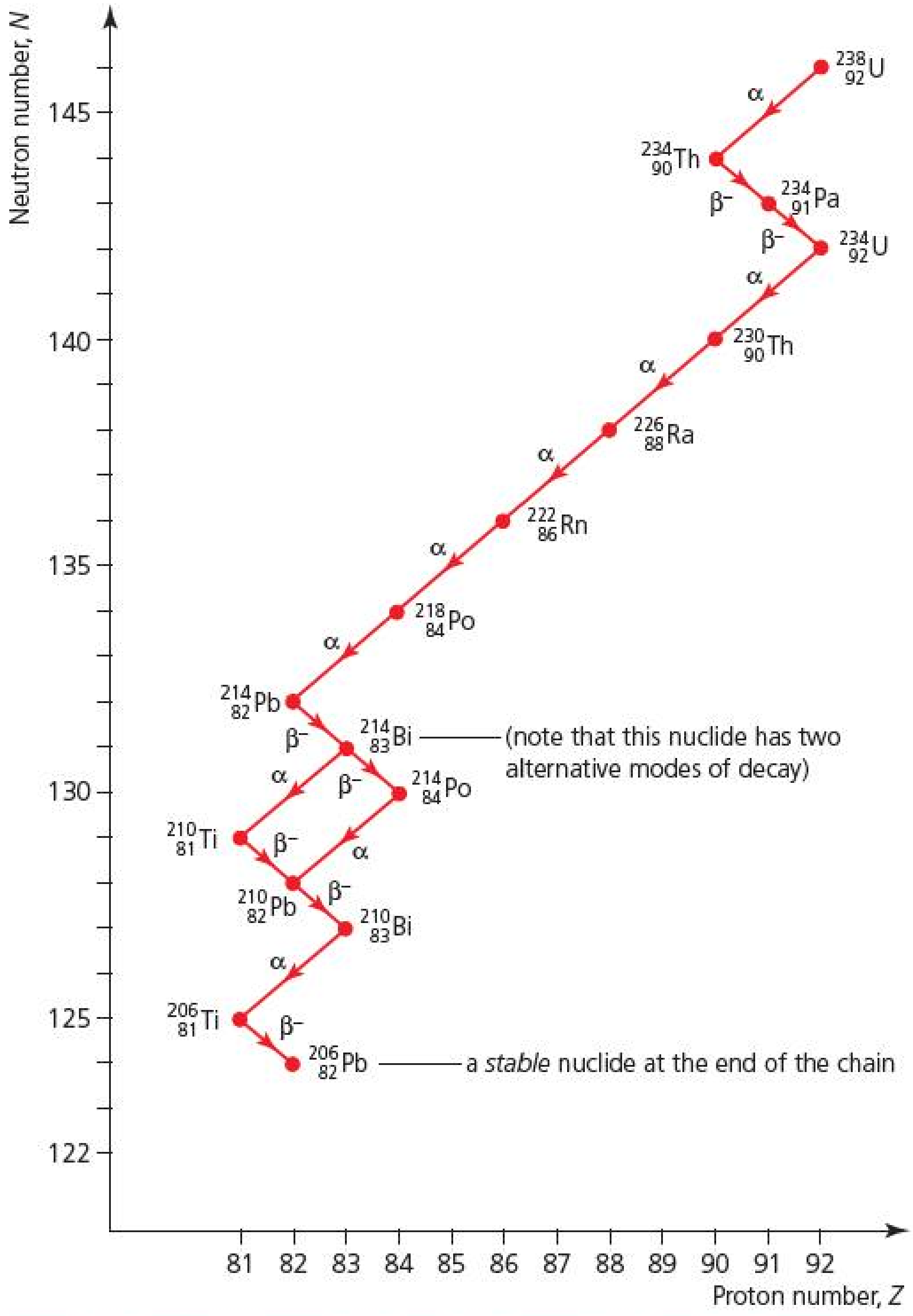


Figure 7.34 An N–Z plot of the decay series for uranium-238. Most of the decay processes shown have short half-lives, measured in seconds, minutes or days

With each decay, the daughter nuclide moves across the N–Z plot. Figure 7.35 summarizes the changes in N and Z that occur with the different types of decay.

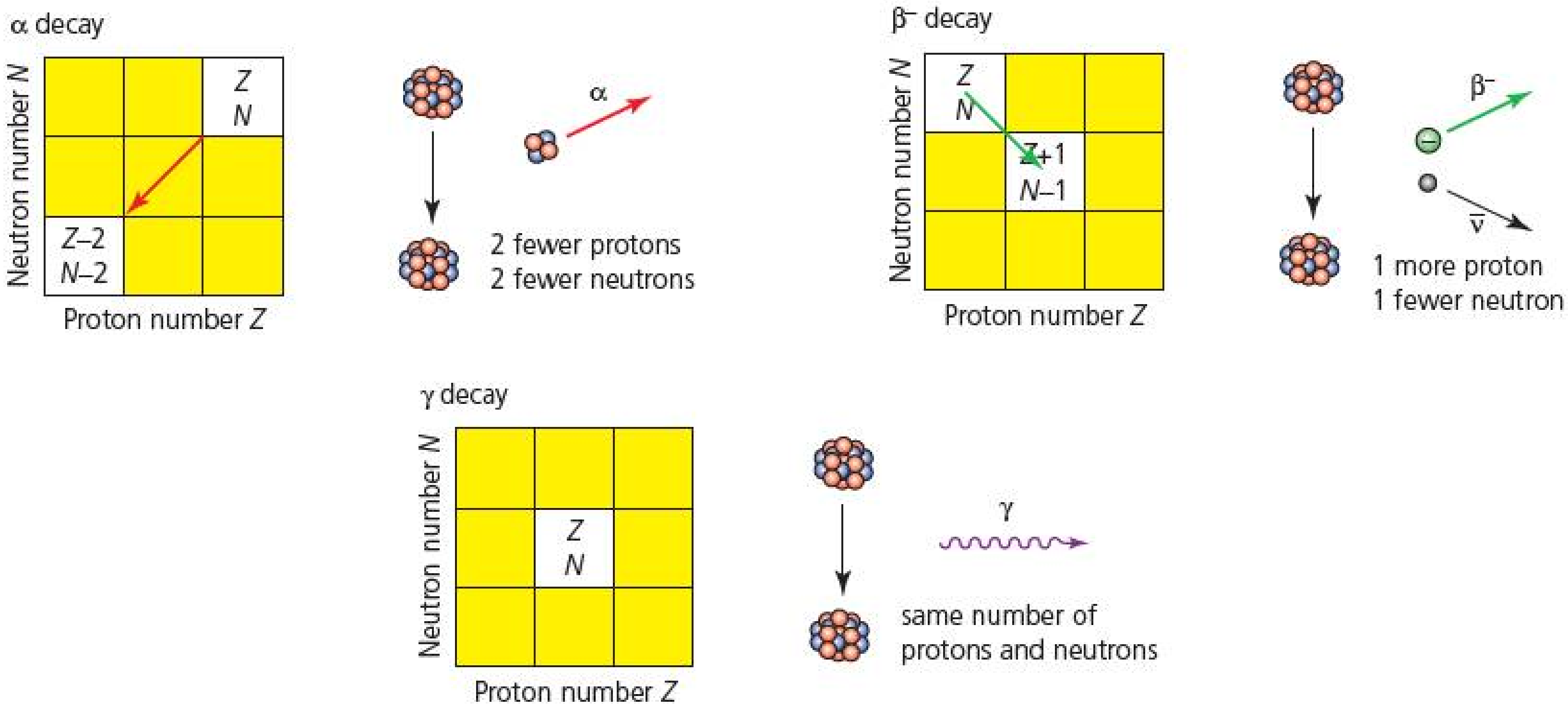


Figure 7.35 Summary of radioactive decay processes

Worked examples

7 Complete the nuclear equations for the transmutations shown below:

a $^{14}_{7}\text{N} + ^{4}_{2}\text{He} \rightarrow ^{17}_{8}\text{O} + ?$

b $^{4}_{2}\text{He} + ^{121}_{51}\text{Sb} \rightarrow ? + ^{1}_{0}\text{n}$

a $^{14}_{7}\text{N} + ^{4}_{2}\text{He} \rightarrow ^{17}_{8}\text{O} + ^{1}_{1}\text{H}$

The addition of the hydrogen atom means that the nucleon numbers balance:
$14 + 4 = 18$; $17 + 1 = 18$.
The proton numbers also balance: $7 + 2 = 9$; $8 + 1 = 9$.

b $^{4}_{2}\text{He} + ^{121}_{51}\text{Sb} \rightarrow ^{124}_{53}\text{I} + ^{1}_{0}\text{n}$

The addition of the neutron means that the nucleon numbers balance:
$4 + 121 = 125$; $124 + 1 = 125$.
The proton numbers also balance: $2 + 51 = 53$; $53 + 0 = 53$.

8 Complete the following nuclear equations. In each case describe the decay process.

a $^{131}_{53}\text{I} \rightarrow \text{Xe} + ^{0}_{-1}\beta$

b $^{241}_{95}\text{Am} \rightarrow \text{Np} + ^{4}_{2}\text{He}$

a $^{131}_{53}\text{I} \rightarrow ^{131}_{54}\text{Xe} + ^{0}_{-1}\beta$ beta decay

b $^{241}_{95}\text{Am} \rightarrow ^{237}_{93}\text{Np} + ^{4}_{2}\text{He}$ alpha decay

26 Fill in the missing values of atomic number and nucleon number for the following part of the thorium decay series:

$$^{232}_{90}\text{Th} \xrightarrow{\text{alpha}} \text{Ra} \xrightarrow{\text{beta}} \text{Ac} \xrightarrow{\text{beta}} {}^{228}\text{Th}$$

27 Write nuclear equations showing the following nuclear reactions: alpha decay of $^{184}_{78}\text{Pt}$, beta decay of $^{24}_{11}\text{Na}$ and gamma decay of $^{60}_{27}\text{Co}$. Use a periodic table to identify the elements formed.

28 Cockcroft and Walton were the first to 'split' the atom, by bombarding lithium-7 atoms with protons from their accelerator. The resulting beryllium-8 atom forms alpha particles. Write a nuclear equation to describe this artificial transmutation.

29 Below is a nuclear equation for a possible fission reaction in a nuclear power station. What particles are required to balance the reaction? Account for their formation – refer to the N–Z curve for nuclides.

$$^{1}_{0}\text{n} + ^{235}_{92}\text{U} \rightarrow ^{236}_{92}\text{U} \rightarrow ^{138}_{53}\text{I} + ^{95}_{39}\text{Y} + ?$$

Unified atomic mass unit

7.3.3 Define the term unified atomic mass unit.

The relative mass of atoms is based on the carbon-12 atom, $^{12}_{6}C$. The **unified atomic mass unit, u**, is defined as one-twelfth of the mass of a carbon-12 atom. Its value u = 1.661×10^{-27} kg.

One mole (see Chapter 3) of carbon-12 has a mass of 12.00000 g and contains $6.022\,141\,29 \times 10^{23}$ carbon-12 atoms (this is a value for N_A to nine significant figures). The molar mass of carbon-12 is 12.00000 g mol^{-1}. Therefore:

$$\text{mass of one carbon-12 atom} = \frac{12.00000}{6.02214129 \times 10^{23}} = 1.992\,647 \times 10^{-23}\,\text{g}$$

$$\text{one unified atomic mass unit (u)} = \frac{1}{12}\,(1.992\,647 \times 10^{-23}\,\text{g}) = 0.166\,0539\,2 \times 10^{-23}\,\text{g}$$

Hence, to four significant figures, 1 u = 1.661×10^{-27} kg (this value is given in the IB *Physics data booklet*).

The **rest masses** of the three fundamental particles in atomic mass units are:

mass of the proton, m_p = 1.007276 u = 1.673×10^{-27} kg
mass of the neutron, m_n = 1.008665 u = 1.675×10^{-27} kg
mass of the electron, m_e = 0.000549 u = 9.110×10^{-31} kg

(All these values are given in the IB *Physics data booklet*.)

The rest mass (m) of a particle is the mass of an isolated particle which is at rest relative to the observer. This comes from the equivalence of mass and energy, described in the next section.

Einstein's mass–energy equivalence

7.3.4 Apply the Einstein mass–energy equivalence relationship.

According to Einstein's special theory of relativity, a mass, m, is equivalent to an amount of energy, E, according to the relationship:

$$E = mc^2 \quad \text{or} \quad m = \frac{E}{c^2}$$

where c represents the speed of light. Einstein's **mass–energy equivalence relationship** implies that whenever a nuclear reaction results in the release of energy there is a corresponding decrease in mass.

For example, when one kilogram of uranium-235 undergoes fission in a nuclear reactor, the energy released is approximately 8×10^{13} joules. This corresponds to a small but measurable decrease in mass:

$$m = \frac{E}{c^2} = \frac{8 \times 10^{13}}{(3.00 \times 10^8)^2} \approx 9 \times 10^{-4}\,\text{kg}$$

Since mass and energy are equivalent, the unified atomic mass unit (u) defined in the previous section can also be expressed in units of energy:

$$E = mc^2$$
$$1\,\text{u} \times c^2 = 1.660\,539\,2 \times 10^{-27} \times (2.99\,79\,264 \times 10^8)^2$$
$$= 1.492\,4182 \times 10^{-10}\,\text{J}$$

An alternative unit of energy is the electronvolt (eV – see Chapter 13), which is equal to 1.6×10^{-19} J. Therefore, the energy equivalent to 1 u of matter can be expressed in electronvolts as:

$$1\,\text{u} = \frac{1.492\,4182 \times 10^{-10}}{1.6021765 \times 10^{-19}}$$
$$= 9.315 \times 10^8\,\text{eV}\,c^{-2} = 931.5\,\text{MeV}\,c^{-2} \text{ (rounded to 4 significant figures)}$$

Since mass and energy are related by $m = E/c^2$, the mass of a particle can be given in units of $\text{MeV}\,c^{-2}$. If $1\,\text{MeV}\,c^{-2}$ worth of mass is converted to energy, then 1 MeV of energy is produced. A related unit is the $\text{GeV}\,c^{-2}$, where $1\,\text{GeV} = 1.6 \times 10^{-10}\,\text{J}$. In general, an energy of x MeV is equivalent to mass of $x\,\text{MeV}\,c^{-2}$.

30 Use Einstein's mass–energy equivalence relationship to calculate the energy equivalent of 500 grams of matter.

31 An atom loses a mass of 2.2×10^{-30} kg after a nuclear reaction. Determine the energy obtained in joules due to this loss of nuclear mass.

32 Calculate the increase in mass when 1.00 kg of water absorbs 4.20×10^4 J of energy to produce a temperature rise of 10.0°C.

33 The mass decrease for the decay of one radium atom is 8.5×10^{-30} kg. Calculate the energy equivalent in MeV.

Mass defect and binding energy

7.3.5 **Define** the concepts of *mass defect, binding energy* and *binding energy per nucleon.*

The relative atomic mass of carbon-12 is 12.000000 u. The nucleus of this atom is composed of six protons and six neutrons. However, the sum of the separate masses of the sub-atomic particles is greater than 12.000000 u, as we can see by using the mass values for the proton and neutron given in the previous section:

$$\underbrace{(6 \times 1.007276\,\text{u})}_{\text{mass of 6 protons}} + \underbrace{(6 \times 1.008665\,\text{u})}_{\text{mass of 6 neutrons}} = 12.095646\,\text{u}$$

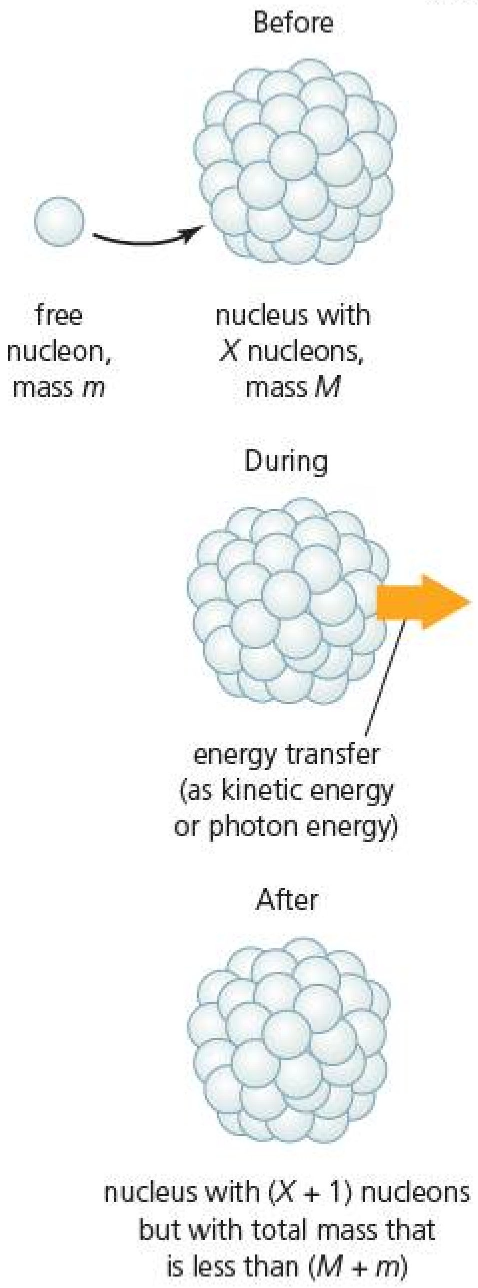


Figure 7.36 The effect of adding a free nucleon to a nucleus

The mass of the separate nucleons is more than the mass of the nucleus itself. Adding an extra nucleon to a nucleus does not increase the total mass by the mass of the nucleon. Instead, the mass of the new nucleus is less than the combined mass of the nucleon and the original nucleus (Figure 7.36).

In the process of combining the nucleon and the nucleus, the total mass decreases and energy is released. This is usually in the form of either kinetic energy of the particle or particles involved, or gamma ray (photon) energy.

If we could bring together six protons and six neutrons (in a particle physics thought experiment), then not only would the total mass decrease, but an equivalent amount of energy would be released. The 'disappeared' mass is termed the **mass defect** of the nucleus.

The mass defect is the difference in mass between the mass of the nucleus and the masses of the individual protons and neutrons.

This amount of mass is released as energy, as predicted by Einstein's mass–energy equivalence relationship: $E = mc^2$.

The **binding energy** of the nucleus is the change in potential energy of the nucleons when they form a nucleus. Since energy is transferred away from the nucleus, it loses energy. The binding energy of nucleons is a *negative* quantity of potential energy, given by:

$$\text{binding energy} = \text{mass defect of nucleus} \times c^2$$

If the carbon-12 nucleus were to break apart into 12 separate nucleons, it would have to gain an amount of energy equal in size to its binding energy. Unless the nucleus receives that energy, its nucleons remain bound together, hence the name 'binding energy'.

The term binding energy can also be defined as the energy needed to separate a nucleus into its individual protons and neutrons.

Defined in this way it is a *positive* quantity (Figure 7.37) and is the energy equivalent of the mass defect ($E = mc^2$).

Nucleons lose energy when they form a nucleus. Their potential energy is negative compared with the zero potential energy they have when they are free. Until they regain this energy, they are trapped. The energy they collectively need to escape from each together is equal to the binding energy.

Protons have to be forced close together when a nucleus is formed. Work must be done to overcome the Coulomb repulsion between them, before the strong nuclear force (see Section 7.1), which operates only over small distances, pulls the nucleons together and a nucleus forms.

The protons are said to 'climb over' the 'hill' that surrounds the 'hole' (Figure 7.38) and then 'fall' into the nucleus. 'Holes' and 'hills' are useful analogies to describe the potential energies that nucleons possess at different distances from the nucleus.

Neutrons do not experience electrical repulsion and so there is no 'hill' for them to climb (Figure 7.38). So, in general, nuclei can absorb neutrons more easily. The absorption and capture of neutrons by uranium-235 nuclei is the basis for the nuclear fission reactor (Chapter 8).

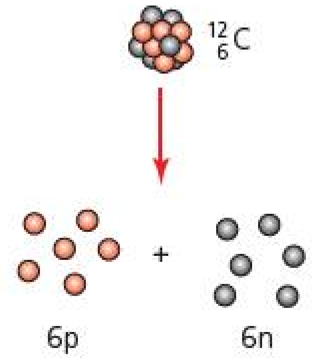


Figure 7.37 Binding energy (positive quantity) must be supplied to separate these carbon-12 nucleons

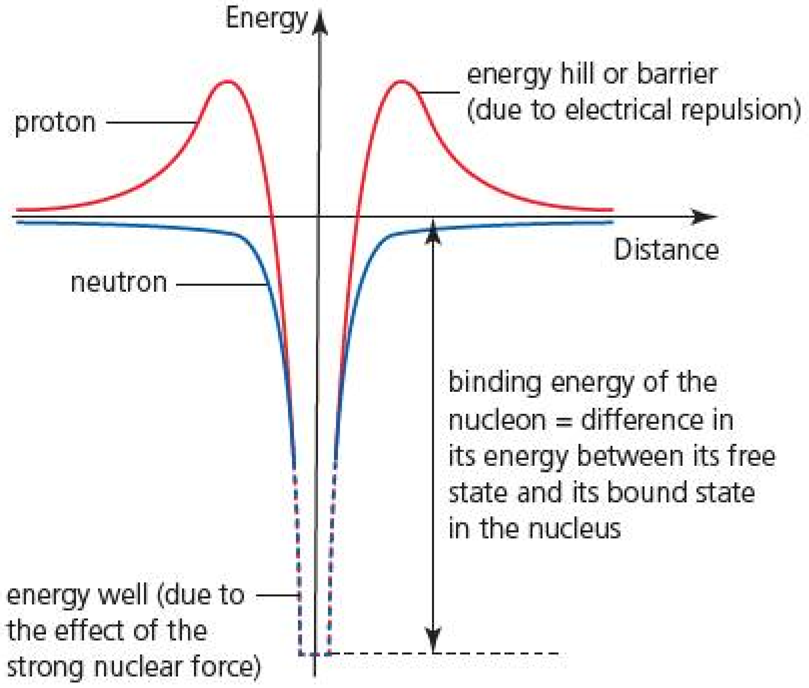


Figure 7.38 Graph showing how potential energy changes with distance from the centre of a nucleus for a proton (red line) and a neutron (blue line)

Worked example

9 Calculate the mass defect (in electronvolts) and binding energy of a helium atom (4.00260 u). It consists of two protons (each of mass 1.007276 u), two neutrons (each of mass 1.008665 u) and two electrons (each of mass 0.000549 u). 1 u = 931.5 MeV c^{-2}.

The total mass of the individual particles = $(2 \times 1.007276) + (2 \times 1.008665) + (2 \times 0.000549) =$ 4.03298 u

Mass defect = 4.03298 − 4.00260 = 0.03038 u

$E = mc^2 = 0.03038 \times 931.5 = 28.30$ MeV

A large nucleus has more binding energy than a smaller nucleus simply because it is formed from a large number of nucleons. However, a more useful measure of the stability of a nucleus than its binding energy is its **binding energy per nucleon**. This is defined as the average energy per nucleon needed to separate a nucleus into its individual nucleons (this definition results in positive quantities):

$$\text{binding energy per nucleon} = \frac{\text{total binding energy}}{\text{nucleon number}}$$

For example, the binding energy per nucleon of a carbon-12 nucleus is its total binding energy divided by 12.

Variation in binding energy per nucleon

7.3.6 Draw and **annotate** a graph showing the variation with nucleon number of the binding energy per nucleon.

A plot of binding energy per nucleon against size of nucleus (number of nucleons) shows that iron nuclei, and others of similar 'medium size', have the most binding energy per nucleon. Nucleons in mid-sized nuclei of the **binding energy curve** (Figure 7.39) are the most stable. Stable nuclei have a low probability of radioactive decay.

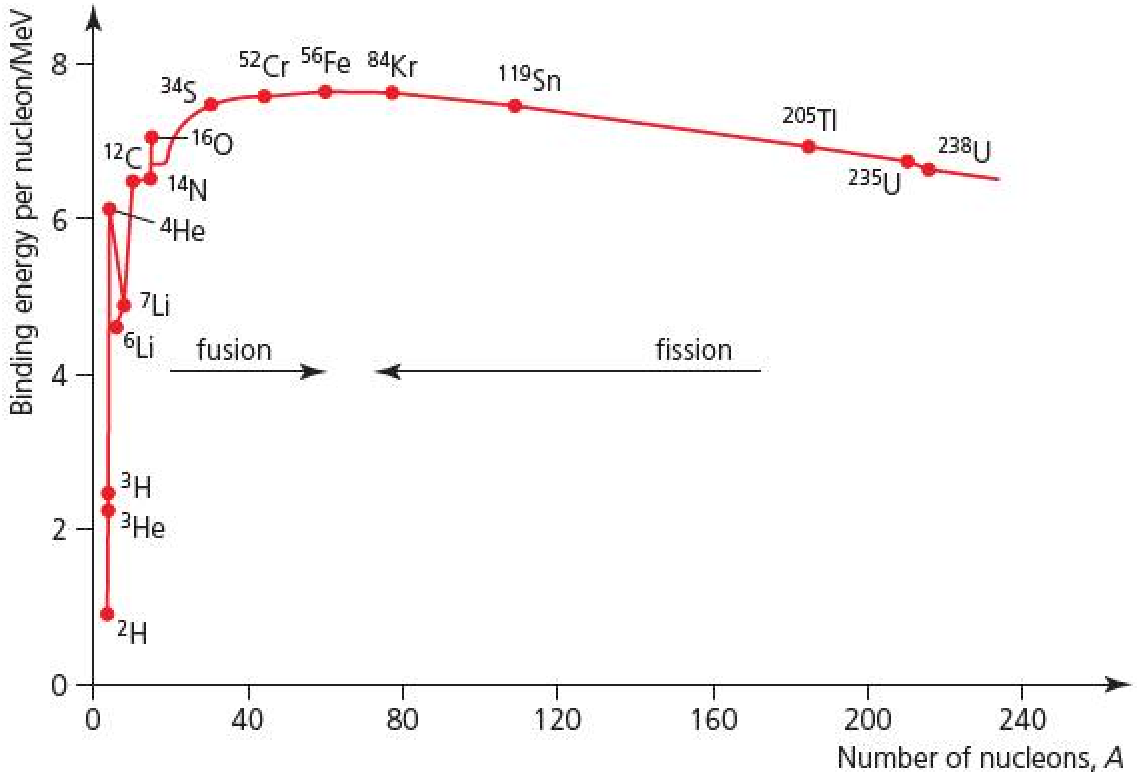


Figure 7.39 A plot of binding energy per nucleon against number of nucleons

Worked examples

10 The binding energy of a helium-4 nucleus is 28.4 MeV. Calculate the binding energy per nucleon.

The helium-4 nucleus has four nucleons. The binding energy per nucleon is therefore $28.4 \div 4 = 7.10$ MeV per nucleon.

11 Determine the binding energy per nucleon of the nucleus of $^{12}_{6}$C ($m_p = 1.007276$ u; $m_n = 1.008665$ u; 1 u = 931.5 MeV)

Total rest mass of individual protons and neutrons $= (6 \times 1.007276\,\text{u}) + (6 \times 1.008665\,\text{u})$
$= 12.095646\,\text{u}$

Mass defect $= 12.095646\,\text{u} - 12.000000\,\text{u} = 0.0956465\,\text{u}$

Binding energy per nucleon $= \dfrac{0.095646 \times 931.5}{12} = 7.425$ MeV per nucleon

7.3.7 Solve Problems involving mass defect and binding energy.

34 The binding energy of carbon-14 nucleus is 102 MeV. Calculate the binding energy per nucleon.

35 Calculate the binding energy, in MeV, of a nitrogen-14 nucleus with a mass defect of 0.108517 u.

Fission and fusion

Nuclear fission

7.3.8 Describe the processes of nuclear fission and nuclear fusion.

Nuclear fission is the splitting of a heavy nucleus into two lighter nuclei accompanied by the release of large amounts of energy. Energy is released because the products are more stable and have less potential energy than the original nucleus. The average binding energy per nucleon of the daughter nuclei is greater than that of the parent nuclide (see below).

A fission reaction occurs after bombardment of uranium-235 nuclei by slow-moving neutrons. This can result in the capture of a neutron and the formation of uranium-236. This is an unstable nuclide and readily undergoes fission.

One common fission reaction that occurs is:

$$^{1}_{0}n + ^{235}_{92}U \rightarrow ^{236}_{92}U \rightarrow ^{141}_{56}Ba + ^{92}_{36}Kr + 3^{1}_{0}n + \text{photons}$$

The neutrons released may be absorbed by other uranium-235 nuclei and induce additional fission reactions. Nuclear fission reactors make use of the *controlled* fission of uranium-235 (see Chapter 8). Atomic bombs make use of the *uncontrolled* fission of uranium-235.

Nuclear fusion

Nuclear fusion is the combination of two light nuclei to produce a heavier nucleus. Energy is released by the process.

An example is the fusion of two hydrogen-2 nuclei to produce a helium-3 nucleus:

$$^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{2}He + ^{1}_{0}n$$

Reactions of this type involving the conversion of hydrogen to helium are the source of the Sun's energy. Temperatures in the region of 10^8 kelvin are required to provide the fusing nuclei with sufficient kinetic energy to overcome their mutual repulsion. This has been achieved in an uncontrolled manner, in the hydrogen bomb. The high temperature required for the fusion reaction is provided by the explosion of an atom bomb. Experimental nuclear fusion reactors make use of controlled nuclear fusion reactions (Chapter 8).

36 One possible reaction taking in the core of a nuclear reactor is:

$$^{235}_{92}U + ^{1}_{0}n \rightarrow ^{95}_{42}Mo + ^{139}_{57}La + 2^{1}_{0}n + 7\,^{0}_{-1}e$$

(Masses: uranium-235, 235.123 u; molybdenum-95, 94.945 u; lanthanum-139, 138.955 u; neutron, 1.009 u. 0.235 kg of uranium-235 contains 6×10^{23} atoms; 1 u = 931.5 MeV.)

For this fission reaction, calculate:

a the mass (in atomic mass units) on each side of the equation (ignore mass of electrons)
b the mass defect (in atomic mass units)
c the energy released per fission of a uranium-235 nucleus (in MeV)
d the energy available from the fission of 0.100 g of uranium-235 (in joules).

37 Two fusion reactions which take place in the Sun are described below.

- A hydrogen-2 nucleus absorbs a proton to form a helium-3 nucleus.
- Two helium-3 nuclei fuse to form a helium-4 nucleus plus two free protons.

For each nuclear reaction, write down the appropriate nuclear equation and calculate the energy released in MeV.

(Masses: hydrogen-2, 2.01410 u; helium-3, 3.01605 u; helium-4, 4.00260 u; proton, 1.00728 u and neutron, 1.00867 u; 1 u = 931.5 MeV).

Fission, fusion and changes in binding energy per nucleon

7.3.9 Apply the graph in 7.3.6 to account for the energy release in the processes of fission and fusion.

When a large nucleus breaks into mid-sized pieces through nuclear fission, the binding energy per nucleon increases (Figure 7.40a). The increase in binding energy occurs because the nucleons are more tightly bound in the two daughter nuclei and involves the release of energy.

For example, in the fission of a nucleus of uranium-236 into barium-141 and krypton-92 (shown at the top of this page), the nucleons are more tightly bound in the daughter nuclei than in the parent nucleus. The binding energy per nucleon has increased in size.

In lithium-6, the binding energy per nucleon is especially small. This means that breaking a lithium-6 nucleus apart is much easier, nucleon for nucleon, than breaking an iron nucleus apart. It also means that if two lithium-6 nuclei were to fuse together to form a carbon-12 nucleus, then the binding energy per nucleon would greatly increase in size (Figure 7.40b) and a relatively large amount of energy would be released.

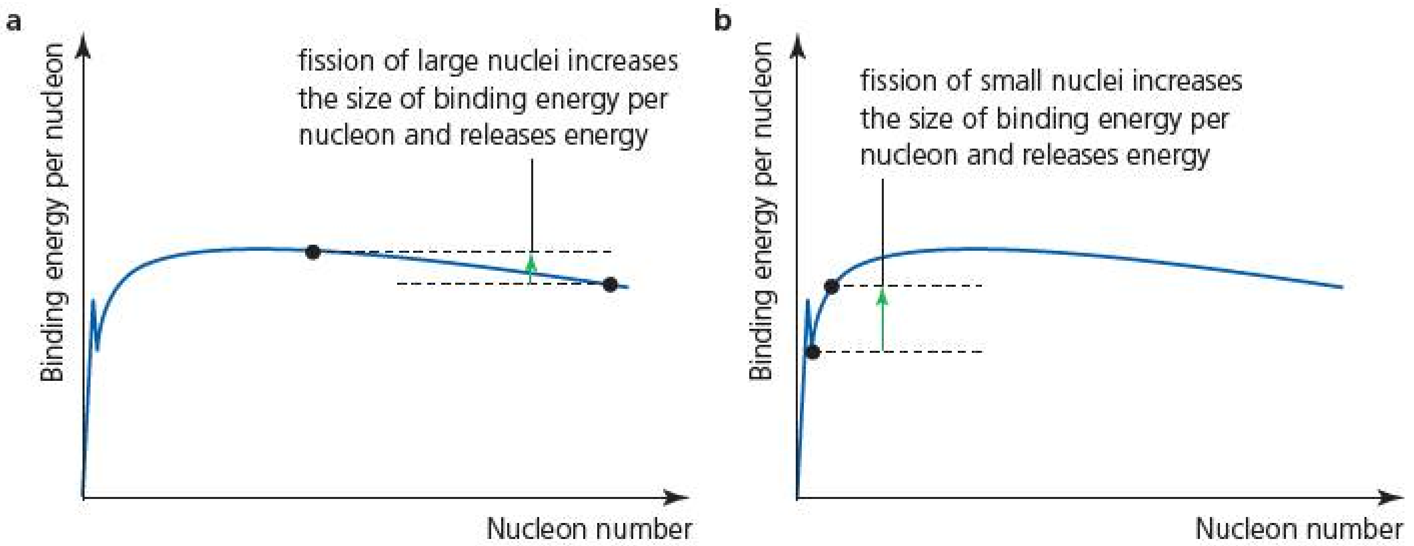


Figure 7.40 Any nuclear process that increases the size of binding energy per nucleon releases energy

This is an example of nuclear fusion of two smaller nuclei into a larger one:

$${}^{6}_{3}\text{Li} + {}^{6}_{3}\text{Li} \rightarrow {}^{12}_{6}\text{C}$$

Fission of large nuclei, like the fusion of small nuclei, makes energy available as kinetic energy, carried by the particles themselves, and as energy that is carried away by photons. These are the energies that form available 'nuclear energy'.

38 Calculate the amount of energy in MeV (to the nearest integer) released when lithium-6 nuclei undergo fusion to form a carbon-12 nucleus. The binding energies per nucleon of lithium-6 and carbon-12 are 5.076 MeV and 7.424 MeV.

39 Determine the amount of energy (in MeV) released if the following nuclear fusion reaction occurs:

$${}^{2}_{1}\text{H} + {}^{2}_{1}\text{H} \rightarrow {}^{3}_{1}\text{H} + {}^{1}_{1}\text{H}$$

The binding energies per nucleon of ${}^{2}_{1}\text{H}$ and ${}^{3}_{1}\text{H}$ are 1.11 and 2.83 MeV, respectively.

Stellar fusion

7.3.10 State that nuclear fusion is the main source of the Sun's energy.

Nuclear fusion is the source of the energy for the Sun and all other stars.

Inside the Sun, nuclei are so close together and have such large amounts of kinetic energy that they can overcome their Coulomb repulsion and get sufficiently close for the strong nuclear force to become important. The most important fusion processes are those involving the smaller nuclei – mostly isotopes of hydrogen. The excess energy is carried away by photons.

A typical fusion reaction involving hydrogen is:

$${}^{3}_{1}\text{H} + {}^{2}_{1}\text{H} \rightarrow {}^{4}_{2}\text{He} + {}^{1}_{0}\text{n}$$

Various fusion processes occur during stars' lifetimes, releasing energy until nuclei as large as iron are synthesized. However, if pairs of iron nuclei (or other larger nuclei) are to fuse together, then energy has to be absorbed from the surroundings (see Option E).

Additional Perspectives

Thermonuclear fusion

The use of high temperatures to give small nuclei the energy required to fuse is called thermonuclear fusion. To calculate the temperature required for two deuterium $({}^{2}_{1}\text{H})$ nuclei to fuse we assume that each nucleus has a potential energy barrier equal to the binding energy per nucleon ($\approx 1\,\text{MeV}$).

To calculate the temperature required for the nuclei to travel at this speed we assume that the nuclei behave as an ideal gas and apply the Boltzmann equation:

$E_K = \frac{3}{2}kT$, where k is the Boltzmann constant and T is the absolute temperature.

To use this expression we need the value of the average kinetic energy of the nuclei (E_K) expressed in joules. $1\,\text{MeV} = 1.6 \times 10^{-13}\,\text{J}$, so substituting this value for E_K gives:

$$1.6 \times 10^{-13}\,\text{J} = \frac{3}{2} \times 1.38 \times 10^{-23}\,\text{J K}^{-1} \times T$$

$$T = \frac{2 \times 1.6 \times 10^{-13}}{3 \times 1.38 \times 10^{-23}} = 7.7 \times 10^{9}\,\text{K}$$

Questions

1 All of the carbon atoms in your body were made by the triple-alpha process. Summarize and describe this process.

2 Find out how iron is formed in stars. Why is there more iron in old stars than young stars?

Reaction energy

7.3.11 Solve problems involving fission and fusion reactions.

Consider the following generalized representation of a nuclear reaction:

$$\text{X} + \text{x} \rightarrow \text{Y} + \text{y} + Q$$

The sum of masses of X and x may be greater than the sum of masses Y and y. Some nuclear mass is converted into energy Q. The energy may appear as the kinetic energy of the product particles and energy of any gamma rays (photons) produced.

The value of Q can be obtained by using the law of conservation of mass and Einstein's mass–energy equivalence relationship:

$$M_X + m_x = M_Y + m_y + \Delta m$$

where M_X and m_x represent the masses of X and x, M_Y and m_y are the masses of Y and y, and Δm is the mass–energy equivalence of the reaction energy Q.

If $M_X + m_x > M_Y + m_y$ then Q is positive and energy is released. However, if $M_X + m_x < M_Y + m_y$ then Q is negative and energy is absorbed: the reaction can only happen if a certain minimum amount of energy is supplied, otherwise the reaction cannot occur.

Worked example

12 A proton collides with a tritium nucleus and the following nuclear reaction occurs:

$$^{3}_{1}\text{H} + ^{1}_{1}\text{H} \rightarrow ^{3}_{2}\text{He} + ^{1}_{0}\text{n}$$

Calculate the amount of energy transferred in the reaction in MeV.

(Masses: $^{3}_{1}$H, 3.016050 u; $^{1}_{1}$H, 1.007276 u; $^{3}_{2}$He, 3.016030 u; $^{1}_{0}$n, 1.008665 u; 1 u = 932 MeV.)

Total mass before reaction, $M_{before} = 4.023326\,\text{u}$; total mass after reaction, $M_{after} = 4.024695\,\text{u}$; $M_{before} = M_{after} + \Delta m$; $\Delta m = -0.001369\,\text{u}$; $Q = -0.001369\,\text{u} \times 932 = -1.276\,\text{MeV}$, hence the reaction needs energy to be supplied.

40 Uranium-238 decays according to the following equation:

$$^{238}_{92}\text{U} \rightarrow ^{234}_{90}\text{Th} + ^{4}_{2}\text{He}$$

The alpha particle has a kinetic energy of 4.20 MeV. Determine the kinetic energy of the thorium nuclide. (Masses: $^{238}_{92}$U, 238.0508 u; $^{234}_{90}$Th, 234.0436 u; $^{4}_{2}$He, 4.0026 u; 1 u = 931.5 MeV)

41 Find out about the proton–proton chain reaction (proton–proton fusion reaction) that occurs in the Sun. Write appropriate nuclear equations and describe the various steps.

SUMMARY OF KNOWLEDGE

7.1 The atom

- A chemical element is a substance that cannot be broken down into simpler chemicals. Elements are made of atoms.
- An ion is formed when an atom loses or gains one or more electrons.
- The nucleus of an atom contains neutrons and protons (nucleons); the electrons surround the nucleus and have different possible energy levels.
- Neutrons are electrically neutral; a proton carries a positive charge equal in magnitude to the negative charge on an electron.
- The electrons have negligible (very small) mass relative to protons and neutrons.
- The mass of an atom is concentrated in the nucleus; most of an atom is empty space.
- The negative electrons and the positive protons in the nucleus exert an electrostatic attraction on each other.
- The Geiger–Marsden experiment involved directing a beam of alpha particles (helium nuclei) onto a thin gold foil.
- Some alpha particles were scattered through large angles, suggesting that an atom has a small, positively charged region (the nucleus). Only a few alpha particles were scattered through large angles, suggesting that the nucleus is very much smaller than the atom as a whole.
- The simple nuclear model of an atom based only on electrostatics is incorrect: accelerating particles radiate high-frequency electromagnetic radiation. The electrons would rapidly lose energy and spiral into the nucleus.
- The simple nuclear model does not explain the stability of many atoms. It does not make any prediction about electronic structure, that is, the arrangement and relative energies of electrons.
- An emission spectrum is produced when excited atoms (or molecules) emit electromagnetic radiation. It consists of a series of coloured lines, on a black background, which become closer as the wavelength decreases (or frequency increases).
- Excited gaseous atoms produce line spectra – the line spectrum of an element is characteristic of that element.
- An absorption spectrum is produced when electromagnetic radiation is absorbed by gaseous atoms or molecules at low pressure. It consists of a series of black lines on a background of the visible spectrum.
- Continuous spectra are produced by hot objects and consist of a continuous range of wavelengths – sunlight and light from a filament bulb both produce continuous spectra.
- A simple model of the hydrogen atom consists of a series of energy levels that the electron may occupy. Transitions between these energy levels produce the characteristic spectrum of the hydrogen atom.
- Excitation involves the absorption of energy (electrical or thermal) by an electron transition to a higher energy level.
- A photon is emitted when the excited electron returns to the ground state.
- Light (and other electromagnetic radiation) behaves as a stream of tiny energy 'packets' called photons.
- The energy of a photon is given by the following relationship: $E = hf$, where E is the energy of the quantum, f is the frequency and h is Planck's constant.
- Nuclides are specific isotopes of atoms described by the following notation: ${}^{A}_{Z}X$, where A is the nucleon number, Z is the proton number and X is the symbol of the element.
- The proton number Z is the number of protons in the nucleus of an atom. Atoms which have the same number of protons are atoms of the same element.
- The nucleon number A is the sum of the number of neutrons and the number of protons.
- The number of neutrons in the nucleus of an atom, N, is given by $A – Z$.
- Two atoms which have the same proton number, but different numbers of neutrons are isotopes of each other. Isotopes have the same proton number, but different nucleon numbers. They have identical chemical properties, but slightly different physical properties.

- There are two main interactions within a nucleus: electrostatic repulsion between protons (Coulomb repulsion) and a short-range attractive force known as the strong nuclear force which operates between neighbouring nucleons.

7.2 Radioactive decay

- An unstable nucleus emits radiation (alpha or beta particles, or gamma rays) which carry energy away from the nucleus, thus making the nucleus more stable.
- Nuclear decay is a spontaneous and random process. This unpredictability means that the emission of radiation tends to fluctuate and average measurements are recorded.
- Alpha particles $\left({}^4_2\alpha \text{ or } {}^4_2\text{He}\right)$ are identical to helium nuclei.
- Beta particles $\left({}^{\ 0}_{-1}\beta \text{ or } {}^{\ 0}_{-1}\text{e}\right)$ are fast-moving electrons.
- Gamma rays $\left({}^0_0\gamma \text{ or } \gamma\right)$ are very short wavelength (high-frequency) electromagnetic radiation.
- The decaying nucleus which releases radiation is termed the 'parent nucleus'; the resulting nucleus is termed the 'daughter' nucleus. Decay processes are described using nuclear equations.
- During alpha decay a new element is formed:

$$\underset{\text{parent nuclide}}{{}^A_Z\text{X}} \rightarrow \underset{\text{daughter nuclide}}{{}^{A-4}_{Z-2}\text{X}} + \underset{\text{alpha particle}}{{}^4_2\text{He}} + \text{energy}$$

- Alpha decay reduces the nucleon number of the parent nuclide by four, and reduces the proton number by two.
- During beta-negative decay a new element is formed:

$$\underset{\text{parent nuclide}}{{}^A_Z\text{X}} \rightarrow \underset{\text{daughter nuclide}}{{}^{\ \ A}_{Z+1}\text{X}} + \underset{\text{beta particle}}{{}^{\ 0}_{-1}\beta} + \underset{\text{antineutrino}}{\bar{\nu}_e} + \text{energy}$$

- Beta-negative decay causes no change to the nucleon number of the parent nuclide, and increases the proton number by one.
- During gamma decay no new element is formed: there are no changes to the proton or nucleon number of the parent nuclide. Gamma rays are released from an excited nucleus.
- A nucleus may undergo a series of radioactive decays, producing a series of different elements by the emission of alpha, beta and gamma radiations.
- Alpha and beta particles may be deflected by magnetic and electric fields. Alpha and beta particles are deflected in opposite directions when passed through a magnetic field.
- Gamma rays are unaffected by magnetic and electric fields. Gamma rays are the most penetrating of the ionizing or nuclear radiations. Alpha particles are the most ionizing and travel at the lowest velocity.
- The cells of the body may undergo physical and chemical changes as a result of exposure to ionizing radiation. Physical changes involve burning; chemical changes may involve damage to DNA molecules (mutation), which can lead to cancer.
- Some nuclei are unstable and undergo radioactive decay. The forces of repulsion are greater than the attractive forces (due to the strong nuclear force).
- Some nuclei are stable and do not undergo radioactive decay. In these nuclei the forces of repulsion are less than the attractive forces (due to the strong nuclear force).
- Radioactivity is a random process – it is not possible to predict when a particular nuclide will decay.
- Radioactivity is a spontaneous process which means a particular nuclide will undergo decay independently of what is happening outside that nuclide.
- Radioactive decay is not affected by chemical reactions or by any changes in physical conditions, such as temperature and pressure.
- For a large number of decaying nuclides radioactive decay is an exponential process with a characteristic half-life.
- The half-life of a radioactive isotope is the time it takes for the level of radioactivity to halve.

- If N is the initial number of radioactive nuclides, then after a period of one half-life $T_{1/2}$ only $\frac{1}{2}N$ radioactive nuclides are left in the sample. After two half-lives $2T_{1/2}$, then there will be only $\frac{1}{4}N$ radioactive nuclides left in the sample, etc.
- The half-life can be determined graphically from a decay curve graph which plots count rate against time.
- The radioactive half-life is independent of the amount of the isotope. Activity is the rate of decay, i.e. the number of disintegrations per second. One becquerel, Bq, is one disintegration per second.
- Investigations to measure radioactivity must take into account the background radiation, ionizing radiation from natural and artificial sources that is always there.

7.3 Nuclear reactions, fission and fusion

- A transmutation is a nuclear reaction involving the formation of a new nuclide.
- Artificial (induced) transmutations occur when nuclides capture one or more nucleons.
- The unified atomic mass unit (u) is defined such that the mass of $^{12}_{6}C$ is 12 u (exactly). $1\,u = 1.661 \times 10^{-27}\,kg = 931.5\,MeV\,c^{-2}$.
- The mass of an atom expressed in unified atomic mass units is numerically equal to its relative atomic mass.
- Einstein's mass–energy equivalence relationship, $E = mc^2$, shows that mass and energy are related. A loss in mass during a nuclear reaction will appear as energy.
- The mass defect of a nucleus is the difference between the total mass of the separate nucleons and the mass of the nucleus.
- The binding energy of a nucleus is the energy needed to separate completely all constituent nucleons.
- The binding energy per nucleon is a measure of the relative stability of a nucleus: a high binding energy per nucleon means the nucleus is stable.
- Binding energy per nucleon rises rapidly with nucleon number to a maximum at iron and then slowly decreases with nucleon number.
- A nuclear process is described by a nuclear equation. The total of nucleon and proton numbers must be the same on both sides of the nuclear equation.
- Nuclear fusion occurs when two or more nuclides join together to make a new nuclide. Energy is released due to an increase in binding energy per nucleon.
- Fusion reactions occur inside stars and are responsible for generating new elements from hydrogen and helium.
- Nuclear fusion can take place inside the Sun and other stars because the temperature is extremely high and the particle density is very high.
- Fission reactions occur when a nucleus breaks apart, producing new nuclides.
- Uranium-235 undergoes fission when bombarded with slow-moving neutrons. Very large quantities of energy are released as the reaction involves an increase in binding energy per nucleon.
- Fission reactions occur inside nuclear reactors and nuclear bombs.

Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 Why are gamma rays not deflected by a strong magnetic field?

A They have no mass.
B They are weakly ionizing.
C They are strongly penetrating.
D They are electrically neutral.

Q2 A freshly prepared sample contains 40 μg of the isotope iodine-131. The half-life of iodine-131 is 8 days.
Which of the following is the best estimate for the mass of the iodine-131 remaining after 24 days?

A 10 μg
B 13 μg
C 5 μg
D zero

Q3 Two neutrons are captured by a nucleus. Which of the following gives the changes in the atomic (proton) number and mass (nucleon) number of the nucleus?

	Atomic number	Mass number
A	unchanged	unchanged
B	unchanged	increases by 2
C	increases by 2	unchanged
D	increases by 2	increases by 2

Q4 Which set of radioactive emissions corresponds to the description given in the table headings?

	High energy photons	High speed helium nuclei	High speed electrons
A	gamma	beta	alpha
B	beta	gamma	alpha
C	gamma	alpha	beta
D	alpha	gamma	beta

Q5 Isotopes of a given element all have the same

A charge-to-mass ratio
B number of neutrons in the nucleus
C number of protons and neutrons in the nucleus (nucleon or mass number)
D number of protons.

Q6 What is particle X in the fusion reaction shown below?

$$^{7}_{3}\text{Li} + ^{2}_{1}\text{H} \rightarrow 2^{4}_{2}\text{He} + \text{X}$$

A a proton
B an electron (beta particle)
C an alpha particle
D a neutron

Q7 The binding energy of a helium-3 nucleus is defined to be the

A energy released when a helium-3 nucleus is formed from its individual constituents
B energy released when the helium-3 nucleus is separated into its individual constituents
C total energy of the protons inside the helium-3 nucleus
D total energy of the helium-3 nucleus.

Q8 A nuclear reaction is represented by the following equation.

$$^{1}_{0}\text{n} + ^{198}_{80}\text{Hg} \rightarrow ^{197}_{79}\text{Au} + ^{2}_{1}\text{H}$$

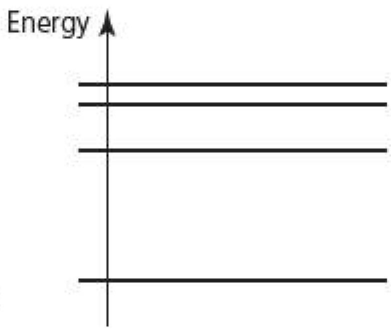
This reaction is an example of

A nuclear fusion
B nuclear fission
C artificial (induced) transmutation
D natural transmutation.

Q9 Which of the following identifies the important interaction(s) between protons and neutrons inside the nucleus of an atom?

A Coulomb only
B nuclear only
C nuclear and Coulomb
D gravitational, nuclear and Coulomb

Q10 The diagram shows four possible electron energy levels in the hydrogen atom.



The number of different frequencies in the emission spectrum of atomic hydrogen that arise from electron transitions between these levels is

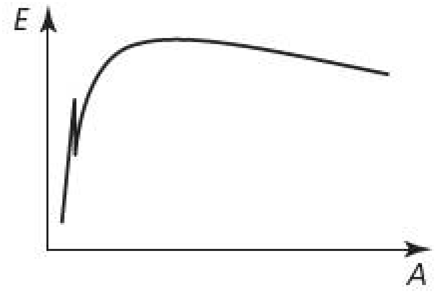
A 0
B 2
C 4
D 6

Paper 2 IB questions and IB style questions

Q1 a i Describe the phenomenon of natural radioactive decay. [3]

ii Ionizing radiation is emitted during radioactive decay. Explain what is meant by the term *ionizing*. [2]

b The sketch graph below shows the variation with mass number (nucleon number) A of the binding energy per nucleon E of nuclei.



One possible nuclear reaction that occurs when uranium-235 is bombarded by a neutron to form xenon-142 and strontium-90 is represented as

$$^{235}_{92}U + ^{1}_{0}n \rightarrow ^{142}_{54}Xe + ^{90}_{38}Sr + 4^{1}_{0}n$$

i Identify the type of nuclear reaction represented above. [1]

ii Copy the sketch graph above. On your copy, identify with their symbols the approximate positions of the uranium (U), the xenon (Xe) and the strontium (Sr) nuclei. [2]

iii Data for the binding energies of xenon-142 and strontium-90 are given below.

Isotope	Binding energy/MeV
Xenon-142	1189
Strontium-90	784.8

The total energy released during the reaction is 187.9 MeV. Determine the binding energy per nucleon of uranium-235.

iv State why the binding energy of the neutrons formed in the reaction is not quoted. [1]

Standard Level Paper 2, May 08 TZ2, QB3 (Part 2)

Q2 A stationary radon-220 ($^{220}_{86}Rn$) nucleus undergoes α-decay to form a nucleus of polonium (Po). The α-particle has kinetic energy of 6.29 MeV.

a i Complete the nuclear equation for this decay [2]

$$^{220}_{86}Rn \rightarrow Po +$$

ii Calculate the kinetic energy, in joules, of the α-particle. [2]

iii Deduce that the speed of the α-particle is $1.74 \times 10^7\,m\,s^{-1}$. [1]

The diagram below shows the α-particle and the polonium nucleus immediately after the decay. The direction of the velocity of the α-particle is indicated.

b i On a copy of the diagram above, draw an arrow to show the initial direction of motion of the polonium nucleus immediately after the decay. [1]

ii Determine the speed of the polonium nucleus immediately after the decay. [3]

iii In the decay of another radon nucleus, the nucleus is moving before the decay. Without any further calculation, suggest the effect, if any, of this initial speed on the paths shown in **b i**. [2]

The half-life of the decay of radon-220 is 55 s.

c i Explain why it is not possible to state a time for the life of a radon-220 nucleus. [2]

ii Define *half-life*. [2]

Standard Level Paper 2, May 05 TZ2, QB1

8 Energy, power and climate change

STARTING POINTS

- The Earth is surrounded by an atmosphere containing nitrogen, oxygen and small percentages of many other gases.
- Burning (non-nuclear) fuels releases carbon dioxide, water vapour and various pollutants into the atmosphere.
- Thermal energy is the transfer of energy from a hotter place to a colder place.
- If the temperature of an object remains constant, it is said to be in thermal equilibrium with its surroundings.
- The three principal means of thermal energy transfer are conduction, convection and radiation.
- Thermal radiation (infrared) is part of the electromagnetic spectrum.
- Heat engines can do useful work when significant amounts of thermal energy flow from hot to cold.
- The total energy of an isolated system remains constant.
- The temperature rise caused by a supply of thermal energy can be calculated from $Q = mc\Delta T$.
- When a mass falls from rest, gravitational potential energy (mgh) is transferred to kinetic energy $\left(\frac{1}{2}mv^2\right)$.
- All macroscopic processes dissipate useful energy into the surroundings.
- Power = energy/time.
- Efficiency = input power/output power.
- Density = mass/volume.
- Nuclear reactions, like fission and fusion, involve very large amounts of energy for small particles.
- Nuclear reactions can be represented by balanced equations, using standard notations for the nuclides.
- The energies of particles are often given in the unit of electronvolts.
- Isotopes have identical chemical properties.

8.1 Energy degradation and power generation

Power stations

8.1.4 **Outline** the principal mechanisms involved in the production of electrical power.

Most power stations which produce electricity transfer the energy from burning fossil fuels or from fission reactions in nuclear fuels. The thermal energy is used to raise the temperature of water in a boiler and turn it into high pressure steam. The steam causes the rotation of turbines, shown in Figure 8.1, which are connected to coils of wire. The process of **electromagnetic induction** produces electrical energy as the coils rotate in strong magnetic fields. When the steam exits the turbine it is cooled, causing it to condense, and the water is pumped back into the boiler. Figure 8.2 shows a simplified representation of a fossil fuel power station.

The same basic ideas apply to *all* power stations which transfer energy from a fuel to electricity using the flow of thermal energy. Machines which are powered by a significant flow of thermal energy are called **heat engines**. Examples include power stations, cars, planes, trains, etc.

Figure 8.1 A turbine used for generation of electricity

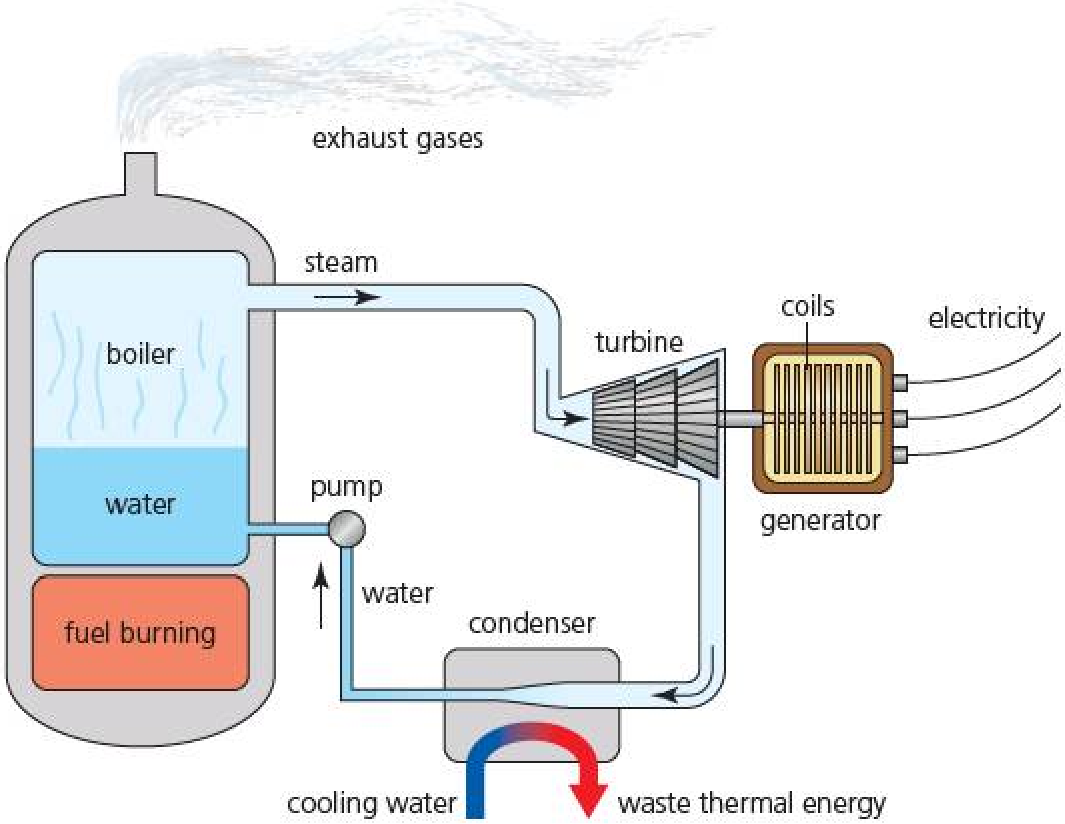


Figure 8.2 A schematic diagram of a fossil fuel power station

Energy conversions and energy degradation

8.1.1 **State** that thermal energy may be completely converted to work in a single process, but the continuous conversion of this energy into work requires a cyclical process and the transfer of some energy from the system.
8.1.2 **Explain** what is meant by degraded energy.

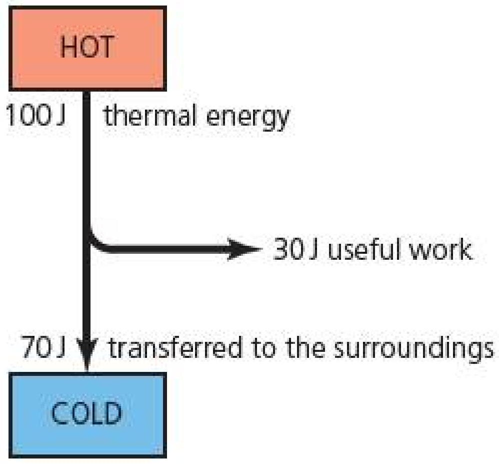


Figure 8.3 Energy flow between 'hot' and 'cold'

Heat engines use the principle that when there is a temperature difference between two places, thermal energy will flow between them. There is then the possibility of using some of that energy to do useful work, for example, turning turbines to generate electricity. Figure 8.3 illustrates the principle by showing an example in which, for every 100 J of energy flowing between a 'hot' and a 'cold' area, 30 J of that energy can be used to do useful work and 70 J is transferred to the cold area, from where it spreads into the surroundings (environment). Of course, 30 J + 70 J = 100 J because of the law of conservation of energy. The efficiency of this process is 30/100 = 0.3, or 30%.

Energy that spreads into the surroundings (**dissipates**) cannot be recovered to do any useful work and it is known as **degraded energy**.

A typical power station may be only 35% efficient in producing electricity (which is a disappointingly low figure), so that 65% of the energy obtained from the fuel is degraded and transferred to the surroundings at the power station. Of course, we want power stations to be as efficient as possible and it is therefore important to understand why energy becomes degraded. As we saw in Chapter 2, in all mechanical processes friction will result in the dissipation of energy, but this is *not* the major reason for power station inefficiency.

To understand the inefficiency of all heat engines, consider the following simplified example. In Figure 8.4, thermal energy is supplied to gas sealed in a cylinder. The increasing pressure inside the cylinder forces the movable piston (assumed to be friction-free) to the right. In this example it is possible, in theory, for all of the thermal energy supplied to the gas to be converted into the kinetic energy of the moving piston. The system can be 100% efficient.

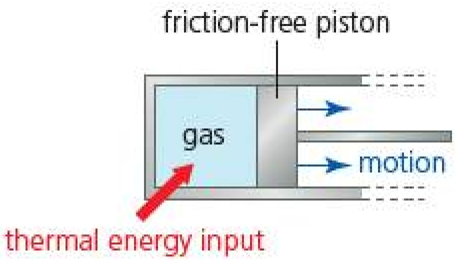


Figure 8.4 Doing work on a piston

But a heated gas in a cylinder cannot expand forever. This means that any practical heat engine must involve a process which repeats itself over and over again in order for it to work continuously. A practical heat engine must work in *cycles* in a confined space, with an expansion followed by compression, followed by expansion, etc. However, it is not possible for any cyclical process to happen without some of the thermal energy being transferred to the surroundings (out of the system). This is the main reason why power stations are so inefficient. The efficiency of a power station (or any other heat engine) improves the higher the temperature input and the lower the temperature of the outlet.

Energy flow (Sankey) diagrams

8.1.3 Construct and **analyse** energy flow diagrams (Sankey diagrams) and identify where the energy is degraded.

Energy transformations can be usefully represented in flow diagrams, such as Figure 8.5. This is another way of representing the simple situation described earlier in Figure 8.3. The width of each section is proportional to the amount of energy (or power), starting with the energy input shown at the left of the diagram. Degraded energy is shown with downwards arrows and useful energy flows to the right.

Diagrams like these are known as **Sankey diagrams** and they can be used to help represent many energy transformations.

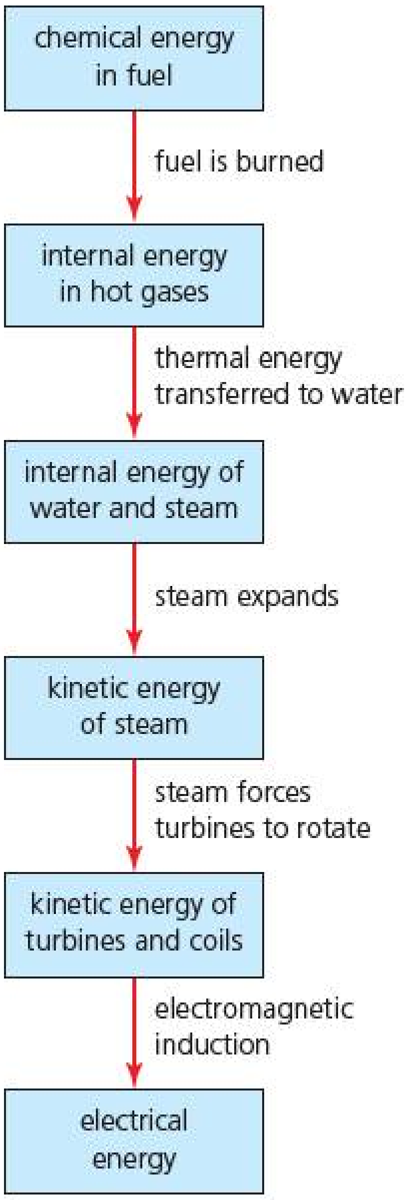


Figure 8.6 Energy transfers in a fossil fuel power station

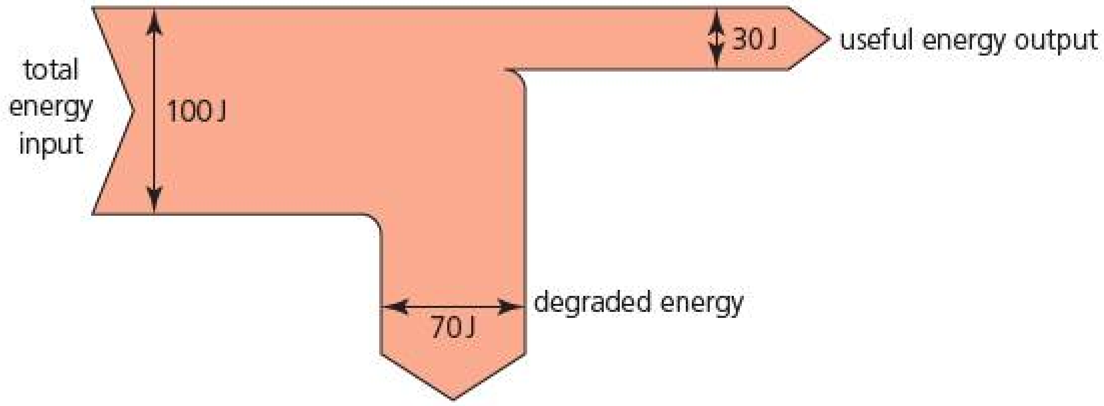


Figure 8.5 A simple Sankey diagram

Figure 8.6 represents the useful energy transformations in a fossil fuel power station (like that shown in Figure 8.2) and Figure 8.7 shows a Sankey diagram representing the energy flow through the same system, including the degraded energy.

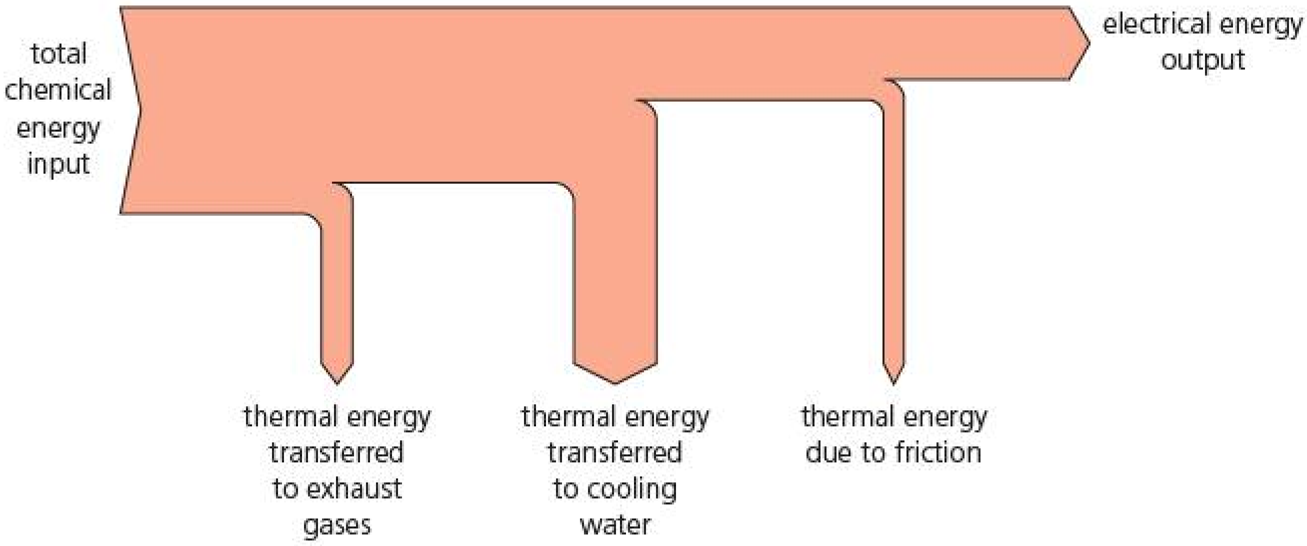


Figure 8.7 Sankey diagram for a fossil fuel power station

Computer simulations are very helpful for observing and comparing Sankey diagrams representing various physical processes.

1 Estimate the overall efficiency of the power station represented in Figure 8.7.

2 Draw Sankey diagrams to represent the energy flows for:
a a car travelling at a constant speed
b a small torch bulb powered by a battery.

3 a Which has more internal energy: 1 kg of water at 25 °C or 1 kg of water at 35 °C?
b Explain why it is theoretically possible to transfer energy from water at 25 °C to do useful work if it is in a room at 15 °C, but useful energy cannot be transferred from the same water if it is in a room at 35 °C.
c Everything around us contains very large amounts of internal energy. Explain why we are not able to extract this energy to do useful work.

8.2 World energy sources

8.2.1 **Identify** different world energy sources.
8.2.5 **State** the relative proportions of world use of the different energy sources that are available.

Data concerning the use of energy throughout the world is changeable over time and can vary in reliability. For this reason the figures quoted throughout this chapter are for guidance only and should not necessarily be assumed to be 100% accurate or up-to-date. It is wise to consult reputable websites and databases for the latest information.

A **fuel** is a widely used term for any substance from which the energy from changes within its atoms and molecules (that is, from chemical or nuclear energy) is used to do useful work. Examples include coal and uranium.

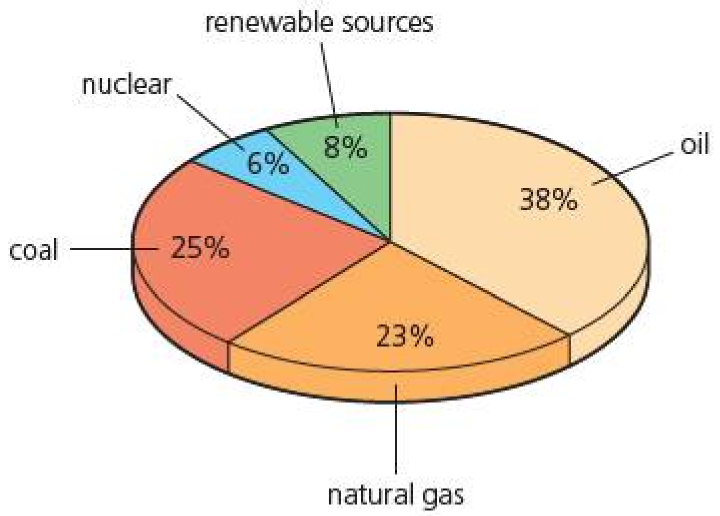


Figure 8.8 The proportion of different energy sources contributing to the total energy consumption of the world

In 2009, the total annual world energy consumption was a little less than 5×10^{20} J. Although this figure cannot be known with a high degree of certainty, it amounts to a global average power consumption of over 2000 W for every person on the planet. Of course, this is only an average and so does not take into account the great differences in energy consumption between richer and poorer countries. This figure includes energy sources used for transport, industry and the generation of electricity but it does not include the use of small-scale individual resources like firewood. About one quarter of the people on Earth do not have electricity in their homes. Sometimes figures quoted on websites or in books can be misleading as they only take into account energy sources used for the generation of electricity.

The different energy sources that are used around the world can be broken down into the following list, as shown in Figure 8.8.

- **Fossil fuels** provide about 86% of the total energy consumed in the world. The products of crude oil are the most widely used (38%), followed by coal and natural gas (25% and 23% respectively). Despite the considerable advantages of using fossil fuels, they have at least two major disadvantages – they release carbon dioxide when they are burned (and almost certainly lead to global warming) and they are non-renewable (see next section).
- **Nuclear power** provides approximately 6% of the total energy consumed in the world. This is also a non-renewable energy source.
- **Renewable energy sources** supply about 8% of the total energy consumed in the world.

Renewable and non-renewable energy sources

8.2.2 **Outline** and **distinguish between** renewable and non-renewable energy sources.

When we call an energy source **renewable** we mean that it is continuously replaced by natural processes, and the source will not be used up (run out). It will continue to be available for our use for a very long time.

Wind and waves, for example, will always be available as long as there is energy coming from the Sun to warm the planet. The main renewable sources are:

- biomass/biofuels (energy transferred from wood and other plants, and also from animal waste)
- hydroelectricity

- solar heating
- wind power
- geothermal (energy transferred from hot rocks under the ground)
- photovoltaic cells
- wave power.

Figure 8.9 Brazil uses a large amount of ethanol extracted from sugar cane as a fuel

This list is in approximate order of total energy use. Hydroelectric power is the most widely used renewable energy source for the generation of electricity. These are fast-developing technologies and the latest information can easily be researched on the Internet.

When we call an energy source **non-renewable** we mean that the source cannot be replaced once it has been used up, and so supplies will run out sometime in the future.

The main non-renewable sources are:

- fossil fuels
- nuclear power.

The original **(primary)** source of most of the energy consumed on the Earth is the Sun. Plants get their energy directly from the Sun's radiation, and plants and animals that died hundreds of millions of years ago are the source of all fossil fuels. Wind and rain are a result of temperature changes caused by the Sun's radiation and waves are formed by the wind.

The main sources from the above lists which do not get their energy from the Sun are nuclear energy (changes inside the nuclei of atoms) and geothermal energy (from radioactivity, which is also from nuclear changes). A third example is tidal energy, which gets the energy from very slight changes in the gravitational potential energy of the Earth/Moon/Sun system.

Energy density

8.2.3 **Define** the *energy density* of a fuel.
8.2.4 **Discuss** how choice of fuel is influenced by its energy density.

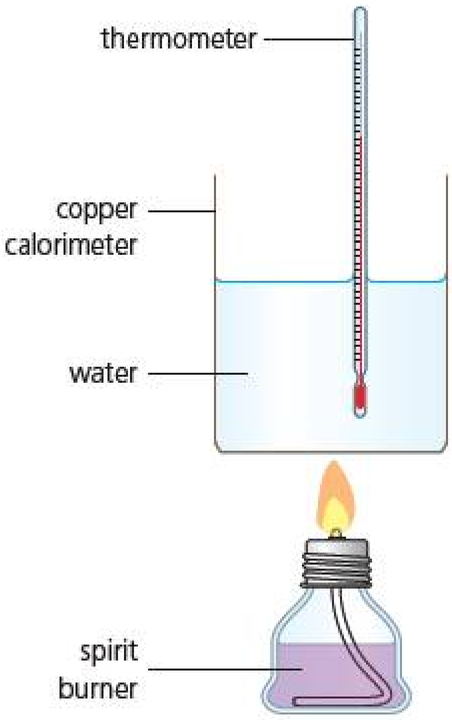


Figure 8.10 Estimating the energy density of a fuel

It is useful to know how much energy can be transferred from a certain amount of fuel so that different fuels can be compared. This can be found in a laboratory by burning a known mass of fuel and transferring the resulting thermal energy to raise the temperature of a known mass of water. (Remember that $Q = mc\Delta T$). Figure 8.10 shows how this might be done for a liquid fuel, although this simple method will not be very accurate (and you should be able to explain why).

Energy density is the energy transferred from a unit mass of fuel. It is measured in joules per kilogram, $J\,kg^{-1}$.

For many renewable energy sources, like wind power for example, it may not be sensible to refer to a numerical value of energy density, but these sources are sometimes described as having *low energy density*.

A major advantage of nuclear power is its high energy density. Of the fossil fuels, natural gas (methane) has the highest energy density.

When using a fuel with a high energy density, less mass of fuel will be needed in order to provide a certain amount of energy compared to using a fuel with a lower energy density.

Table 8.1 Energy densities

Material	Energy density/$MJ\,kg^{-1}$
Nuclear fusion	580 000 000
Nuclear weapons	90 000 000
Uranium used in nuclear reactors	3 500 000
Hydrogen	143
Natural gas (methane)	55
Gasoline	46
Crude oil	46
Vegetable oil	42
Ethanol	30
Coal	28
Wood	17
Typical carbohydrate (food)	17
Cow dung	15
Household waste	9
Torch battery	0.1
Water at height of 100 m	0.001

4 In an experiment similar to that shown in Figure 8.10, when 1.24 g of fuel was completely burned in air the temperature of 54.8 g of water rose from 18.0 °C to 83.4 °C.
- **a** Calculate a value for the energy density of the fuel.
- **b** Would you expect the actual value to be higher or lower than your answer? Explain.

5 A country has a population of 46 million people and a total energy consumption of 5.2×10^{18} J in one year.
- **a** Calculate the average energy consumption per person in one year.
- **b** What is the average power consumption per person?
- **c** An average home has 2.8 people and a total power consumption of 2.7 kW. What percentage of the country's energy consumption occurs in people's home?
- **d** Suggest what the rest of the energy is used for.

6
- **a** How much gasoline is needed to accelerate a 1500 kg car from rest to $20\,m\,s^{-1}$ if the overall efficiency of the process is 32%?
- **b** Estimate the rate at which the same car consumes gasoline if it is travelling at a constant speed of $20\,m\,s^{-1}$ against a total resistive force of 2000 N.
- **c** What is the energy density of a torch battery containing 7.9 g of chemicals which supplies enough energy to run a 0.85 W light bulb for 110 minutes?

7 A coal-burning power station has an efficiency of 36% and an output power of 312 MW.
- **a** Calculate the mass of coal burned:
  - **i** every second
  - **ii** every week.
- **b** Approximately how much uranium would be needed each week for a nuclear power station of the same output but with an efficiency of 42%?

8
- **a** Find out which countries of the world have the highest average power consumptions (per person/capita).
- **b** Suggest possible reasons why the people of those countries use so much energy.

9 Suggest in what ways a list of the energy sources used in large-scale electrical power generation might be different from a list of overall energy sources used in the world.

10 Find out from the Internet the latest figures for the use of renewable energy sources:
  a in the world
  b in the country where you live.

Advantages and disadvantages of different energy sources

8.2.6 **Discuss** the relative advantages and disadvantages of various energy sources.

The most important factors to consider when discussing the advantages and disadvantages of individual energy sources are:

- greenhouse gas emissions and possible effects on global warming
- risks to human health/life
- possible pollution and environmental effects
- whether the source is renewable or non-renewable
- expense
- energy density
- whether the energy is continuously available, or dependent on factors such as weather conditions or the time of the day/night.

These factors will be considered throughout this chapter when each energy source is discussed.

8.3 Fossil fuel power production

Fossil fuels

8.3.1 **Outline** the historical and geographical reasons for the widespread use of fossil fuels.

The first heat engines which usefully transferred chemical energy into kinetic energy (using steam) were developed in the early 18th century in the United Kingdom. They were very inefficient and used mostly for pumping water out of mines. Over the next century intensive efforts went into improving their design and engines became widely used in factories and then for transport. The change from humans doing hard physical work to the use of machines to do the same work much more easily and quickly is usually known as the Industrial Revolution.

At the start of the Industrial Revolution there was a rapidly increasing demand for the fuels needed to run the engines. In the early years of development coal was the obvious choice for a fuel (wherever it was available) because it was energy dense, cheap and plentiful. Industries were usually located close to the mines so that the coal did not have to be transported large distances.

It was not until the end of the 19th century that science and technology had advanced to the point where power stations could be built in order to burn fuels to generate electricity for large numbers of people (see Figure 8.11). Before that time, coal and wood were used for heating and gas, oil or candles were used for lighting. Even today almost half of the world's population still uses local wood or other plant or animal materials for providing some of their energy.

Figure 8.11 A fossil fuel power station

Coal is an example of a fossil fuel. Oil and natural gas are also fossil fuels. Fossil fuels are formed underground by the action of high pressure and temperatures in the absence of air over many millions of years. Coal is formed from dead plants, and oil and natural gas are formed from dead marine creatures. Although the processes which form fossil fuels are still happening today, due to the very long time needed for their formation, they are known as non-renewable energy sources. We are using them at a very much quicker rate than they are being formed.

8.3.2 **Discuss** the energy density of fossil fuels with respect to the demands of power stations.

Energy density of fossil fuels

All fossil fuels have high energy densities compared to most other energy sources except nuclear energy (as shown in Table 8.1). This is one of the main reasons that the world has become so dependent on them. The following worked example shows a typical calculation involving the fuel consumption of a gas-fired power station.

Worked example

1 Estimate the useful output power of a gas-fired power station that uses natural gas at a rate of $15.0\,\mathrm{kg\,s^{-1}}$. Assume that the energy density of the gas is $55\,\mathrm{MJ\,kg^{-1}}$ and the efficiency of the power station is 45%.

Power output, P = efficiency × mass of fuel burned every second × energy density

$$P = 0.45 \times 15.0 \times (55 \times 10^6)$$
$$P = 3.7 \times 10^8\,\mathrm{W}$$

Transportation of fossil fuels

8.3.3 **Discuss** the relative advantages and disadvantages associated with the transportation and storage of fossil fuels.

Figure 8.12 The oil spill in the Gulf of Mexico, 2010

Because of its high energy density, the cost of transporting coal over quite large distances by rail is not too great when compared with the other expenses of power generation. For similar reasons, oil and gas (being fluids) are commonly moved around the world through very long pipelines. Large quantities of oil are also moved around the world in large ships (tankers). The advantages of moving fossil fuels from where they are extracted from the ground to where they are used are considerable, but so too are the disadvantages, especially when accidents result in oil being spread widely into the environment. This kind of pollution can be particularly harmful when an oil spill occurs at sea, such as the accident off the coast of the USA in 2010 (see Figure 8.12). Because of their high energy density, it is possible to store enough fossil fuels to provide for the energy demands for many weeks ahead.

Efficiency of fossil-powered power stations

8.3.4 **State** the overall efficiency of power stations fuelled by different fossil fuels.

Natural gas power stations are the most efficient types of power stations. They are able to convert nearly half of the chemical potential energy in the gas into electrical energy (approximately 45% efficient). Coal-fired power stations are typically no more than 35% efficient. Oil-fired power stations have efficiencies of about 40%. (These figures are for rough guidance only and they have been rounded-off to the nearest 5% to make them easy to remember.) Chapter 10 uses the laws of thermodynamics to explain why it is impossible to achieve much higher efficiencies than these values.

How long will fossil fuels reserves last?

There are many reasons why it is not easy to predict how many years fossil fuel reserves will last. However, we can be sure that sometime within the next 100 years the world will have to face the major problem of very significant reductions in the availability of coal, crude oil and natural gas as the Earth comes to the end of the 'fossil-fuel age' and cheap energy. At the same time there will be a beneficial decrease in the emission of greenhouses gases.

Figure 8.13 The USA has the greatest reserves of coal

At the present rate of usage, existing coal reserves will probably last about another 150 years, natural gas about 60 years and crude oil about 40 years. But these figures can be very misleading because we cannot be sure about:

- how much fossil fuel remains to be discovered
- whether existing sources now considered uneconomic to extract from the ground will later become viable
- to what extent the use of renewable sources and nuclear power will increase
- by how much the world's consumption of energy will increase.

On top of these factors, economic and political pressures will probably have unforeseen and all-important effects. Of course, some countries are 'rich' in fossil fuel reserves (for example, the USA has the most coal reserves, see Figure 8.13), while other countries have very little or none.

Advantages and disadvantages of using fossil fuels

8.3.5 Describe the environmental problems associated with the recovery of fossil fuels and their use in power stations.

Table 8.2 Advantages and disadvantages of using fossil fuels

Advantages	Disadvantages
■ High energy density. ■ Fuel is relatively cheap (although economic and political factors may result in significant and sudden changes in price). ■ Power stations are relatively inexpensive to construct and maintain (when considering their large power outputs). ■ Power stations can be built in almost any location (which has good transport links and a plentiful water supply). ■ These are established technologies – power stations, transport and storage systems already exist.	■ Greenhouse gas emissions and global warming. ■ Chemical pollution during mining and burning (including acid rain). ■ Non-renewable sources. ■ Extraction/mining can damage the environment and be dangerous and hazardous to health. ■ Leakage from oil tankers or pipelines can cause considerable harm to the environment.

11 An old-fashioned steam train had an output power of 2.3 MW but an efficiency of only 8.4%. If the coal used had an energy density of 29 MJ kg^{-1}, how much coal had to be burned every minute?

12 a How much fuel (kerosene) does a jet airliner consume every second when travelling at a constant speed of 240 m s^{-1} with a power output of 69 MW? Assume that the fuel has an energy density of 43 MJ kg^{-1} and the engines have an efficiency of 32%.
b What is the resistive force acting on the airliner?
c Estimate the amount of fuel needed to travel a distance of 5000 km at this constant height and speed.

13 A 220 MW oil-fired power station has an efficiency of 40% and uses fuel at a rate of 13 kg s^{-1}. What is the energy density of the fuel?

14 a What is the efficiency of a coal-fired power station which produces an average 560 MW of output power when burning fuel at a rate of 4.8×10^6 kg every day? (Take the energy density of the coal to be 30 MJ kg^{-1}.)
b If the discarded thermal energy was removed from the power station by cooling water, which should not increase in temperature by more than 4.0 °C, calculate the minimum rate of flow of cooling water that would be needed. The specific heat capacity of water is 4180 J kg^{-1} °C^{-1}.
c What mass of coal must be burned every year to supply the needs of a home which uses an average power of 3 kW?

15 Use the Internet to find out the latest information on how long fossil fuel reserves are expected to last.

8.4 Non-fossil fuel power production

Fossil fuels provide about 86% of the world's energy. The remainder is provided by nuclear power (≈ 6%) and renewable sources (≈ 8%). In this section we will first look at nuclear power and then take a closer look at four of the renewable energy alternatives: solar, hydroelectric, wind and wave powers.

Nuclear power

There are a total of approximately 500 nuclear power stations in about 30 different countries around the world. The USA has the greatest number of reactors but Lithuania and France have the highest percentage of their total electrical energy generated from **nuclear reactors**. China and Russia are planning to build the largest number of reactors. The use of nuclear power is a controversial issue which stirs up strong feelings in many people, but it is not sensible to form an opinion without first understanding some of the physics involved.

Chain reactions and energy transfers

8.4.1 **Describe** how neutrons produced in a fission reaction may be used to initiate further fission reactions (chain reaction).
8.4.2 **Distinguish between** controlled nuclear fission (power production) and uncontrolled nuclear fission (nuclear weapons).
8.4.4 **Describe** the main energy transformations that take place in a nuclear power station.

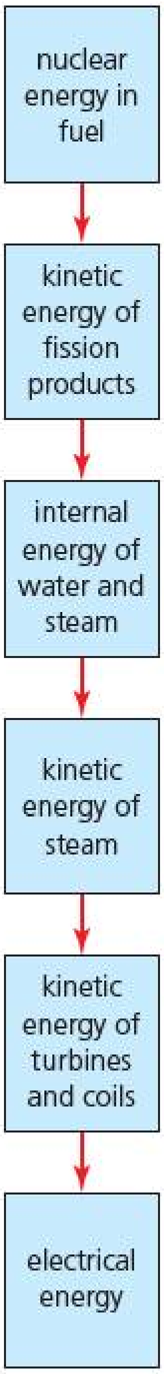


Figure 8.14 Energy flow in a nuclear power station

The energy released in a nuclear reactor comes from the *fission* of nuclei, not from radioactive decay. When a fission reaction in a nucleus of ^{235}U is initiated (started) by the *capture* of a neutron, the process will release about 200 MeV (= 3.2×10^{-11} J) of energy (for more details see Chapter 7). The nuclear potential energy is transferred to the kinetic energy of the resulting nuclei (the **fission fragments**) as well as to the neutrons and photons. This is an enormous amount of energy from one tiny nuclear reaction.

If the same process can be repeated with a very large number of uranium nuclei, the resulting increased kinetic energy of so many particles will effectively be a very large rise in the internal energy of the material. The uranium will get much hotter with the possibility of large-scale production of electrical energy in much the same a way as in a conventional fossil-fuelled power station. The material used is called the **nuclear fuel** and it is usually used in the shape of *rods*. Obviously, a nuclear fuel is different from a fossil fuel because it is not burned to transfer energy. Figure 8.14 shows the main energy transformations in a nuclear power station.

The following equation is just one of several possible fission reactions that nuclei of ^{235}U may undergo when they interact with slow moving neutrons. It is also shown diagrammatically in Figure 8.15.

$$^{1}_{0}n + ^{235}_{92}U \rightarrow ^{144}_{56}Ba + ^{89}_{36}Kr + 3^{1}_{0}n + \text{about } 200\,\text{MeV energy}$$

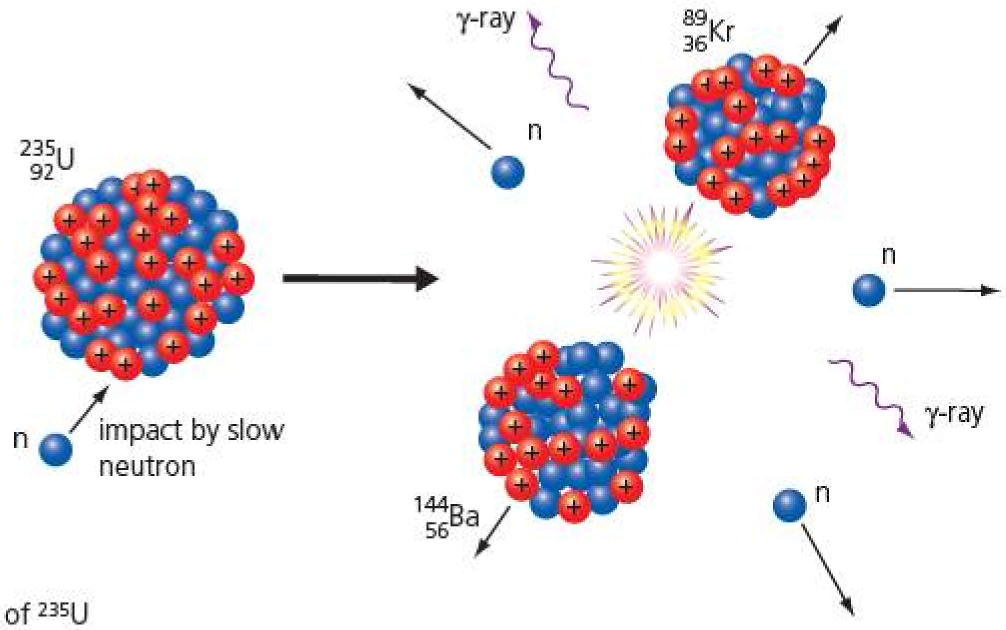


Figure 8.15 Fission of ^{235}U

When a neutron collides with a nucleus of ^{235}U, fission is not certain, in fact it is unlikely, and typically only a small percentage of interactions will initiate nuclear fission. In the equation above it is clear that, in this example, each nuclear fission produces three more neutrons (the average for all ^{235}U fissions is about 2.5) and these three neutrons may themselves then go on to cause further fissions. If on, average, at least one of the neutrons causes further fission, then the process will continue to occur and it is described as being a **chain reaction**. This type of reaction is difficult to achieve.

If, on average, each fission reaction leads to one further fission reaction (and so on), then the number of fissions per second is *constant* and the process can be described as **controlled** or **self-sustained** nuclear fission. This process provides the basis for nuclear reactors generating electricity in power stations. However, if each fission produces more than one further fission the reaction is **uncontrolled** and the number of fissions per second will rise rapidly. This will result in the release of an enormous amount of energy very quickly. This is what happens in a nuclear weapon.

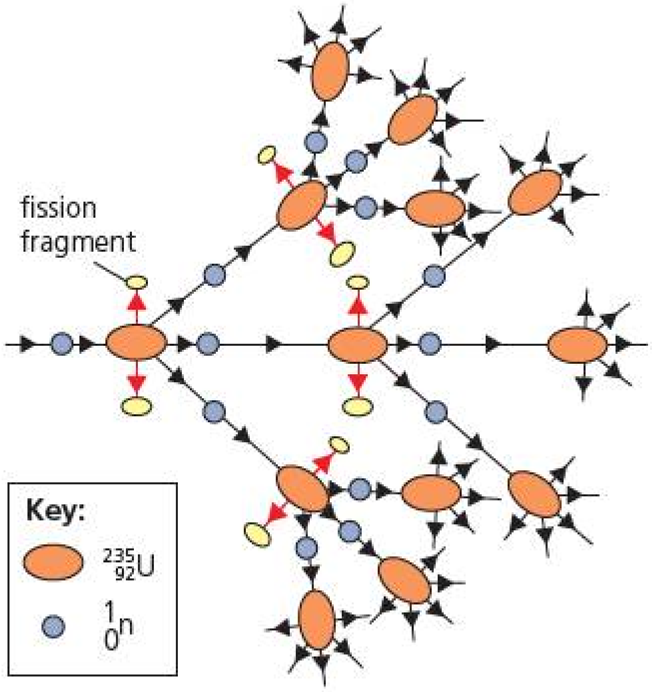


Figure 8.16 The start of a chain reaction

Figure 8.16 shows an idealized example of the start of a chain reaction in which all neutrons cause further fissions.

A relatively small number of fission reactions occur randomly in any sample of ^{235}U all the time. These reactions provide the neutrons needed for further fission reactions. In order to encourage and control nuclear fission, scientists need to understand the main factors which affect the probability of fission occurring. These are as follows:

1 Neutrons are penetrating particles (because they are uncharged) and many or most of them will usually pass out of the material without interacting with any nuclei.
2 Uranium-235 atoms are only a small percentage of all the atoms in the material.
3 When a neutron does interact with a ^{235}U nucleus, it will only cause fission if it is travelling relatively slowly. Fast neutrons do not cause fission.

Critical mass

The ratio of volume to surface area of a solid increases as it gets larger (consider solid cubes of different sizes). This means that the more massive a material is, the smaller the percentage of neutrons that will reach the surface and escape. That is, a greater percentage will cause fission. The **critical mass** of the material is the minimum mass needed for a self-sustaining chain reaction. (Uranium that contains 20% ^{235}U has a critical mass of over 400 kg.)


Figure 8.17 Uranium fuel rods

Fuel enrichment: increasing the percentage of uranium-235

8.4.3 **Describe** what is meant by fuel enrichment.

Uranium-235 is the only nuclide on the planet which occurs in significant quantities and can sustain a chain reaction. However, when uranium ore is extracted from the ground and refined, the uranium atoms are approximately in the ratio 99.3% ^{238}U and only 0.7% ^{235}U (with traces of other uranium isotopes). All the isotopes of uranium are radioactive, but the half-life of ^{238}U is very long (4.5×10^9 years), similar to the age of the Earth, whereas ^{235}U has a half-life of 7.0×10^8 years. For a chain reaction and power generation, the percentage of ^{235}U has to be increased at the very least to about 3%, although higher percentages are preferable. (Nuclear weapons require a much higher percentage.) This process is called **fuel enrichment**. Figure 8.17 shows a photograph of enriched uranium fuel rods.

Uranium-238 nuclei can absorb neutrons without causing fission, so too much ^{238}U will also discourage a chain reaction. Enrichment cannot be done chemically, because isotopes of the same element have identical chemical properties, so physical processes need to be involved (for example, using the diffusion of gaseous uranium compounds) but these are difficult and expensive technologies. The remaining uranium is called **depleted uranium** and it has physical properties which have made it useful in military engineering, but this has been controversial (because it is radioactive).

Moderator, control rods and heat exchanger

8.4.5 **Discuss** the role of the moderator and the control rods in the production of controlled fission in a thermal fission reactor.
8.4.6 **Discuss** the role of the heat exchanger in a fission reactor.

The neutrons released in nuclear fission have typical energies of about 1 MeV, which means that they travel very fast. This is usually too fast to initiate another fission reaction. The slower a neutron travels, the higher the probability it has of it being captured by a uranium nucleus. Therefore, before a chain reaction can occur, the neutrons need to be slowed down to energies of about 1 eV or less (they are often described as thermal neutrons). This process is called moderating the neutrons and the material used is called a **moderator**.

In order for the fast neutrons to lose so much of their kinetic energy they need to collide many times with the nuclei of atoms (but not be captured). In general, when particles collide there is a greater transfer of kinetic energy between them if they have approximately the same mass (see Chapter 2). The mass of a neutron is always smaller than the mass of a whole nucleus, but the difference is less for nuclei with low mass. This is why atoms with nuclei of small mass are preferable for this process of moderation. Commonly, graphite or water is used as the moderator.

In the pressurized water-cooled reactor shown in Figure 8.18, the nuclear fission occurs in the **fuel rods** in the **reactor vessel**. As the cold water flows past the hot fuel rods it removes thermal energy and at the same time acts as the moderator, slowing down the neutrons.

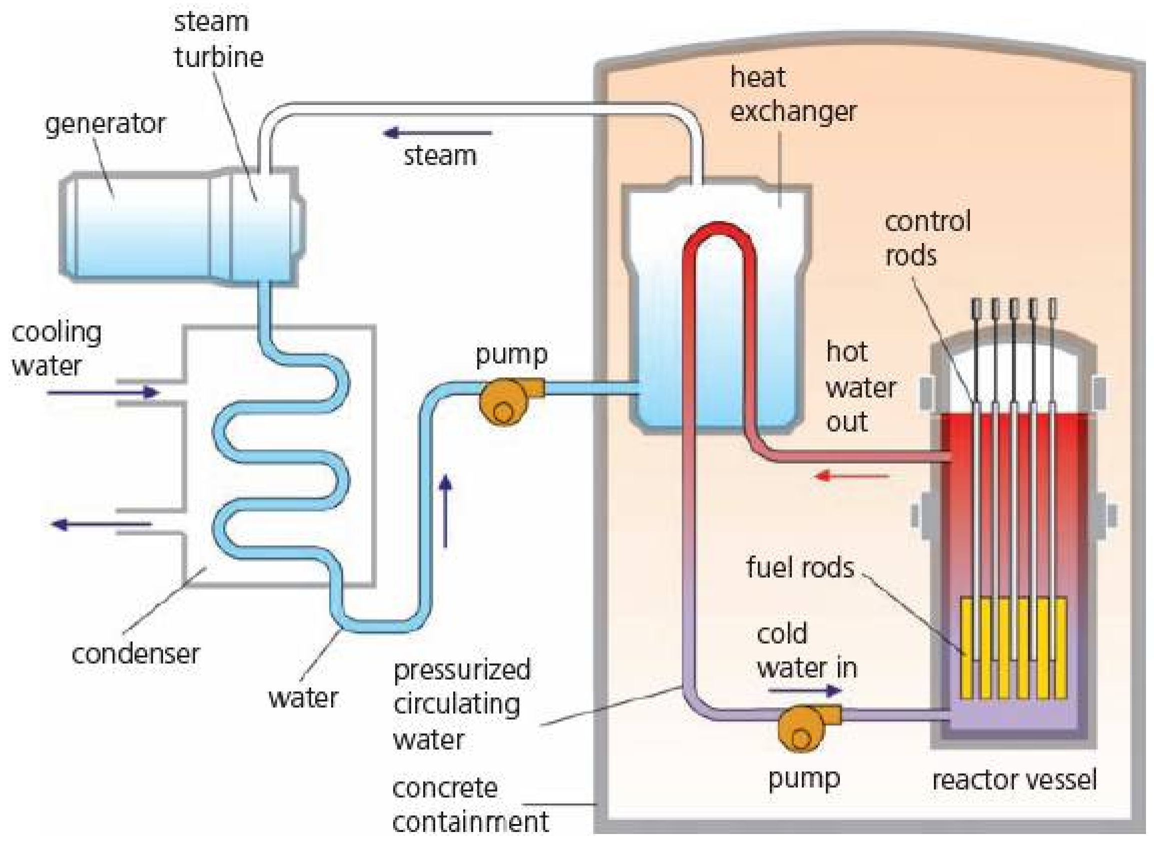


Figure 8.18 A pressurized water-cooled reactor

The hot water that flows through the reactor vessel is in a sealed system and it never leaves the concrete containment. Its temperature may be as high as 300°C, but it does not boil because it is under very high pressure. It is pumped to a **heat exchanger** in which the thermal energy in the water is transferred to more water in a *separate* system (this is an important safety measure). Steam is generated, which turns the turbines to generate electricity.

The **control rods** (typically made with boron) are used for adjusting the rate of the fission reactions by absorbing neutrons. This is done by moving the rods into or out of the system as necessary. Figure 8.19 represents the use of control rods to produce controlled fission.

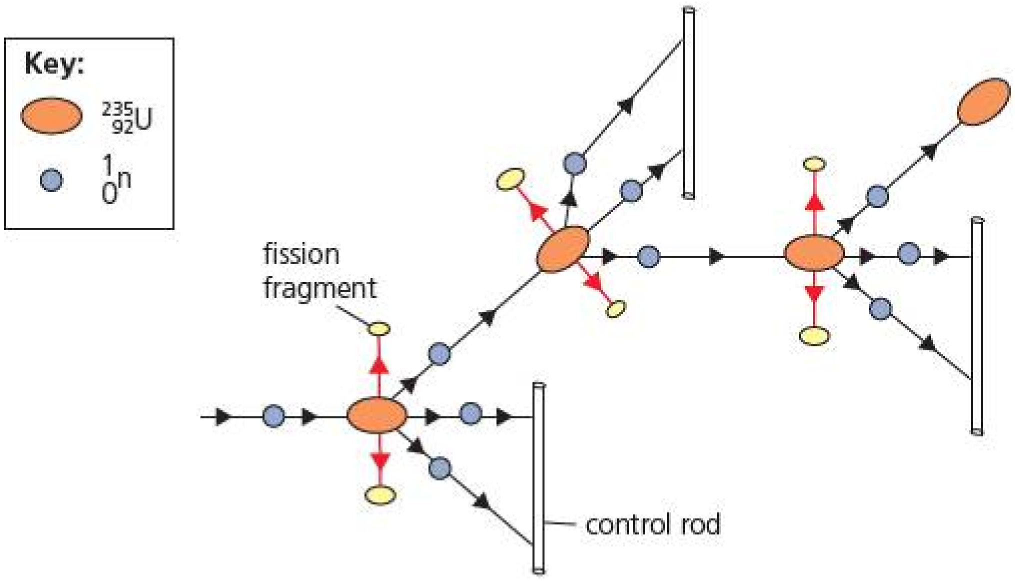


Figure 8.19 How control rods are used to control a chain reaction

Students are advised to use computer simulations to model the working of nuclear power stations and other nuclear processes.

16 Calculate the speeds of:
a a 1 MeV neutron
b a 1 eV neutron.

17 Suggest why isotopes of uranium are so difficult to separate by physical methods.

18 A nuclear power station using ^{235}U has an efficiency of 43% and a useful power output of 1 GW.
a If the average energy released per fission is 190 MeV, calculate how many fissions occur every second.
b What mass of ^{235}U is needed every day?

19 **a** Calculate the energy density of pure ^{235}U (in $J\,kg^{-1}$).
b If ^{235}U is only 3% of the total mass of the fuel, calculate the energy density of the fuel.
c Compare your answer for **b** with a type of coal which can transfer 29 $MJ\,kg^{-1}$.

20 Another possible fission of ^{235}U results in the formation of $^{138}_{55}Cs$ and $^{96}_{37}Rb$. Write a nuclear equation for this reaction.

21 **a** If an individual uses electrical energy at an average rate of 1 kW, what is their annual energy consumption?
b What mass of ^{235}U atoms has to undergo fission to provide the energy needed for a year (assume 100% efficiency)?

Plutonium

8.4.7 **Describe** how neutron capture by a nucleus of uranium-238 (^{238}U) results in the production of a nucleus of plutonium-239 (^{239}Pu).
8.4.8 **Describe** the importance of plutonium-239 (^{239}Pu) as a nuclear fuel.

Uranium-235 is the only naturally occurring nuclide that can be used in nuclear reactors, but the production of **plutonium-239** provides another possibility. Plutonum-239 does not exist in significant quantities naturally but it can be produced in the following way.

Uranium-238 nuclei can absorb neutrons without causing fission. The resulting nucleus, ^{239}U, then radioactively decays by beta particle emission to ^{239}Np, which then decays to ^{239}Pu, also by beta particle emission.

$$^{1}_{0}n + ^{238}_{92}U \rightarrow ^{239}_{92}U$$

$$^{239}_{92}U \rightarrow ^{239}_{93}Np + ^{0}_{-1}e + \bar{\nu}_e$$

$$^{239}_{93}Np \rightarrow ^{239}_{94}Pu + ^{0}_{-1}e + \bar{\nu}_e$$

These two decays have very short half-lives compared with ^{239}Pu (which has a half-life of 24 000 years). This means that the capture of neutrons by ^{238}U nuclei is effectively producing ^{239}Pu, which is a fissionable material and can be used as a fuel in a nuclear reactor. Plutonium-239 has the advantages of being more easily fissile than ^{235}U and releasing slightly more energy per fission.

Safety issues and nuclear weapons

Nuclear accidents

8.4.9 Discuss safety issues and risks associated with the production of nuclear power.

It is fair to say that the risks of nuclear power are very well understood and, as a result, safety standards are very high. But, for many people, this is not reassuring enough because no matter how careful nuclear engineers are, accidents and natural disasters can happen. Safety standards do vary from country to country, but the consequences of a nuclear accident anywhere could be really disastrous.

The world's worst nuclear accidents at Chernobyl in Ukraine in 1986 and Fukushima (see Figure 8.20) in Japan in March 2011 remain as vivid warnings of the possible risks of nuclear power. The radiation leaks and explosions at Fukushima followed the damage caused by a tsunami, whereas the 'meltdown' at Chernobyl is widely recognized to be the result of poor design and management. Modern designs and safety procedures have been vastly improved but may still be insufficient in the face of an extreme natural disaster. Hundreds of thousands of people had to be evacuated because of these accidents, and the number of long-term illnesses and deaths caused by them will take many years to be confirmed.

A **thermal meltdown** is probably the most serious possible consequence of a nuclear accident. If, for some reason (for example, the loss of coolant or the control rods failing to work properly), the core of the reactor gets too hot or even melts, the reactor vessel may get badly damaged. Fires and explosions may happen as extremely hot materials are suddenly exposed to the air. Highly concentrated and dangerous radioactive materials may then be released into the ground, water or air so that they are spread over large distances by geographical and weather conditions.

Figure 8.20 Fukushima nuclear reactor after explosions, in the wake of the tsunami (2011)

When considering the dangers of nuclear power, it should always be remembered that the uses of all other kinds of energy resources also have their various risks. In particular, coal mining has been responsible for an extremely large number of serious injuries, long-term health problems and deaths for over 200 years.

Nuclear waste

The products of fission reactions are highly radioactive and can therefore be dangerous if safety standards are not high enough. Anything associated with the reactor must be considered as a potential health risk, although the greatest dangers are obviously posed by the fuel rods themselves. Nuclear waste products, like the fuel rods after they are no longer useful, are highly radioactive and contain some isotopes with very long half-lives, which means that they will be dangerous for many thousands of years. As an example, ^{99}Tc has a half-life of 210 000 years.

Figure 8.21 Storage of nuclear waste

There is no way to stop a radioactive material from emitting nuclear radiation, so safety measures have to concentrate on preventing people from being exposed to the radiation. Radioactive substances need to be surrounded by materials which will absorb the radiation safely. Thick concrete, water and lead are widely used for this purpose, depending on the circumstances. Exposure to radiation is also reduced by keeping people as far away as possible or, if they need to be close, by limiting the length of time of their exposure.

The disposal of highly dangerous nuclear waste remains a serious problem which we are leaving for generations to come. At present, some of this waste is stored deep underground in very strong and thick containers designed to prevent leakage, and some is stored above ground where it can be more easily monitored (see Figure 8.21).

Mining

Figure 8.22 McClean Lake uranium mine in Saskatchewan, Canada

Uranium is widespread as a trace element in many minerals around the world, but more than half of the world's uranium mines are located in only three countries: Canada (see Figure 8.22), Australia and Kazakhstan. A lot of rock has to be extracted from large, open-cast mines to obtain only a small amount of uranium, which is usually in the form of uranium oxide.

One of the products from the decay of uranium nuclei is the radioactive gas, radon. This makes uranium mines potentially dangerous places. Because it is a gas, any radon present in the air can enter the lungs, and once it is inside the body the radiation that it emits is far more dangerous than similar radiation sources outside the body.

Nuclear weapons

The first nuclear bomb was tested in 1945 in the USA. The latest country to develop nuclear weapons was North Korea in 2006. It is believed that nine different countries have nuclear weapons; they all claim that these are just for national defence and that the weapons would never be used except under the extreme circumstances of being attacked by another country.

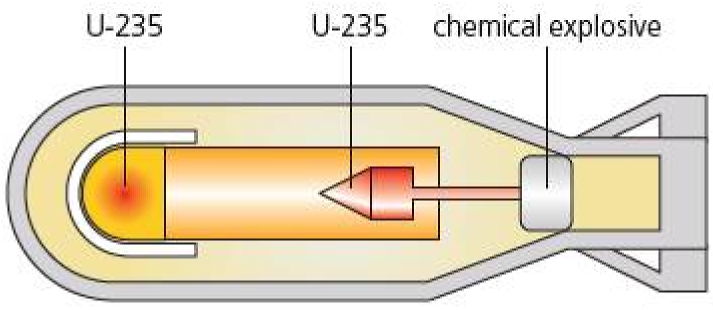


Figure 8.23 Sub-critical masses in a nuclear bomb. A conventional explosion forces the two sub-critical masses together

The destructive power of just a single nuclear bomb was horrifyingly demonstrated at both Hiroshima and Nagasaki in Japan in 1945. The basic requirement is a critical mass (but not all together in one piece) of ^{239}Pu or highly enriched ^{235}U (at the very least enriched to 20%, but preferably much more). If the bomb is to be detonated, two separate halves of the critical mass have to be brought together by a relatively small conventional explosion, as shown in Figure 8.23. This then produces an uncontrolled chain reaction.

The production of enriched 'weapons grade' uranium is technically extremely difficult and expensive. The alternative, ^{239}Pu, is a product of neutron capture by ^{238}U nuclei in a nuclear power station and it can later be recovered from the fuel rods. This is why there is concern that a few counties may use their development of a nuclear power programme to (secretly) develop nuclear weapons.

Many people claim that the threat of nuclear weapons becoming available to the 'wrong' countries/organizations/people is one of the biggest threats facing the world in the 21st century. However, some people believe that the threat of the use of nuclear weapons has actually helped prevent major wars, although many others believe that there could never be any justification for the construction or use of nuclear weapons under *any* circumstances. Regretfully, in the modern world we need to be aware of the possibilities of terrorist attacks on nuclear power stations, or the transportation of nuclear materials, or the use of nuclear or 'dirty' bombs by terrorists.

Advantages and disadvantages of nuclear power

Table 8.3 Advantages and disadvantages of nuclear power

Advantages	Disadvantages
■ Extremely high energy density.	■ Dangerous and very long-lasting radioactive waste products.
■ No greenhouse gases emitted during routine operation. (Some scientists think that nuclear power may be the only realistic solution to global warming.)	■ Expensive.
■ No chemical pollution during operation.	■ Efficiency is not high when the whole process is taken into account.
■ Reasonably large amount of nuclear fuels are still available.	■ Threat of serious accidents.
■ Despite a few serious incidents, statistically over the last 50 years, nuclear power has overall proven to be a reasonably safe energy technology.	■ Possible target for terrorists.
	■ Linked with nuclear weapons.
	■ Not a renewable source.

Figure 8.24 The 3.6 GW nuclear power station in Cruas, France, supplies about 5% of the country's electricity needs

Figure 8.25 Protestors try to stop the shipment and storage of nuclear waste

Nuclear fusion

8.4.10 Outline the problems associated with producing nuclear power using nuclear fusion.
8.4.11 Solve problems on the production of nuclear power.

Nuclear fusion is the source of the energy in all stars, including the Sun (see Chapter 7). A typical fusion reaction like that shown in the following equation would transfer about 18 MeV.

$$^{2}_{1}H + ^{3}_{1}H \rightarrow ^{4}_{2}He + ^{1}_{0}n$$

The two nuclei of hydrogen (known as deuterium and tritium) are both positively charged and they will only fuse together if they have enough kinetic energy to overcome the repulsive forces between them. To achieve this high kinetic energy requires high temperatures of more than 10^8 K. At this high temperature electrons move around separately from hydrogen ions (protons) and the gas is called a **plasma**.

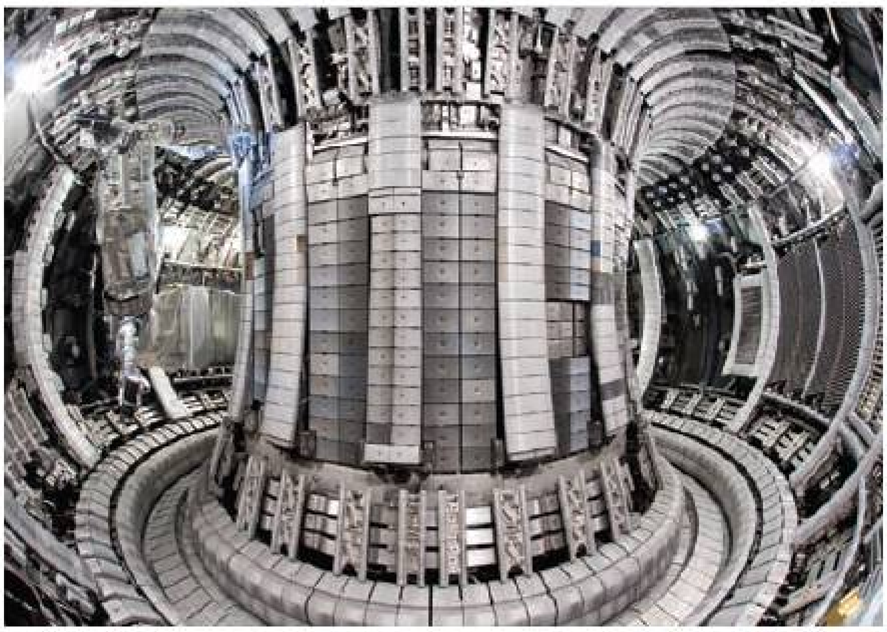


Figure 8.26 Russian Tokamak thermonuclear reactor

Nuclear fusion at these temperatures has been achieved on Earth, but the main problem is continuing (sustaining) the reactions over any significant lengths of time. The hot plasma cannot be allowed to come in contact with the walls of a container because that would cause it to cool down and the container would also become contaminated.

To avoid this problem the plasma is confined (contained) using a very strong magnetic field, such as in the Russian designed Tokamak thermonuclear reactor, which uses a doughnut (torus) shaped field (Figure 8.26).

If energy could be released on a large scale by nuclear fusion, it would have the advantage of a more plentiful fuel supply with fewer radioactive waste problems, compared to nuclear fission. These advantages are enormous, but despite years of considerable scientific efforts, the technology needed to sustain nuclear fusion reactions is still proving too difficult.

22 All the isotopes of uranium are radioactive and decay into other elements. Explain why there is still a significant amount of uranium left on Earth.

23 If, after a nuclear accident involving an isotope with a half-life of 12 years, the radiation level in the environment was 33 times greater than normal background count, approximately how much time would have to pass before people could re-enter the area?

24 a Dangerous nuclear waste is to be stored underground for more than 100 000 years. Suggest ways in which this plan could go wrong.
b Is putting the waste on a rocket and firing it into the Sun a sensible alternative? Explain your answer.

25 a Use Coulomb's law to calculate the repulsive force between a tritium nucleus and a deuterium nucleus when they are separated by a distance of 5.0×10^{-15} m.
b Calculate the acceleration with which the deuterium nucleus would start to move away.

26 a Use the Internet to gather data on the total number of deaths and serious illnesses that have been attributed to the Chernobyl accident since it occurred in 1986.
b You will probably find that the information from various web sites is different. Suggest reasons why consistent information is difficult to obtain.
c How can you decide which web site gives you the most reliable information?

27 Use the Internet to find out the latest developments in nuclear fusion: what power was achieved, and for how long? What nuclides were used? Where did it take place? In particular, what and where is ITER?

28 a What is your opinion about the use of nuclear power? In 100 words or fewer explain why you think that your country should, or should not, operate nuclear power stations.
b Conduct a survey of the opinions of your fellow students. Are opinions of physics students different from those of non-physics students?

29 a Under what circumstances (if any) do you think that it would be acceptable for one country to use a nuclear weapon to bomb another?
b Do you believe that every country has the right to decide for itself whether to build nuclear weapons? If not, who should control any development of the world's nuclear weapons?

Solar power

In general, the power of radiation per unit area is known as **intensity (I)**.

$$\text{intensity} = \frac{\text{power}}{\text{area}}$$

$$I = \frac{P}{A}$$

This equation is in the IB *Physics data booklet*.

The intensity of the Sun's radiation arriving at the Earth is approximately 1400 W m^{-2}. Although this figure can vary by approximately 7%, it is known as the **solar constant**. It is the power passing through each square metre perpendicular to the Sun's 'rays' as they reach the outer limits of the atmosphere, as shown in Figure 8.27. (This power is across all wavelengths of the Sun's emission spectrum, although about half is in the visible part of the spectrum.)

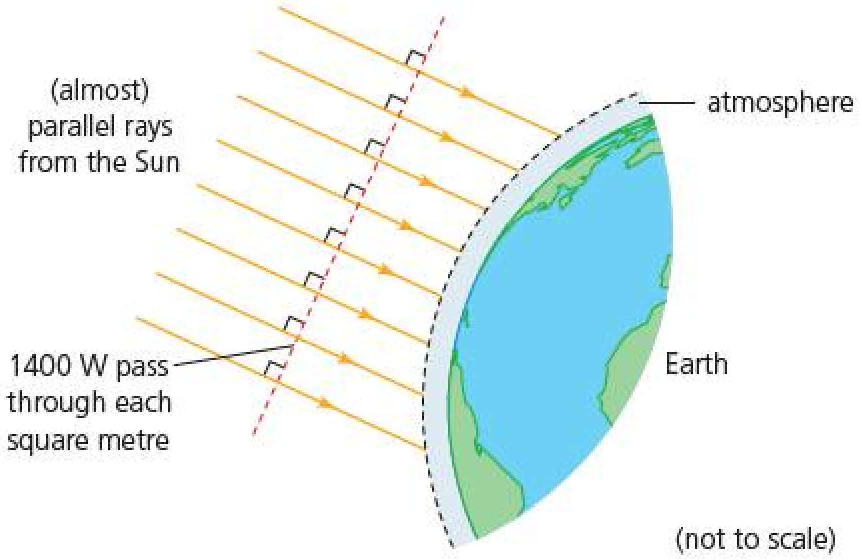


Figure 8.27 Solar radiation arriving at the Earth

This is an enormous amount of energy incident on the Earth, although some of the energy from the Sun is absorbed, reflected and scattered (reflected randomly) in the atmosphere. For example, on a clear day with the Sun directly overhead, a figure for the intensity on the Earth's surface is sometimes quoted to be about 1 kW m^{-2}.

A more useful value, averaged over 24 hours and over the whole planet, is about 235 W m^{-2}. Of course, the energy transferred by radiation from the Sun is essential for all life on Earth and it is important to remember that it is, or has been, the indirect origin of the majority of the energy resources discussed in this topic. However, surprisingly little use has been made of the *direct* transfer of solar energy to other energy forms.

Photovoltaic cells and solar heating panels

8.4.12 Distinguish between a photovoltaic cell and a solar heating panel.

Solar radiation can be used directly in two different ways: in **solar heating panels** the radiant (solar) energy is transferred to raise the internal energy of water; and in **photovoltaic cells** (also called photo cells or solar cells) the solar energy is transferred to electrical energy. Groups of photovoltaic cells are commonly called solar panels, but this can cause confusion with solar *heating* panels.

Photovoltaic cells

When radiation falls on the semiconducting material of a photovoltaic cell, free electrons are released within the material and a potential difference is produced across it.

When used in a circuit, both the current and the potential difference from a single cell tend to be low and there is also significant internal resistance. So, individually, they are only sources

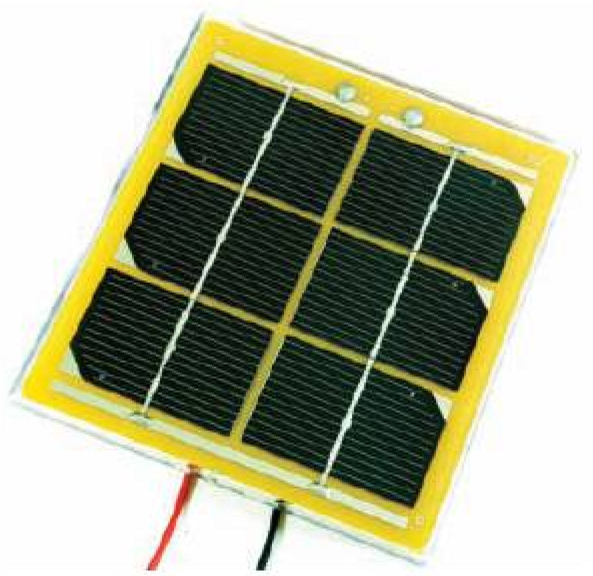


Figure 8.28 Photovoltaic cells

of small amounts of electrical power (see Figure 8.28). A large number can be connected together to provide much greater power. Figure 8.29 shows the energy transformation that occurs in a photovoltaic cell.

Photovoltaic cells are probably the world's fastest growing renewable energy technology. Costs are falling quickly and efficiencies are rising. Most solar cells operate at less than 20% efficiency, although the best are about 25%, with future hopes aiming at over 40%. Because of their potential convenience and low cost, a lot of research is going into the design of thin-film photovoltaic cells.

Figure 8.29 Energy flow for a photovoltaic cell

Figure 8.30 Germany has made significant investments in photovoltaic power

Photovoltaic cells are used (together with rechargeable batteries) mostly for low power devices like calculators and for isolated single devices (like emergency phones), but improving technology and lower costs now mean that they are also being used widely to provide supplementary electric power to homes and offices. Figure 8.30 shows a typical example. Solar power stations are also being built to provide 100 MW (or more) of power to electrical grid systems around the world.

Solar heating panels

In a solar heating panel the radiation from the Sun is absorbed by black-painted copper pipes through which water flows (Figure 8.31). The hot water is then pumped through a heat exchanger in which thermal energy is transferred from the hot water to a larger amount of water stored in an insulated tank. When the water in the tank is hot enough it can be used in the home for washing, etc. Homes with solar

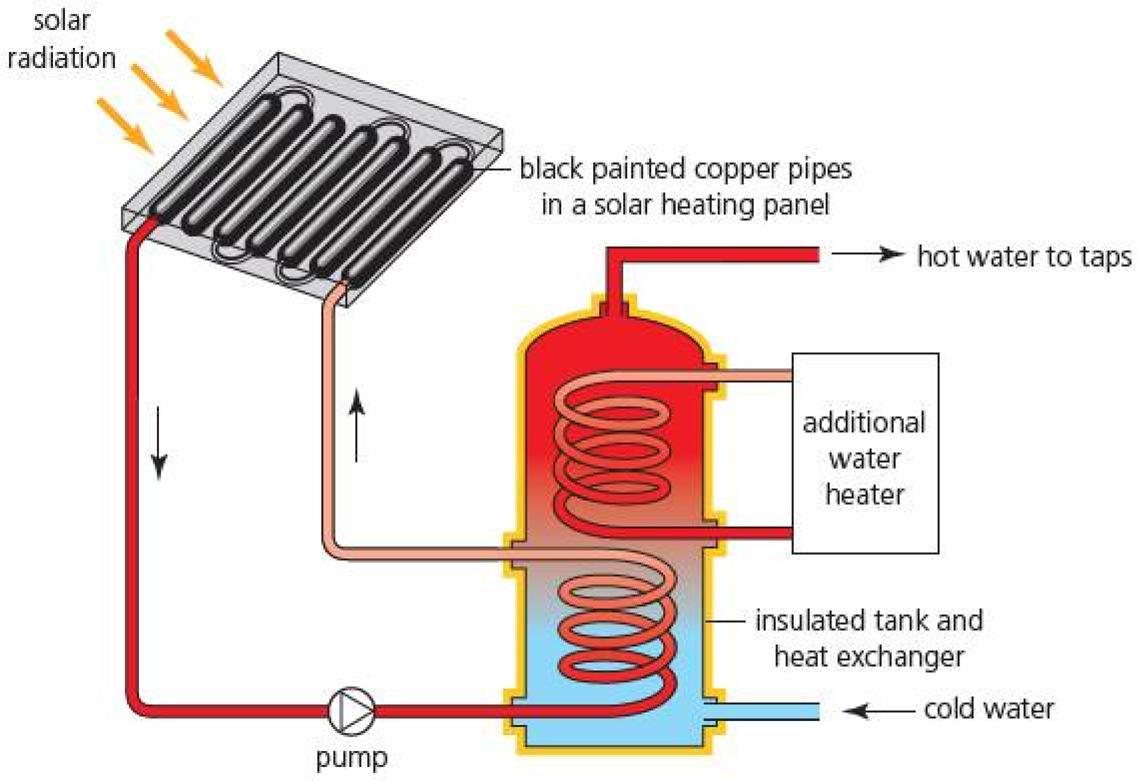


Figure 8.31 A solar heating system

Figure 8.32 Energy flow for a solar panel

Figure 8.33 Solar heating panels on homes in Turkey

heating panels will usually also have some other means of heating water for the times when the water is not hot enough. Figure 8.32 shows the energy transfer in a solar heating panel.

Solar heating panels are mostly used for individual homes and they are typically placed on a roof. They are used widely in many countries of the world, most usefully where there is a hot and/or a sunny climate (see Figure 8.33). Many governments encourage their use, or even require that all new homes have solar heating panels.

Variations in radiation intensity reaching the Earth's surface

8.4.13 **Outline** reasons for seasonal and regional variations in the solar power incident per unit area on the Earth's surface.

The intensity of the radiation reaching a particular place on the Earth's surface will depend on:

- the weather/climate
- the geographical latitude of the location
- time of day/night
- the angle of the surface to the horizontal at that location.

Figure 8.34 illustrates that if the same radiated solar power falls on areas A and B, the intensity will be less at A because the area is larger. This diagram could be representing places at different latitudes on the Earth or, for a given location, it could be representing variations in intensity at different times of the day or at different times of the year.

To reduce the effect of these variations, photovoltaic cells or solar heating panels are not usually placed horizontally on the ground but are positioned so that they receive radiation perpendicularly at a carefully chosen (average) time of the day and year. A more expensive option would be to provide the machinery necessary for the devices to move.

It is important to realize that if a panel were positioned so that it received the maximum radiation at midday in winter, the received power would still be less than that received at midday in summer. This is because the radiation has to pass through a greater length of the Earth's atmosphere, due to the tilt of the Earth. This can be seen by comparing the lengths P and Q in Figure 8.35.

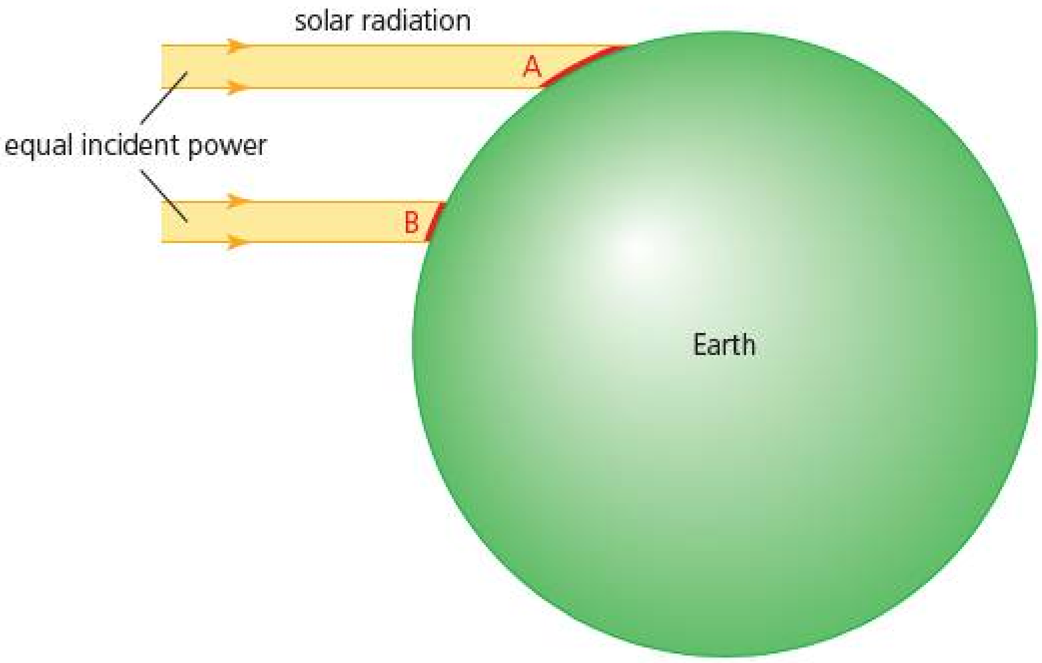


Figure 8.34 The received intensity of solar radiation varies with the angle of incidence

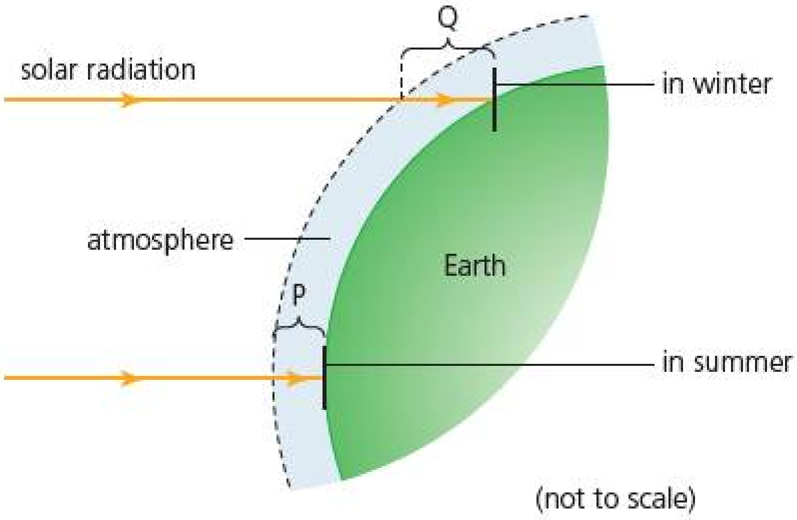


Figure 8.35 Effect of the atmosphere on received intensity of radiation at different times of year

Advantages and disadvantages of solar power

Table 8.4 Advantages and disadvantages of solar power

Advantages	Disadvantages
■ Renewable source of energy. ■ Free source of energy. ■ Pollution free during use. ■ Low maintenance costs, typical 20+ years lifetime.	■ Variation of output with time of day/night, weather, time of year, etc (meaning that individual systems need supplementary power/ batteries). ■ Photovoltaic cells and solar heating panels are expensive to construct and install (but they are becoming significantly cheaper). ■ Photovoltaic cells have some pollution issues during construction and end-of-use recycling. ■ Low energy density – large areas needed for photovoltaic power stations.

8.4.14 Solve problems involving specific applications of photovoltaic cells and solar heating panels.

30 a Calculate the total amount of energy radiated from the Sun that falls on the Earth's upper atmosphere in one year. (Assume that the distance from the centre of the Earth to the limit of the atmosphere is 6.5×10^6 m.)
b Compare your answer with the world's total annual energy consumption, estimated to be approximately 5×10^{20} J.

31 Figures 8.27 and 8.35 are not drawn to scale. The radius of the Earth is 6.4×10^6 m and the height of the atmosphere is often assumed to be 100 km. Use a compass to draw a scale diagram of the Earth surrounded by its atmosphere.

32 Make copies of Figure 8.34 and label them (a) variation of intensity with time of the day, and (b) variation of intensity with time of the year. On each diagram indicate the position of the Earth's axis and the North Pole. On (b) also label the direction of rotation.

33 Suggest two reasons why the value of the solar constant varies (by up to 7%).

34 Explain the ways in which the design of a solar heating system (with black copper pipes) shown in Figure 8.31 tries to maximize the amount of thermal energy transferred to the water in the tank.

35 a Draw an electrical circuit which would enable you to investigate how light intensity affects the potential difference across a photovoltaic cell and the current through it.
b How could you use the data to determine the internal resistance of the cell?

36 A photovoltaic cell of area $1.8\,cm^2$ is placed where it receives radiation at a rate of $700\,W\,m^{-2}$.
a What electrical power is produced if its efficiency is 18%?
b If the voltage across it is 0.74 V, calculate the current through the cell.
c How many cells would be needed for a power output of 50 W?
d What would be the total area of the cells?

37 a Use the Internet to get an up-to-date estimate of the world's total electrical power consumption.
b Use estimates of the average radiated power received on the Earth's surface, and typical areas and efficiencies of photovoltaic cells, to calculate an approximate value for the area of cells that would be needed to supply all of the world's electrical power needs.

38 Imagine that a solar heating panel is being placed in a fixed position on the roof of a home in your town/city. Discuss which would be the best direction for it to face.

39 A solar heating panel of area $3.4\,m^2$ is placed in a position such that on one day it received an average intensity of $640\,W\,m^{-2}$ during 12 hours of daylight.
a What is the total energy incident on the panel during that day?
b The system is designed to transfer this energy to a $0.73\,m^3$ tank of water and at the start of the day the temperature of the water in the tank was 2 °C. What is the maximum possible temperature of the water at the end of the day? (Assume that the density of water is $1000\,kg\,m^{-3}$ and that the specific heat capacity of water is $4200\,J\,kg^{-1}\,°C^{-1}$.)
c What assumption did you make?

40 a Make a copy of Figure 8.36, which shows how the total energy received per day by a fixed solar panel (or solar cell) is predicted to vary during the year at a location with a latitude of 30° in the northern hemisphere.

b Add lines to your graph to show the energy/day that would be predicted for the same panel if it was **i** at a latitude of 60° N, **ii** at a latitude of zero, on the equator, and **iii** at a latitude of 30° S.

c Give two reasons why the total energy received per day decreases with increasing latitude.

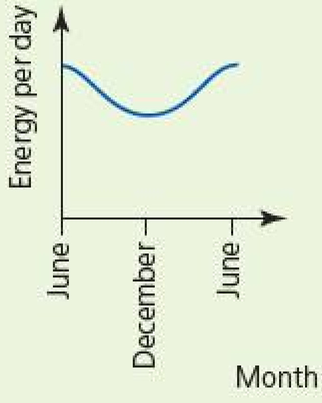


Figure 8.36

Hydroelectric power

8.4.15 Distinguish between different hydroelectric schemes.

When water is free to flow downwards, gravitational potential energy is transferred into kinetic energy, which can be used to drive turbines and generate electrical energy. Electrical power generated in this way is known as **hydroelectric power**.

In order for this idea to be a useful long-term source of electricity, there must be a process (*cycle*) continually involving the transfer of the energy needed to return the water to a higher level. The following three hydroelectric schemes can be identified.

- **Water storage in lakes**. Water is continuously evaporating from the oceans and seas due to the transfer of thermal energy from the Sun. The water vapour rises due to convection currents, forms clouds and later the water falls as rain (or snow), which can be collected and stored at high altitude in lakes or in artificial reservoirs behind dams.

  When the water is allowed to drop downwards, often nearly vertically, electrical energy can be generated by turbines at the bottom of the fall (see Figure 8.37). It is also possible to generate smaller amounts of electrical power directly from fast flowing rivers. Some of the most powerful power stations in the world are hydroelectric, for example the Itaipu hydroelectric power station on the Brazil–Paraguay border (Figure 8.38), but there are also a very large number of small-scale schemes providing electrical power for small communities in remote locations. Approximately 20% of the world's electricity is generated by hydroelectric power stations and it is the world's most widely used source of renewable energy.

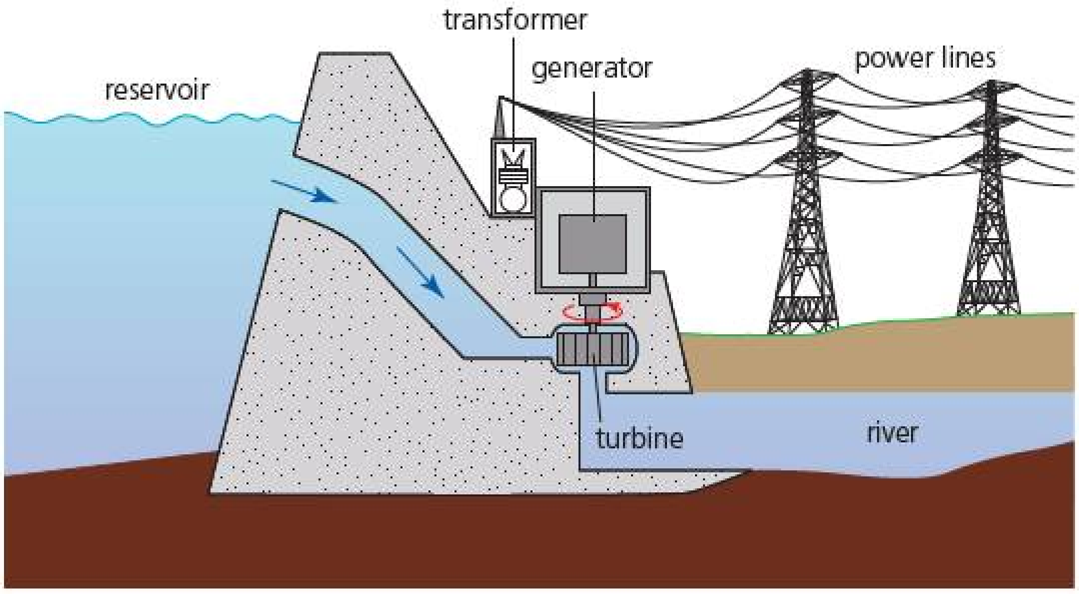


Figure 8.37 Cross-section of a hydroelectric power station

Figure 8.38 The Itaipu hydroelectric power station on the Brazil–Paraguay border

- **Tidal water storage.** As the Earth spins on its axis every day, the gravitational attraction of the Moon (and the Sun) raises the level of the oceans on the side of the Earth closest to the Moon and also on the opposite side. This means that the level of the seas and oceans rise and fall approximately twice every day. If water can be stored behind a barrage (dam) at high tide, it can be released at low tide to generate electricity, as shown in Figure 8.39. Energy can also be transferred when water flows the other way.

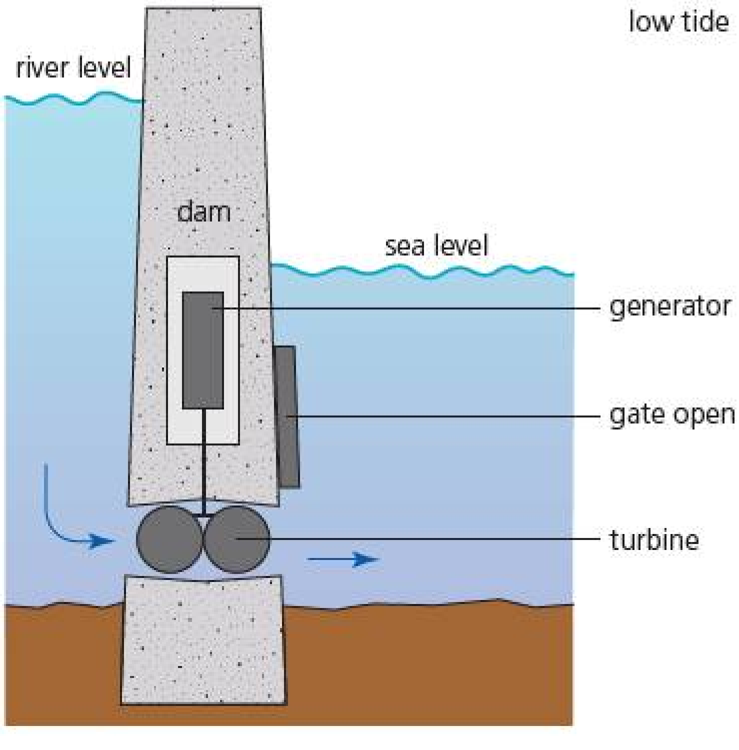


Figure 8.39 Cross-section of a tidal barrage scheme

The difference between the water levels at high tide and at low tide varies considerably at different locations around the world and there are not many places where it is worth the considerable expense of building a tidal power station. However, there are many locations where there are strong tidal currents in the oceans and it is possible to generate electricity from the kinetic energy of the water using underwater turbines.

- **Pump storage.** In hilly locations, hydroelectric power offers a large-scale solution to the storage of energy from an electrical power station. Electrical power stations are designed to be operated at a particular power output at which they have their maximum efficiency, but at night, when most people are asleep, there is obviously less demand for electrical energy (see Figure 8.40). If power stations are operated at a lower power at night, their efficiency is reduced and a higher proportion of energy is wasted as thermal energy transferred into the surroundings. It is better to keep a power station running at approximately the same power 24 hours a day by storing the excess energy generated at night and transferring it back into the system during the day. One way of doing this is to use the excess power to pump water up to a lake at a greater height during the night and then release it to generate extra power the next day.

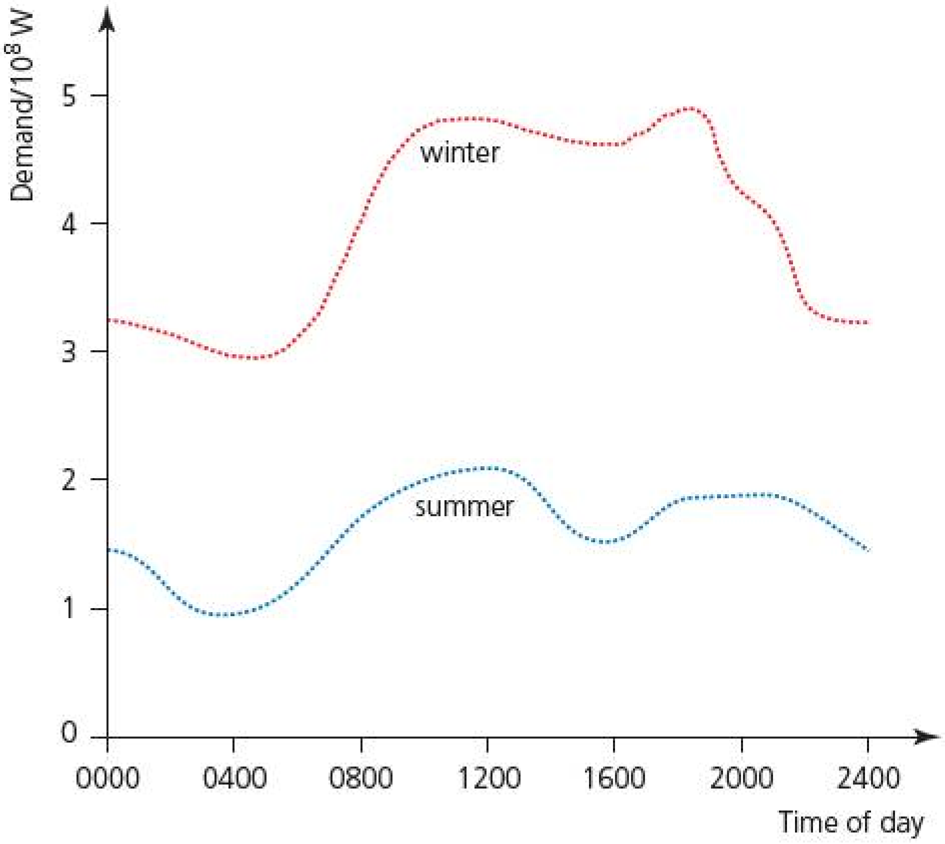


Figure 8.40 Typical daily variation in energy consumption from a large power station in summer and winter

Energy transfers in hydroelectric schemes

8.4.16 Describe the main energy transformations that take place in hydroelectric schemes.

The energy transformations that take place in a hydroelectric power station are shown in Figure 8.41.

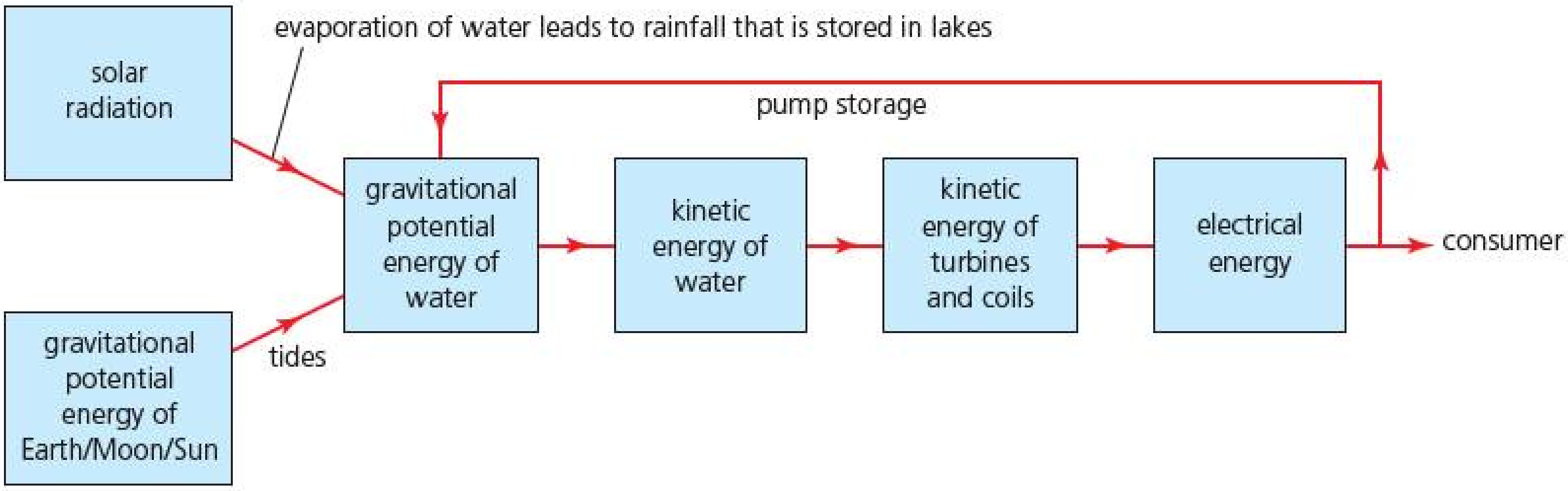


Figure 8.41 Energy flow diagram for hydroelectric power stations

Advantages and disadvantages of hydroelectric power

Table 8.5 Advantages and disadvantages of hydroelectric power

Advantages	Disadvantages
■ Renewable source of energy. ■ No greenhouse gas emissions (although there will probably be some increase in the release of methane from the enlarged lakes). ■ Free source of energy. ■ No significant pollution during operation. ■ Dams may also be used to control river flow, improve irrigation and prevent flooding. ■ Newly created lakes can be a recreational resource and provide a new habitat for some plants and animals. ■ Can be the ideal energy resource for remote, hilly locations.	■ The environment will be affected and the natural habitat of many plants and animals may be destroyed. ■ Newly built dams will form lakes which may cover land which was previously villages, towns, farmland, etc. ■ Can only be used in certain locations (mountainous/hilly places, places with high tides or powerful tidal flows). ■ Large-scale projects can be very expensive to construct, although maintenance is more reasonable. ■ The natural flow of rivers may be interrupted, which can have many undesirable consequences. ■ Hydroelectric power station may be a long way from centres of population, so that the power needs to be transmitted large distances. ■ If a dam bursts it can cause considerable damage and loss of life.

Calculations involving hydroelectric schemes

8.4.17 Solve problems involving hydroelectric schemes.

The loss of gravitational energy when a mass m of water falls a distance Δh can be determined from $\Delta E_P = mg\Delta h$.

The average theoretical power output during time Δt will be $\frac{\Delta E_P}{\Delta t}$.

Because they do not rely on the transfer of thermal energy, hydroelectric power stations will be much more efficient that fuel-burning power stations. Typically, the larger the power station, the more efficient it will be.

Worked example

2 What is the average output power from a small hydroelectric power station operating at 92.0% efficiency if 242 000 kg of water pass through its turbines every hour, having fallen a vertical distance of 82.4 m?

Energy transferred from falling water in one hour $\Delta E_P = mg\Delta h = 242\,000 \times 9.81 \times 82.4$
$= 1.96 \times 10^8\,\text{J}$

$$\text{Maximum theoretical power} = \frac{\Delta E_P}{\Delta t} = \frac{1.96 \times 10^8}{3600} = 5.43 \times 10^4\,\text{W}$$

Output power = input power × efficiency = $5.43 \times 10^4 \times 0.92 = 5.00 \times 10^4\,\text{W}$

41 Calculate the maximum theoretical power available from a hydroelectric power station in which 2.0 m^3 of water falls a vertical distance of 112 m every second. Assume that the density of water is 1000 kg m^{-3}.

42 What is the total gravitational potential energy available from a lake of area 5.8 km^2 and average depth of 33 m which supplies a hydroelectric power station that is 74 m below the bottom of the lake? Think carefully about what height you should use in the calculation.

43 Water flows from a lake through a certain hydroelectric power station at an average annual rate of 4.3 $m^3 s^{-1}$.
a If the water in the lake comes from the rainfall over an area of 138 km^2, what is the average rainfall (in cm y^{-1}) over that area?
b What assumption did you have to make? Is it reasonable?

44 A family wants to install a small hydroelectric generator to provide for their own electrical needs, which they estimate to be an average of 3 kW. The generator's efficiency is 85%. If their home is at an altitude of 1420 m and they want to use water falling from a small lake at an altitude of 1479 m, what average mass of water must pass through the generator every second?

45 Make a list of the desirable features of a possible location for a new hydroelectric power station.

46 Suggest some 'undesirable consequences' which may occur if the natural flow of a large river is interrupted by one or more dams.

47 In a pumped storage system, water is pumped up to a lake/reservoir using electrical energy and then later allowed to fall back to generate electricity again. Of course some energy is wasted in this process. Explain why this process can still be a good idea.

48 Figure 8.39 shows a simplified sketch of a tidal power station when the sea is at low tide, which is 8.9 m below high tide.
a Remembering that the water can flow both ways, sketch a graph to show how the power output from the system might vary over a period of 24 hours.
b If the water behind the dam covers an area of 23 km², calculate the gravitational potential energy stored in it.


Figure 8.42 Traditional windmills in Greece


Figure 8.43 A wind farm in Denmark

Wind power

8.4.18 **Outline** the basic features of a wind generator.

Radiation from the Sun causes differences in temperature that result in changes in air density. These changes produce convection currents and differences in air pressure. Wind is air moving between areas of high pressure and low pressure. For thousands of years the energy transferred from windmills has been used around the world for work such as grinding crops (see Figure 8.42). There is an enormous amount of kinetic energy in the winds across the planet, but it has low energy density and it is only in recent years that technological advances have encouraged the widespread use of large wind-driven electrical generators. It has now become a very rapidly expanding technology.

There are many different designs of **wind generators**, but we will concentrate our attention on the type most commonly used for large scale electrical power generation. They have a horizontal axis of rotation, as shown in Figure 8.43.

The blades (usually there are three) move in a vertical circle and are designed to be struck by wind moving parallel to the Earth's surface. The kinetic energy of the blades is then transferred to make coils spin in electrical generators. The number of blades, their width and angle to the wind are all carefully chosen to get the maximum amount of power transferred from the wind. For the same reason, the whole blade mechanism is often able to rotate in the horizontal plane so that the wind always strikes the blades perpendicularly. The blades are designed to rotate at a particular speed with winds within a certain range, typically about 5–15 m s^{-1}. Figure 8.44 shows the energy transformations involved in wind power.

Winds flowing close to the ground will lose energy due to friction and from striking obstacles, so generators are usually located near to coasts or on tree-less hills and the towers supporting the rotating blades are as tall as possible. Placing wind generators in shallow water off-shore has obvious advantages, but these are expensive to construct. The largest off-shore wind farms (already operational and those in construction) are off the shores of the UK and Denmark.

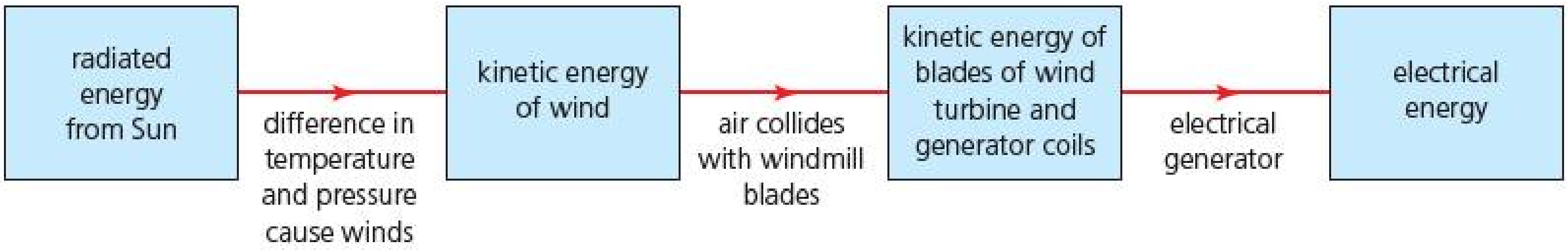


Figure 8.44 Energy flow diagram for wind generators

Smaller wind generators may be ideal for remote locations, for example on boats or farmhouses, which have no access to mains electricity.

Advantages and disadvantages of wind power

Table 8.6 Advantages and disadvantages of wind power

Advantages	Disadvantages
■ No greenhouse gas emissions. ■ Renewable source. ■ Free source of energy. ■ No pollution during use. ■ Small generators are ideal for remote locations.	■ Expensive to construct. ■ Low energy density, large area needed (although the land around them can still be used for farming). ■ Wind speed (and power output) is unreliable. ■ Emits some noise. ■ Best locations are often far away from cities and towns. ■ Many people consider that they are ugly and spoil the environment.

Calculations involving wind power

8.4.19 **Determine** the power that may be delivered by a wind generator, assuming that the wind kinetic energy is completely converted into mechanical kinetic energy, and explain why this is impossible.
8.4.20 **Solve** problems involving wind power.

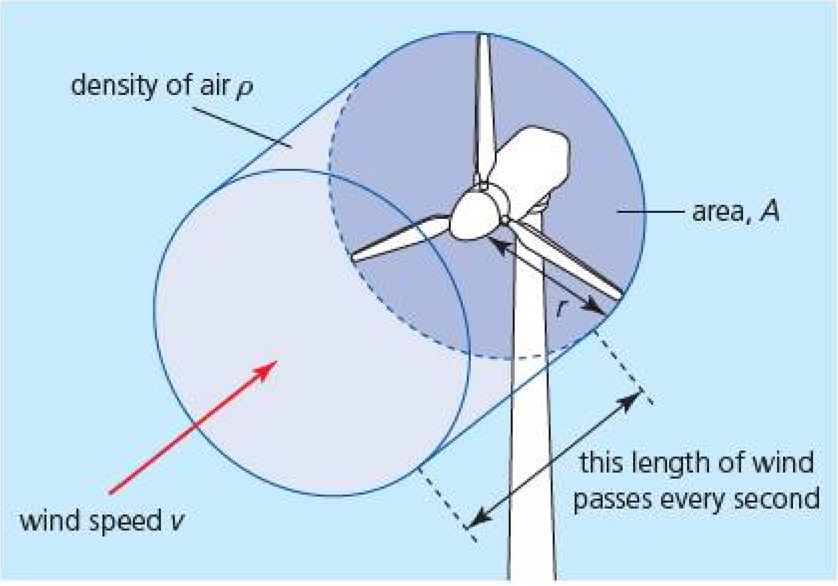


Figure 8.45 Cross-section of wind striking the blades of a wind turbine

An equation for the power of a wind generator (Figure 8.45) can be obtained by considering the loss of kinetic energy of the wind striking the blades.

The volume of air passing the blades every second, V, is equal to the speed of wind, v, multiplied by the area 'swept out' by the rotating blades, A.

$$V = vA$$

Since mass = volume × density, ρ, the mass m passing every second is given by:

$$m = vA\rho$$

Using the equation for kinetic energy $\left(E_K = \frac{1}{2}mv^2\right)$, the kinetic energy of this moving mass of air passing the blades every second is:

$$E_{K(air)} = \frac{1}{2} \times vA\rho \times v^2$$

We know that:

$$\text{power} = \frac{\text{energy transferred}}{\text{time}}$$

So, if all of the wind's kinetic energy is transferred to the wind generator (which would mean that the air behind the generator is not moving), then its maximum theoretical power P is given by:

$$P = \frac{1}{2}A\rho v^3$$

This equation is given in the IB *Physics data booklet*.

It is not surprising that the power is proportional to the area and density, but the fact that it depends on the wind speed *cubed* should be noted. A doubling of the wind speed could theoretically produce eight (2^3) times the power.

The transfer of kinetic energy of the air to kinetic energy of the blades cannot be 100% efficient, considering that some of the wind must pass between the blades and that the air which strikes the blades will not come to rest.

49 Explain what it means to say that wind power has a low 'energy density'.

50 A wind generator has blades of length 18 m and operates with an efficiency of 21%. What power output would you expect with a $12\,\text{m s}^{-1}$ wind? (Density of air = $1.3\,\text{kg m}^{-3}$.)

51 A wind generator has an output of 10 kW when the wind speed is $8\,\text{m s}^{-1}$.
a What wind speed would be needed for an output of 20 kW?
b What power output would you predict for another generator of similar design, but with blades of twice the length, when the wind speed was $16\,\text{m s}^{-1}$?

52 A small farmhouse needs an average power of 4 kW. A wind generator is to be placed on the top of a tower where the effective average wind speed is $7.5\,\text{m s}^{-1}$. Calculate the length of blades needed to generate this power if the system is 24% efficient.

53 Large wind generators can produce 5 MW or more of electrical power.
a Make an estimate of the diameter of the circle through which the blades rotate. List the assumptions you made.
b A total output of 20 MW is required. Discuss whether it might be better to build four similar generators or one generator of twice the size.

54 **a** How many large wind generators (as described in question **53**) would be needed to provide 2 GW (equivalent to a large fossil-fueled power station)?
b If the wind generators have to be at least 300 m apart, what is the minimum area of land needed?
c Suggest a reason why the generators need to be so far apart.

55 Wind speeds vary considerably and quoting an 'average' value may be misleading, especially if a value for the average speed is used directly to predict an average power, as in question **52**. Explain why this would almost certainly lead to an underestimate of the available power.

56 Draw a Sankey diagram to represent the energy transfers involved with a wind generator.

Wave power

8.4.21 Describe the principle of operation of an oscillating water column (OWC) ocean-wave energy converter.

Winds flowing across water surfaces disturb the water and create waves. The transverse waves of the oceans and seas transfer an enormous amount of energy, which is usually dissipated into the environment. It is not difficult to design wave power-driven electrical generators and many different possibilities have been tested over the years. However, it has been difficult to get a high enough overall efficiency to justify the high installation and maintenance costs. The best locations, with the largest waves, are off-shore, but oscillating water column (OWC) generators, such as that as shown in Figure 8.46, can be built both off-shore and on-shore.

As the crest of a wave enters the chamber it pushes air out past the turbine blades and a potential difference is generated. When a trough of a wave enters the chamber, air is forced to flow in the opposite direction and a potential difference in the reverse direction can be generated.

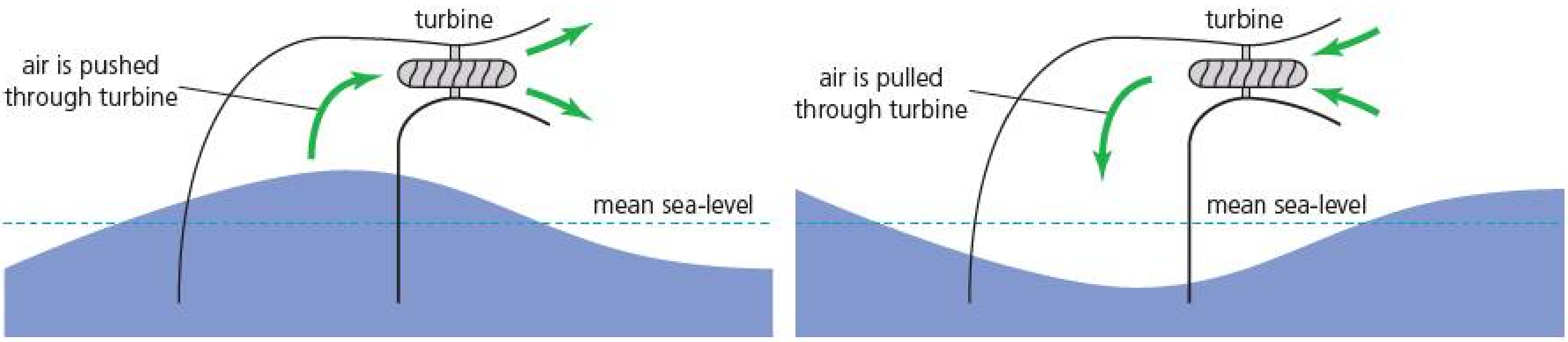


Figure 8.46 Simplified diagram of a OWC generator

Figure 8.47 shows a large OWC generator built in Australia. Figure 8.48 shows the energy transformations involved with wave power.

Figure 8.47 An OWC generator in Australia

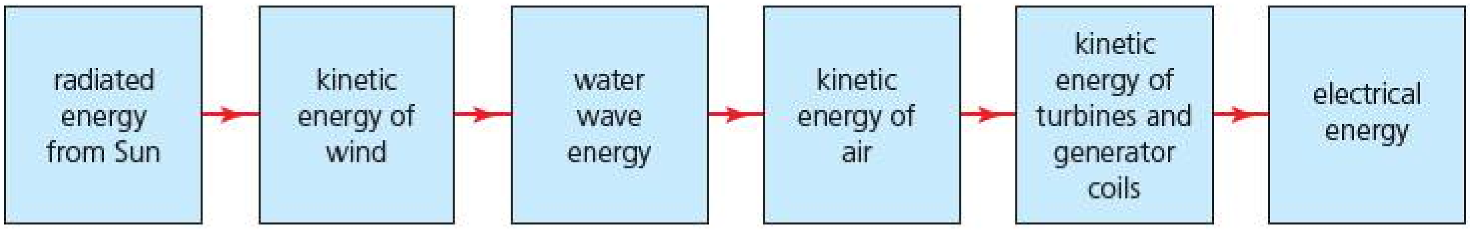


Figure 8.48 Energy flow for a wave power generator

Advantages and disadvantages of wave power

Table 8.7 Advantages and disadvantages of wind power

Advantages	Disadvantages
■ Renewable energy source. ■ Free energy source. ■ Non-polluting. ■ No greenhouse gas emissions. ■ Enormous amounts of energy are transferred by ocean waves and the energy density is reasonable for a renewable source.	■ Very expensive to construct. ■ Only worthwhile in places that regularly have large waves. ■ The damage caused by large waves means that maintenance costs are high. ■ Possible locations will often be a long way from centres of population. ■ The size of the waves and the power output will be unreliable. ■ Overall efficiency is low.

Calculating the energy transferred by water waves

8.4.22 **Determine** the power per unit length of a wavefront, assuming a rectangular profile for the wave.
8.4.23 **Solve** problems involving wave power.

To simplify our understanding, we need to assume that all the waves have the same regular shape. A sine wave would be the obvious choice, but we will choose a square wave, as shown in Figure 8.49, because it simplifies the mathematics.

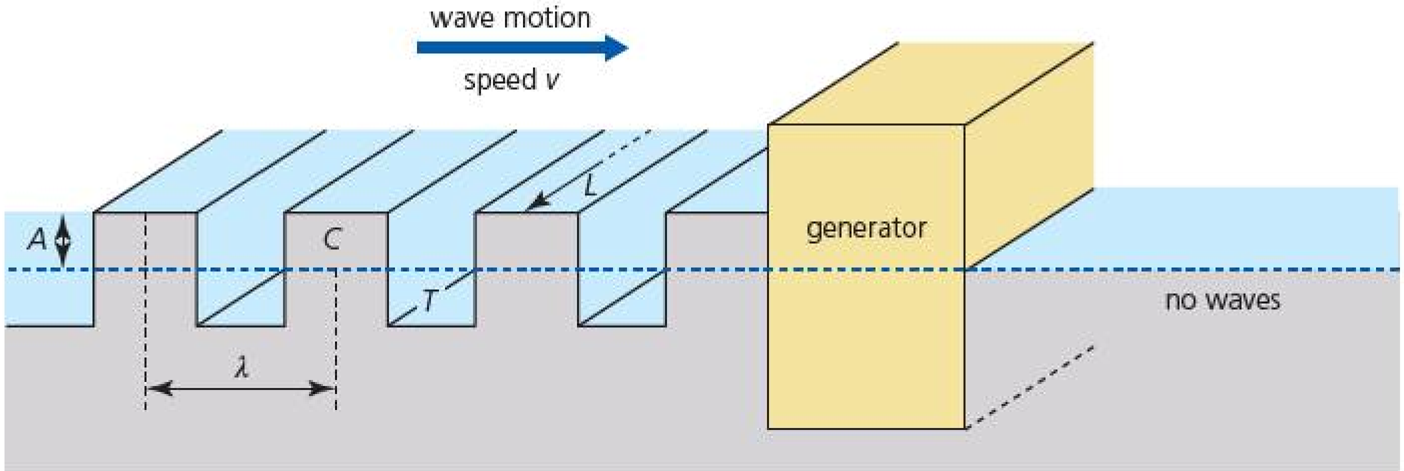


Figure 8.49 Representing water waves by a square waveform

In Figure 8.49 the waves are moving to the right at speed v and entering an OWC generator. If all the energy was removed from the waves, then there would be no waves to the right of the generator. To determine the maximum theoretical energy of the waves we need to calculate the difference in gravitational potential energy before and after passing through the generator. The easiest way to think about this is by considering that all the top halves of the waves (like the crest C) were moved to fill in the lower gaps (like the trough T).

Volume of water in a crest of length (L), $V = \frac{A \times \lambda}{2 \times L}$

Mass of water in this volume equals the volume multiplied by density (ρ), $m = \frac{\rho A L \lambda}{2}$

The water falls through a height, $\Delta h = A$

So that the loss of gravitational potential energy $= mg\Delta h = \left(\frac{\rho A L \lambda}{2}\right) \times gA = \frac{A^2 \rho g \lambda L}{2}$

To determine the maximum theoretical power available we need to multiply this energy by the number of waves passing though the generator every second, which is the frequency of the waves ($f = v/\lambda$):

$$\text{power} = \left(\frac{A^2 \rho g \lambda L}{2}\right) \times \frac{v}{\lambda} = \frac{A^2 \rho g v L}{2}$$

Or, power per unit length $= \frac{1}{2} A^2 \rho g v$

This equation is given in the IB *Physics data booklet*.

Worked example

3 What is the wave power carried by 1 km long ocean waves of amplitude 1.9 m moving at 2.5 m s^{-1} (density of sea water = 1025 kg m^{-3})?

Power per metre $= \frac{A^2 \rho g v}{2} = \frac{1.9^2 \times 1025 \times 9.81 \times 2.5}{2} = 4.5 \times 10^4$ W

Power per km $= 4.5 \times 10^4 \times 10^3 = 4.5 \times 10^7$ W

57 The equation used above to calculate wave power was derived assuming that the wave form was square. Because the actual waveform of water waves is more like a sine wave, do you think that the power predicted by the equation is too high or too low? Explain your answer.

58 Waves of average amplitude 0.46 m are moving towards a 1.6 km long beach at a rate of 10 waves every 60 s.

a What is the frequency of the waves?
b If their wavelength is 1.9 m, what is the wave speed?
c Calculate the wave power per metre delivered by the waves to the beach.
d What is the total wave energy delivered in one hour?
e Suggest what happens to all this energy.

59 The chamber on an OWC generator allows waves of width 3.4 m and amplitude 1.25 m to enter. If the waves travel at a speed of 1.6 m s^{-1} what is the output of the power station if its efficiency is 30%?

60 The following list contains some of the many questions that might be asked about different types of power station. (Sometimes the questions do not have easy or straightforward answers.)

- Is the energy source renewable?
- Can it provide large amounts of power to large numbers of people?
- Are there any greenhouse gases emitted?
- Is the source of energy free, or expensive?
- Is it expensive or difficult to transfer the energy source to the power station?
- Does the source have a high energy density?
- Is the supply of energy reasonably constant and predictable?
- Is the process energy efficient?
- Is there any pollution during operation?
- Is there any significant pollution during construction or disposal?
- Are there any other adverse effects on the environment or people?
- Is it difficult or expensive to construct and install?

- Is the equipment expensive to maintain?
- How long before the equipment needs replacing?
- Are there plenty of suitable locations for power stations of this type?
- Are the probable locations of power stations close to towns and cities?

a Choose five of these questions which you consider to be the most important.
b Make a chart or spreadsheet of the questions in a vertical column and a horizontal row of the six types of power stations that have been discussed in detail in this topic (fossil, nuclear, solar, hydroelectric, wind and wave). Place a suitable comment in each box of the chart, comparing the advantages and disadvantages of the different choices.

8.5 The greenhouse effect

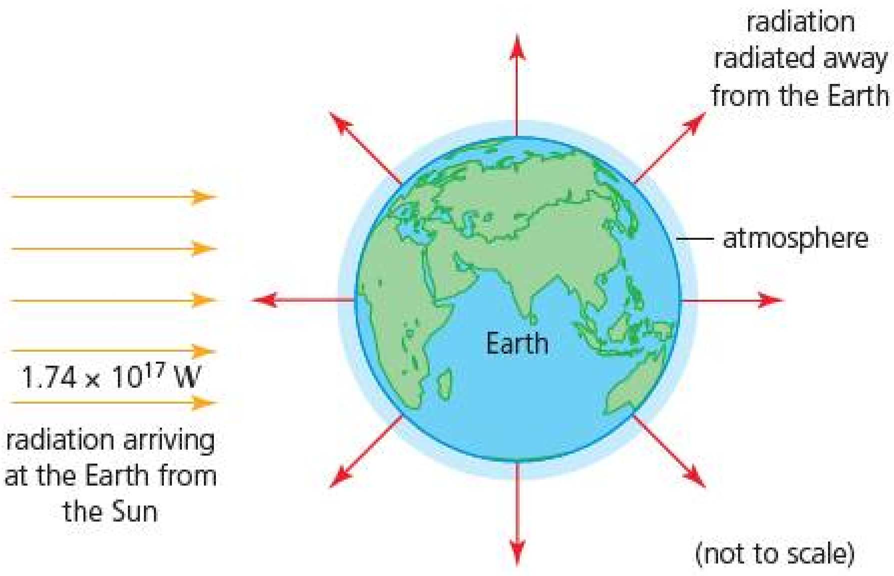


Figure 8.50 Earth receiving and emitting radiation

The Earth receives an average of about 1.74×10^{17} J of radiated energy from the Sun every second (Figure 8.50). If the Earth reflects or radiates energy back into space at the same rate, then it will be in **thermal equilibrium** and its temperature will remain constant. If the Earth reflects or radiates less energy than it receives it will get hotter because the energy will be absorbed by the atmosphere. If it reflects or radiates more than it receives, it will become cooler. Although there have been some natural changes slowly occurring in the Earth's temperature in the distant past, it will be difficult for the people of the world to adjust to the more rapidly changing temperatures and climate change that most scientists are predicting.

Solar radiation

8.5.1 **Calculate** the intensity of the Sun's radiation incident on a planet.

The Sun emits radiation at an average rate, P, of 3.84×10^{26} W (this is the Sun's power, also called its **luminosity**). If we assume that this radiation spreads out equally in all directions without absorption, then at a distance r from the Sun, the same power is passing through an area $4\pi r^2$ (the surface area of an imaginary sphere), as shown in Figure 8.51.

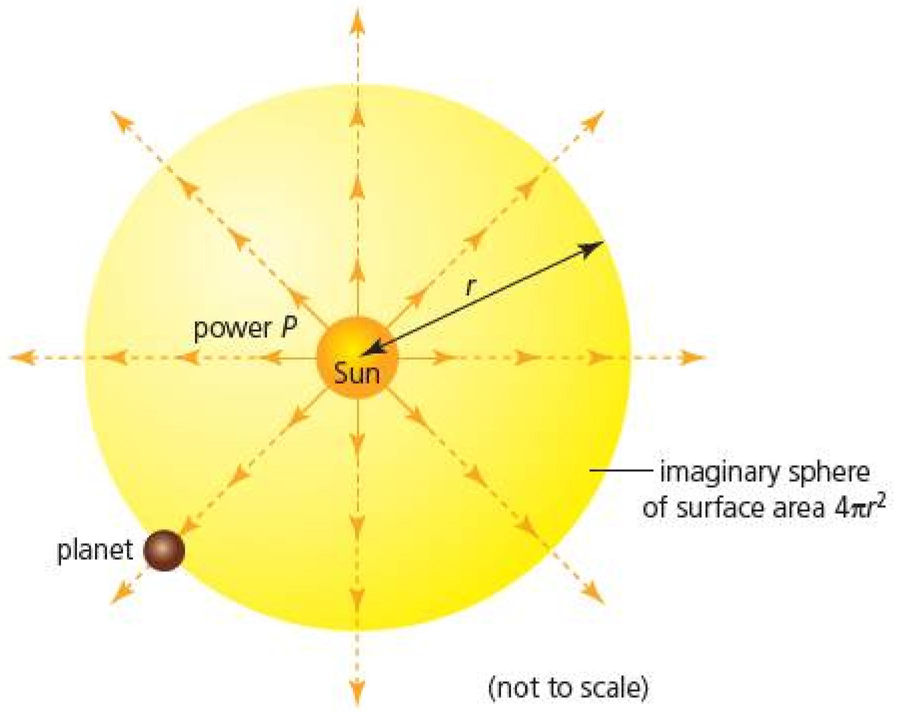


Figure 8.51 The Sun's radiation spreading out

The intensity at different distances, r, from the Sun can be calculated from:

$$I = \frac{P}{4\pi r^2}$$ Intensity is inversely proportional to the square of the distance from the source.

Worked example

4 Calculate the intensity of the Sun's radiation arriving at:

a the Earth

b Mercury

a Earth is an average distance of 150×10^6 km from the Sun, so

$$I = \frac{P}{4\pi r^2} = \frac{3.84 \times 10^{26}}{4 \times \pi \times (150 \times 10^9)^2}$$

$I = 1.36 \times 10^3\,\text{W m}^{-2}$ ($\approx 1400\,\text{W m}^{-2}$)

b Mercury is an average distance of 58×10^6 km from the Sun, so

$$I = \frac{P}{4\pi r^2} = \frac{3.84 \times 10^{26}}{4 \times \pi \times (58 \times 10^9)^2}$$

$I = 9.1 \times 10^3\,\text{W m}^{-2}$

The result for Earth calculated in Worked example 4 was used earlier in this chapter when discussing solar power. It represents the average intensity falling on an area above the Earth's atmosphere which is perpendicular to the direction in which the radiation is travelling. It has been measured on orbiting satellites to have an average value of $1366\,\text{W m}^{-2}$.

Reflected radiation: albedo

8.5.2 **Define** *albedo*.
8.5.3 **State** factors that determine a planet's albedo.

On average about 30% of the radiation incident on the Earth is directly reflected or scattered back into space, but there are considerable variations from place to place and at different times of the day and year.

The ratio of the total scattered or reflected power to the total incident power is known as the **albedo**.

$$\text{albedo} = \frac{\text{total scattered or reflected power}}{\text{total incident power}}$$

This equation is given in the IB *Physics data booklet*.

Figure 8.52 Snow has high albedo but water has low albedo

Albedo is sometimes given the symbol α. An albedo of 1 (100%) would mean that all radiation was scattered or reflected. An albedo of zero would mean that no radiation was scattered or reflected. White and bright surfaces (like snow, ice and clouds) have a high albedo because they reflect a lot of the incident radiation. Darker surfaces (like water) have a lower albedo as they absorb the incident radiation (Figure 8.52).

Albedo also varies with the angle of incidence (which changes with inclination of the surface, season, latitude and time of the day). The Earth's average albedo is approximately 0.3 (30%).

61 Use the information on page 295 and the radius of the Earth (6.37×10^6 m) to show that the total radiated energy arriving at the Earth from the Sun is 1.74×10^{17} J.

62 The total radiated power arriving from a Sun at a planet is 2.82×10^{17} W. The orbit of the planet around the Sun has a radius of 1.23×10^7 m.
a Calculate the intensity of the radiation arriving at the planet (perpendicular to the direction of the radiation).
b If the average albedo of the planet is 0.22, calculate the average intensity of radiation absorbed by the planet.

63 One of the planets in our solar system receives solar radiation with an intensity of about 15 W m^{-2}. Calculate its distance from the Sun and find out which planet it is.

64 A small light bulb is the only lamp in an otherwise dark room. It emits visible radiation with a power of 5 W. What is the intensity received at a book placed 90 cm from the bulb?

65 Suggest possible reasons why the average albedo at a particular place is
a different in winter than in summer
b different at different times of the same day.

66 Two planets are orbiting a distant star. Planet A is 100 million kilometres from the star and receives an intensity of 860 W m^{-2}. What is the intensity received at planet B if it is 120 million kilometres from the star?

The greenhouse effect

8.5.4 **Describe** the greenhouse effect.

The **greenhouse effect** is the name that has been given to the *natural* effect a planet's atmosphere has in increasing the temperature of the planet to a value higher than it would be without an atmosphere.

The Earth has had a beneficial greenhouse effect as long as it has had an atmosphere and if the Earth did not have an atmosphere it would be about 30°C cooler. However, most scientists now believe that the greenhouse effect has become **enhanced** (increased). It is believed that this is mainly because human activity has changed the composition of the atmosphere.

If the Earth *did not* have an atmosphere, the average intensity of radiation received from the Sun over the whole of the Earth's surface (over an extended period of time) could be calculated as follows:

$$I = \frac{\text{total power received by Earth from the Sun}}{\text{total surface area of the Earth}} = \frac{1366 \times \pi r_E^2}{4\pi r_E^2}$$

where r_E is the radius of the Earth, 6.37×10^6 m.

$$I = 342\ \text{W m}^{-2}$$

But the Earth *has* an atmosphere. Of the 342 W m^{-2} incident on the Earth and its atmosphere, 107 W m^{-2} is immediately reflected back into space (77 W m^{-2} from the atmosphere and 30 W m^{-2} from the Earth's surface). These values give the albedo of the Earth as $107/342 \approx 30\%$, which agrees with the figure quoted earlier for the average albedo of the Earth.

This leaves a global average of 235 W m^{-2} absorbed by the Earth and its atmosphere. Figure 8.53 shows what happens to this energy: 67 W m^{-2} is absorbed by the atmosphere and 168 W m^{-2} is absorbed by the Earth's surface.

(Actual values are quoted, although these figures may vary slightly depending on the sources of information used; these do not need to be learnt.)

For the Earth to remain in thermal equilibrium it is necessary for the incident 235 W m^{-2} to be matched by an equal intensity radiated back into space.

Radiation is the only way that thermal energy can travel away from the Earth through space. (But thermal energy is transferred from the Earth's surface to the atmosphere by a number of other processes as well, involving conduction, convection and evaporation.) The atmosphere absorbs a much higher percentage of this outgoing radiant energy than it does incoming radiant energy from the Sun. This is because the Earth is much cooler than the Sun and the radiation it emits is mostly at lower wavelengths (all infrared, rather than visible) and the greenhouse gases in the atmosphere absorb some of these infrared wavelengths.

The energy absorbed by the atmosphere is also transferred – the radiated energy is emitted in all directions and some of it goes directly into space. However, most of the energy is transferred back down to the Earth's surface (by radiation and other processes) and so the process continues, and thermal equilibrium is established at a certain temperature. (On average this is about 15 °C at present on the Earth's surface.) Figure 8.53 attempts to simplify and summarize these complex processes.

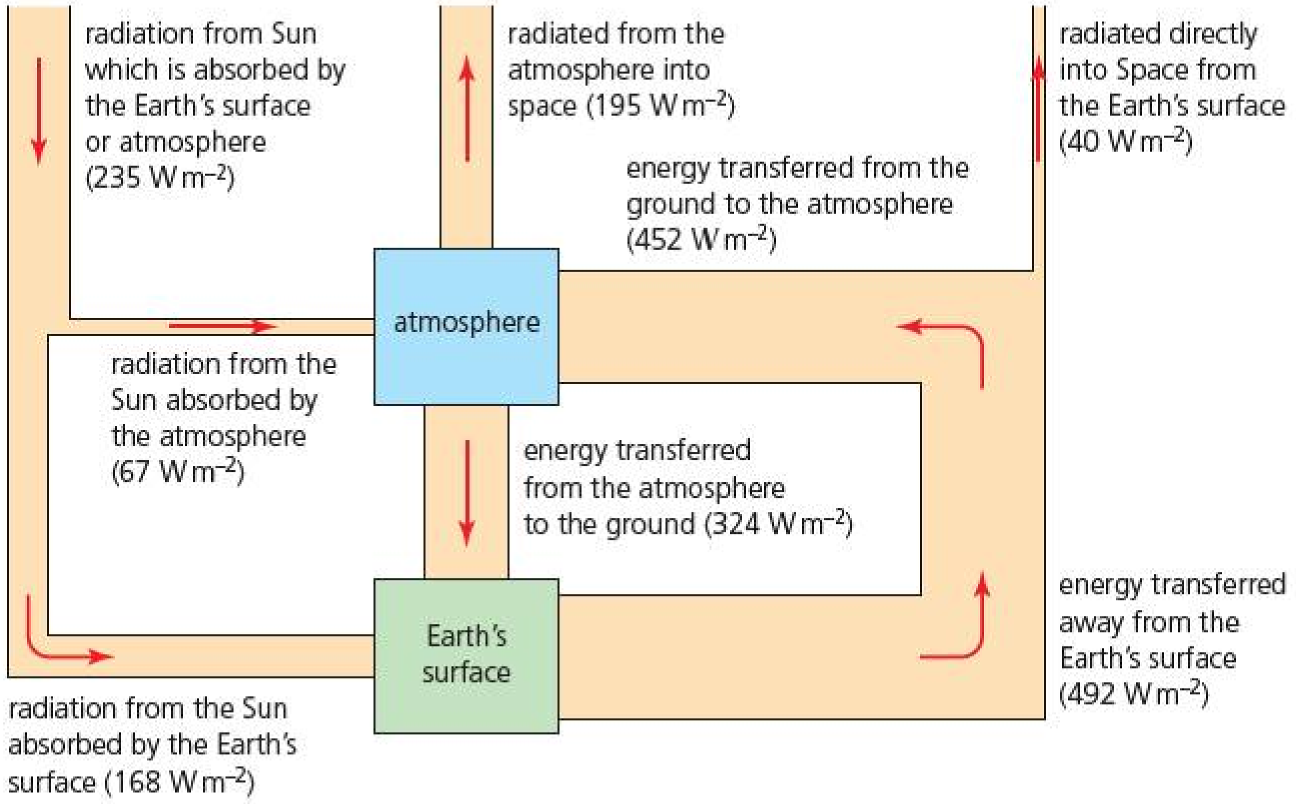


Figure 8.53 Energy flow through the Earth's atmosphere

Greenhouse gases

8.5.5 **Identify** the main greenhouse gases and their sources.
8.5.6 **Explain** the molecular mechanisms by which greenhouse gases absorb infrared radiation.
8.5.7 **Analyse** absorption graphs to **compare** the relative effects of different greenhouse gases.

The Earth's atmosphere has been formed over millions of years by naturally occurring volcanic and biological processes and from collisions with comets and asteroids. The air in the atmosphere contains approximately 78% nitrogen, 21% oxygen and 0.9% argon. There are also naturally occurring traces of many other gases, including carbon dioxide and water vapour. Some of these trace gases are called **greenhouse gases** because they play a very important part in controlling the temperature of the Earth and the greenhouse effect. Greenhouse gases absorb (and emit) infrared radiation. Nitrogen, oxygen and argon have no greenhouse effect (because they are non-polar).

There are many greenhouse gases, but the four most important are, in order of their contribution to the greenhouse effect:

- water vapour
- carbon dioxide
- methane
- nitrous oxide (dinitrogen monoxide).

The relative importance of these gases in causing the greenhouse effect depends on their relative abundance in the atmosphere as well as their ability to absorb infrared radiation. Each of the gases has natural as well as man-made origins.

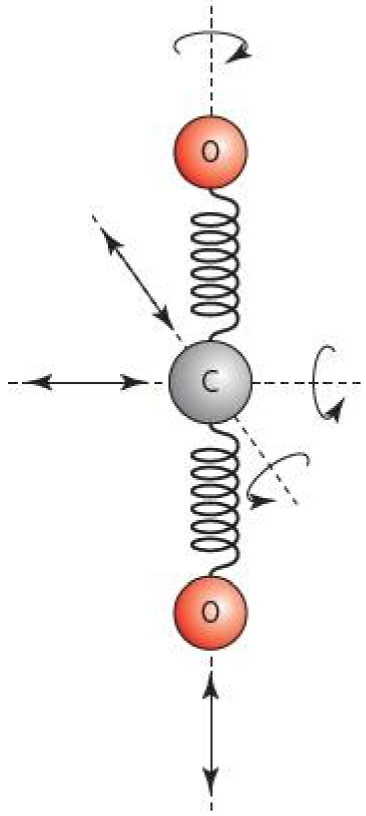


Figure 8.54 A few possible molecular vibrations in carbon dioxide

Water vapour is by far the most abundant of the greenhouse gases, but since the overall concentration of water vapour in the atmosphere does not appear to be much affected by human activity (such as burning fuels), it does not play a significant part in discussions about global warming.

Of the other gases carbon dioxide probably contributes between 15 and 20% of the overall greenhouse effect and is the major concern over global warming. Its concentration in the atmosphere is increasing mainly from the burning of fuels and changes to land use.

Methane and nitrous oxide absorb infrared radiation more strongly than carbon dioxide, but their concentrations in the atmosphere are very much lower. However, their increasing concentrations due to changes in land use and increases in animal waste are concerning. Changing diets around the world mean that more land is being used for animals and less for crops.

Molecules of greenhouse gases absorb infrared radiation because of *resonance* (see Chapter 4). Atoms within a molecule are not at rest – they oscillate with simple harmonic motion, like masses connected by springs. Figure 8.54 shows a much simplified example of possible modes of vibration of carbon dioxide.

The frequencies of oscillation can be calculated from knowledge of the masses of the atoms and the strength of the (covalent) bonds between them. For example, one of the natural frequencies at which carbon dioxide molecules oscillate can be calculated to be about 6×10^{13} Hz (which corresponds to a frequency of infrared radiation). Molecules of the other greenhouse gases oscillate at natural frequencies of about the same order of magnitude, and the molecules of each gas will have more than one frequency at which they oscillate.

Resonance occurs when energy is transferred to an oscillating system from another oscillation of the same (or very similar) frequency. Infrared frequencies are in the approximate range 0.1×10^{13} Hz to 40×10^{13} Hz.

This means that some infrared photons will have the right frequency to be absorbed by greenhouse gases due to resonance. It also means that the gases can emit photons of the same frequencies. Obviously, the molecules of the different greenhouse gases are not identical, so they will each absorb photons of different frequencies.

As an example, the sketch in Figure 8.55 indicates what happens to infrared radiation of different frequencies as they pass though carbon dioxide. A **transmittance** of 100% means that the radiation of that frequency passes through the gas without being absorbed. A transmittance of zero means that all radiation of that frequency is absorbed.

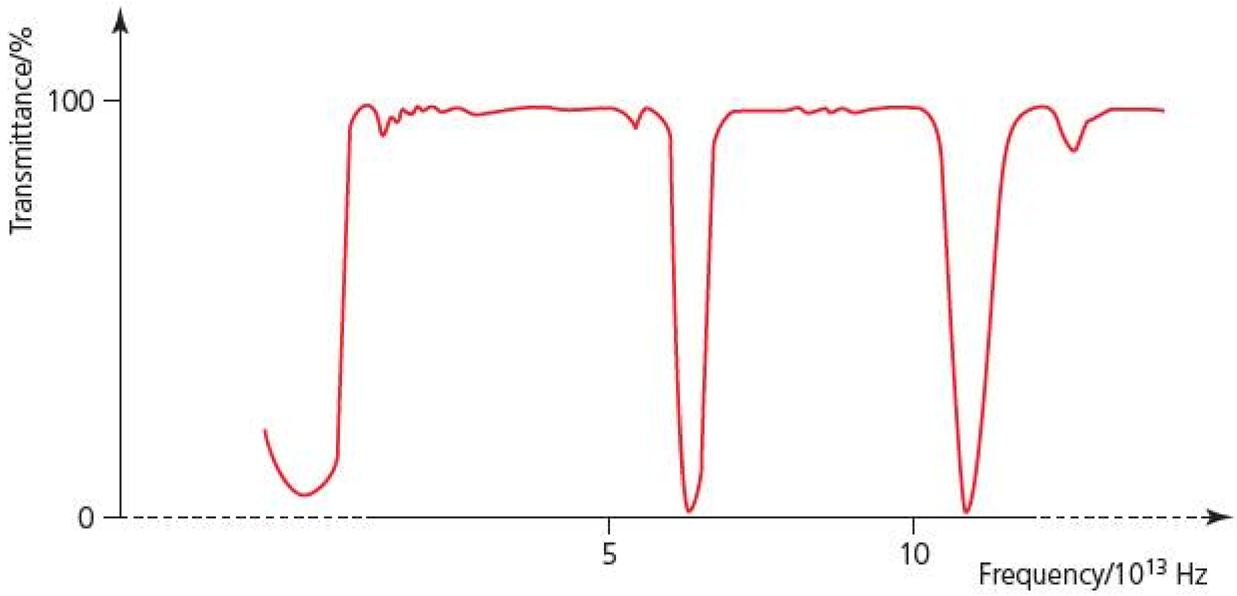


Figure 8.55 Transmittance of radiation through carbon dioxide

The graph shows, for example, that frequencies of about 6×10^{13} Hz and 11×10^{13} Hz are strongly absorbed (because carbon dioxide molecules oscillate at those frequencies).

Black-body radiation

8.5.8 **Outline** the nature of black-body radiation.
8.5.9 **Draw** and **annotate** a graph of the emission spectra of black bodies at different temperatures.

To fully understand the greenhouse effect it is necessary to understand in what ways the radiation emitted by the Sun is different from the radiation emitted from the Earth. Everything emits electromagnetic radiation, mostly in the infrared part of the spectrum. Differences in the radiation from different objects can only be in the total power and the range of frequencies emitted.

A perfect **black body** is an idealized object which absorbs *all* the electromagnetic radiation that falls upon it. Since it does not reflect light, it will appear black (unless it is very hot).

A good absorber of radiation is also a good emitter of radiation. The radiation emitted from a 'perfect' emitter is called **black-body radiation**. Remember that, if the object is hot enough to emit visible light, it will not appear black.

An object cannot emit all of its energy over all frequencies in an instant, so we can define (perfect) black-body radiation by using graphs which show the intensity distribution of the frequencies emitted. Figure 8.56 compares the radiation emitted from black bodies at different temperatures.

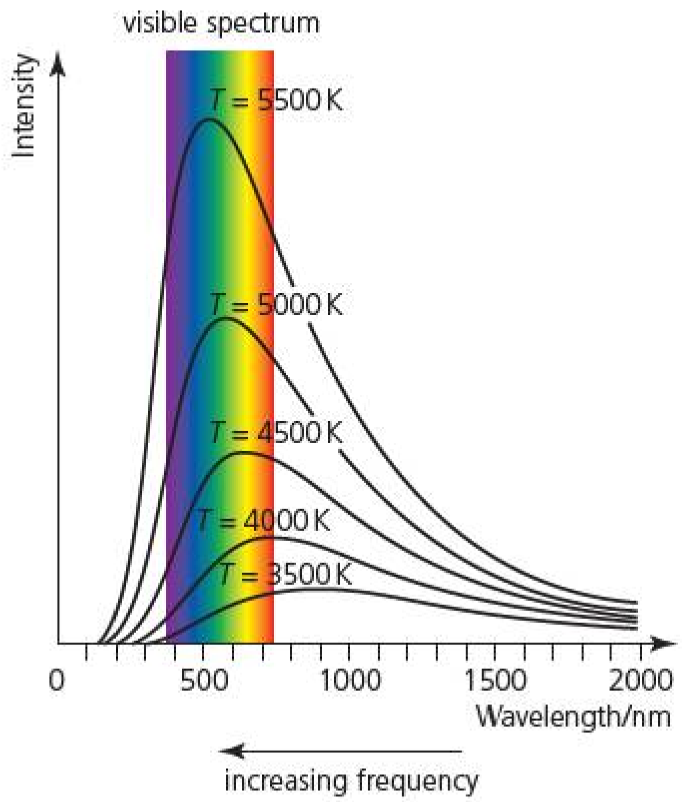


Figure 8.56 Spectra of black-body radiation at different temperatures

As any object gets hotter it emits more radiation at all frequencies, so that the graph is shifted upwards for higher temperatures. However, the distribution of frequencies/wavelengths also changes, with greater increases occurring at lower wavelengths. The peak wavelength moves to a lower value (higher frequency).

The vertical axis on the graph represents the intensity (power emitted per unit area) from a surface over a small range of wavelengths. The area under the graph is an indication of the total power emitted per unit area.

The surface temperature of the Sun is about 5800 K and the Earth's average surface temperature is approximately 288 K (which is too low to be shown clearly on the scale used in Figure 8.56). If we assume that they both behave approximately like perfect black bodies, we can compare the radiation emitted from the Sun and the Earth. The solar radiation arriving at the Earth from the Sun is approximately 50% visible light and 50% infrared, with small amounts of ultraviolet and other radiations. The radiation coming away from the Earth is much lower in intensity and all in the infrared part of the electromagnetic spectrum. As already mentioned, because of this difference in the spectra of the incoming and outgoing radiations, the greenhouse gases in the Earth's atmosphere will absorb a much greater proportion from the radiation emitted from the Earth than from the radiation arriving from the Sun.

Calculating radiated power emitted

8.5.10 **State** the Stefan–Boltzmann law and apply it to compare emission rates from different surfaces.
8.5.11 **Apply** the concept of emissivity to **compare** the emission rates from different surfaces.

The total power radiated from any surface depends on only three things:

1. surface area, A
2. surface temperature, T
3. nature of the surface.

Note that the power emitted depends on the nature of the surface, but *not* the substance itself.

Emitted power varies considerably with changes in temperature and this is shown by the **Stefan–Boltzmann law**. This states that for a perfect black body, the total power emitted per unit area is proportional to the fourth power of the (absolute) temperature:

$$\frac{\text{power}}{\text{area}} \propto T^4$$

The power, P, emitted from a perfect black body is given by the following equation:

$$P = \sigma A T^4$$

This equation is in the IB *Physics data booklet*.

Here σ is a constant. It is called the **Stefan–Boltzmann constant** and has the value $5.67 \times 10^{-8}\,\text{W m}^{-2}\,\text{K}^{-4}$. (This value is given in the IB *Physics data booklet*.)

This equation can be easily adapted for use with any surface which does not behave like a perfect black body by using the concept of the **emissivity**, *e*, of a surface.

Emissivity is a number between zero and 1 – a body with an emissivity of 1 represents a perfect black body emitting the maximum power. An emissivity of 0.5 represents a surface which emits 50% of the maximum power.

Emissivity is defined as the power radiated by a surface divided by the power radiated from a black body of the same surface area and temperature. Emissivity is a ratio and therefore does not have a unit.

$$e = \frac{\text{power radiated by a surface}}{\text{power radiated from a black body of same temperature and area}}$$

Dark surfaces have high emissivity (and a low albedo). Light surfaces have a low emissivity (and a high albedo). The Stefan–Boltzmann law for the power emitted from *any* surface then becomes:

$$P = e\sigma A T^4$$

This equation is given in the IB *Physics data booklet*.

Worked example

5 The power of our Sun is 3.9×10^{26} W and its surface temperature is 5800 K. Assuming that it behaves like a perfect black body, calculate the radius of the Sun.

$P = e\sigma A T^4$

$3.9 \times 10^{26} = 1 \times (5.67 \times 10^{-8}) \times (4\pi r^2) \times 5800^4$

$r^2 = \dfrac{3.9 \times 10^{26}}{8.06 \times 10^{8}}$

$r = 7.0 \times 10^{8}$ m

67 A star has a radius of 8.3×10^{7} m and a surface temperature of 7500 °C.
- **a** Calculate the power that it emits.
- **b** What assumption did you make?
- **c** What is the intensity at a distance of 2.4×10^{12} m from the star?

68
- **a** Estimate an average surface area for a naked adult human body.
- **b** If skin has an emissivity of 0.95, calculate the power radiated away from a naked body with a skin temperature of 33 °C.
- **c** In what other ways will the body lose heat?
- **d** Your answers may suggest that a naked body would lose heat very quickly. Explain why this is probably not true.

69
- **a** If the temperature of an object increases from 285 K to 312 K, what is the percentage increase in its emitted radiation?
- **b** An object is at 20 °C. To what temperature would it have to be heated to double the power of the radiation emitted?

70 When the current through an incandescent light bulb is turned on, the temperature of the filament will rise until the thermal energy is radiated away at a rate equal to the electrical power provided. Calculate the operating temperature of a 40 W light bulb if the filament has a surface area of 1.2×10^{-5} m². (Assume it behaves as a perfect black body.)

Planet in thermal equilibrium

A planet is in thermal equilibrium if the total incoming power equals the total outgoing power. A planet can only receive energy by radiation, so the average incoming radiation intensity must equal the average outgoing radiation intensity. If more energy is received by a planet than is radiated away, then the average temperature of the planet will rise. This is called **global warming**. As we shall see, it is possible to develop mathematical/computer models to predict possible temperature rises.

If the planet has no atmosphere we can simply equate the energy received from the Sun to the energy radiated away from the surface, $P = e\sigma AT^4$, but the situation is more complicated for a planet like the Earth, which has an atmosphere.

71 A planet is 3.4×10^{11} m from a star of power 2.3×10^{27} W.
- **a** Estimate a value for the surface temperature of the planet.
- **b** What assumptions did you make?
- **c** Explain why a planet of twice the radius would have the same temperature (at the same distance from the star).

72 The radius of the Earth is 6.37×10^6 m and it receives energy from the Sun at a rate of 1.74×10^{17} W. Assuming the surface temperature of the Earth is constant at 15 °C, estimate an effective overall value for the emissivity of the planet and its atmosphere.

Surface heat capacity

8.5.12 **Define** *surface heat capacity*, C_s.
8.5.13 **Solve** problems on the greenhouse effect and the heating of planets using a simple energy balance climate model.

In Chapter 3 we considered the temperature rise when an electrical heater was used to transfer energy to a metal block. For example, if energy was supplied at a rate P (say, 40 W) for a time t (100 s) to a block of metal of mass m (2.0 kg) with a specific heat capacity c (750 J kg^{-1} K^{-1}), the temperature rise ΔT can be calculated as follows:

$$\Delta T = \frac{Pt}{mc} = \frac{40 \times 100}{2.0 \times 750} = 2.7\,\text{K}$$

A possible temperature rise of the Earth could be calculated in a similar way, but it would not be sensible to consider the whole of the Earth as a single mass (like the block in the example above). This is because when energy is absorbed by a massive object (like a planet), it does not spread out evenly throughout the object, rather it has the greatest effect on the surface. It is therefore convenient to use the concept of *surface* heat capacity.

Surface heat capacity, C_s, is defined as the amount of energy needed to raise the temperature of a unit area of a planet's surface by one kelvin (unit: J m^{-2} K^{-1}).

To calculate a value for a surface heat capacity it would have been necessary to make an estimate of the depth over which the absorbed energy would be distributed and also make assumptions about the average specific heat capacities and densities of the soil and rocks that were heated. This means that surface heat capacities are approximate values. The average value for the Earth's surface is usually quoted to be 4.0×10^8 J m^{-2} K^{-1}.

$$\text{surface heat capacity } (C_s) = \frac{\text{thermal energy transferred } (Q)}{\text{area } (A) \times \text{temperature rise } (\Delta T)}$$

$$C_s = \frac{Q}{A\Delta T}$$

This equation is given in the IB *Physics data booklet*.

This equation can be rearranged to give:

$$Q = AC_s\Delta T$$

In this form, the equation can be compared to $Q = mc\Delta T$, used for calculations involving specific heat capacity in Chapter 3.

If we want to calculate changes in temperature over a period of time (Δt) we need to consider power and intensity:

$$Q = \text{power } (P) \times \text{time } (\Delta t)$$

Therefore:

$$P\Delta t = AC_s\Delta T$$

Rearranging, and remembering that P/A is intensity (I), we get:

$$\Delta T = \frac{I\Delta t}{C_s}$$

Or, if we use I_{in} to represent the incoming radiation intensity and I_{out} for the outgoing radiation intensity, this can be rearranged to give:

$$\Delta T = \frac{(I_{in} - I_{out})\Delta t}{C_s}$$

This equation is in the IB *Physics data booklet*.

It may be alternatively interpreted as:

$$\frac{\text{change of planet's temperature}}{\text{time}} = \frac{\text{overall radiation intensity absorbed}}{\text{surface heat capacity}}$$

$$\frac{\Delta T}{\Delta t} = \frac{I_{in} - I_{out}}{C_s}$$

Worked example

6 As stated above, the Earth receives an average intensity of $342\,\text{W m}^{-2}$ and, if it is in thermal equilibrium, it reflects or radiates an equal amount back out into space. As an example, suppose that human activity reduced the outgoing intensity to an average of $340\,\text{W m}^{-2}$. (That is, only $2\,\text{W m}^{-2}$ less.) What annual temperature rise would be predicted by the equation above?

$$\Delta T = \frac{(I_{in} - I_{out})\Delta t}{C_s}$$

$$\Delta T = \frac{(342 - 340) \times (365 \times 24 \times 3600)}{4.0 \times 10^8}$$

$\Delta T = 0.16\,\text{K}$ in one year.

This would (wrongly) suggest, for example, a temperature rise of 1.6 K every ten years.

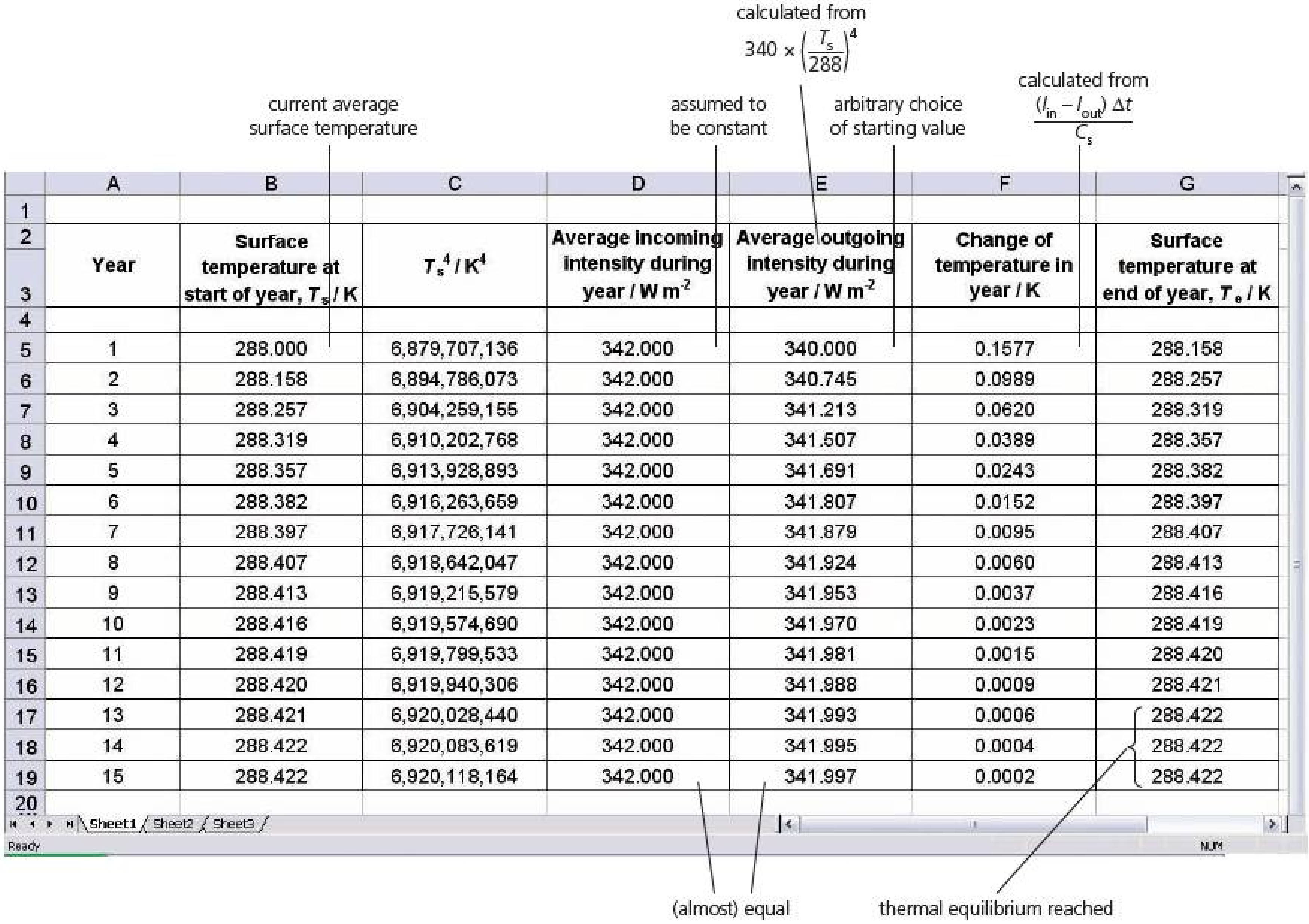
The calculation in Worked example 6 is presented as an example only, and the value used for a reduced radiated intensity ($340\,\text{W m}^{-2}$) should *not* be assumed to be accurate. However, the calculation does suggest that a small change (from $342\,\text{W m}^{-2}$ to $340\,\text{W m}^{-2}$, which is only –0.6%) *might* have a serious effect on the Earth's climate. But, even if we use the same simplified energy balance mathematical model with the latest accurate input data to attempt to predict what might happen to the Earth's temperature in the next ten years, we would be making at least four very important mistakes.

1 Any increase in surface temperature would also increase the power radiated away (consider the Stefan–Boltzmann law). Overall, this is an example of **negative feedback,** because any change produced by the effect reduces the size of subsequent changes.

2 The calculation ignores the possible effects of any changes to the Earth's climate that may be produced by the increasing temperatures (increased cloud cover, melting ice, etc.).
3 This simple model does not consider that human activities may continue to make the situation worse.
4 The model does not take into account any other natural processes that periodically affect the temperature of the Earth.

Computer modelling

Consider how the first of the four problems mentioned above could be analysed by using a computer spreadsheet to make multiple calculations.



	A	B	C	D	E	F	G
1							
2–3	Year	Surface temperature at start of year, T_s / K	T_s^4 / K^4	Average incoming intensity during year / W m^{-2}	Average outgoing intensity during year / W m^{-2}	Change of temperature in year / K	Surface temperature at end of year, T_e / K
4							
5	1	288.000	6,879,707,136	342.000	340.000	0.1577	288.158
6	2	288.158	6,894,786,073	342.000	340.745	0.0989	288.257
7	3	288.257	6,904,259,155	342.000	341.213	0.0620	288.319
8	4	288.319	6,910,202,768	342.000	341.507	0.0389	288.357
9	5	288.357	6,913,928,893	342.000	341.691	0.0243	288.382
10	6	288.382	6,916,263,659	342.000	341.807	0.0152	288.397
11	7	288.397	6,917,726,141	342.000	341.879	0.0095	288.407
12	8	288.407	6,918,642,047	342.000	341.924	0.0060	288.413
13	9	288.413	6,919,215,579	342.000	341.953	0.0037	288.416
14	10	288.416	6,919,574,690	342.000	341.970	0.0023	288.419
15	11	288.419	6,919,799,533	342.000	341.981	0.0015	288.420
16	12	288.420	6,919,940,306	342.000	341.988	0.0009	288.421
17	13	288.421	6,920,028,440	342.000	341.993	0.0006	288.422
18	14	288.422	6,920,083,619	342.000	341.995	0.0004	288.422
19	15	288.422	6,920,118,164	342.000	341.997	0.0002	288.422
20							

Figure 8.57 Computer spreadsheet of Earth's temperature

Figure 8.57 shows a spreadsheet that recalculates the outgoing radiation every year based on the temperature at the start of that year. In this way the temperature changes each year become less and less as the planet warms up. This spreadsheet shows that after about 15 years, thermal equilibrium would be almost reached and the temperature becomes constant at 0.422 K higher than at the start of the calculations.

The calculation could be made more accurate by getting the computer to recalculate every week (or even every day, hour, minute or second) rather than every year. It is also easy to change the starting assumption (340 W m^{-2}) to another value. For example, if 330 W m^{-2} is used the temperature stabilizes at a higher temperature of 290.6 K.

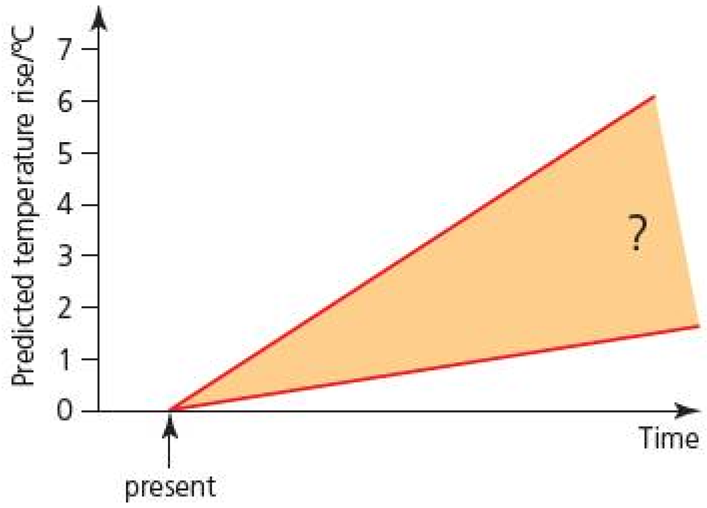


Figure 8.58 Graph to show the change in temperature predictions over the next 100 years: the future is uncertain

The problems about predicting the future temperature and climate of the Earth are enormous because there are so many interlinked variables and unknowns. An average value for the current surface temperature for the Earth is also difficult to determine with a high degree of accuracy and there is some uncertainty over the current value of the average outgoing radiation intensity.

Computer modelling and simulations can be used to make predictions of complex situations more easily. In recent years, an enormous amount of time, expertise and resources have gone into trying to predict possible future climate changes. Despite this, there is still a great deal of uncertainty and much disagreement over the various temperature rises predicted by different computer models. Most predictions are in a range from about 1 °C to 6 °C for the next 100 years (Figure 8.58).

TOK Link: Using computers to expand human knowledge

Trying to predict the future, or to answer the question 'what would happen if….' has always been a very common and enjoyable human activity, but it seems that our predictions are usually much more likely to be wrong than right. This is partly because, in all but the simplest of examples, there are just too many variables and unknown factors. Of course, the delightful inconsistencies of human nature play an important part when we are dealing with people's behaviour, but accurately predicting events governed mostly by the laws of physics, such as next week's weather, can also be extremely difficult.

When a situation can be represented by equations and numbers, mathematical modelling is a powerful tool to help understanding. But even in the simplest situations, there are nearly always simplifications and assumptions which result in uncertainty in predictions. When dealing with complex situations, like predicting next month's weather, the value of a financial stock next year, or the climate in 50 years' time, even the most able people in the world will struggle with the complexity and amount of data. The rapid increase in computing power in recent years has changed this.

Modern computers have computing power and memory far greater than human beings. They are able to handle masses of data and make enormous numbers of calculations that would never be possible without them. They are ideal for making predictions about the future climate, but that does not necessarily mean that the predictions will turn out to be correct. Computer predictions are limited by the input data provided to them and, more particularly, by the specific tasks that human beings have asked them to perform. To check the accuracy of predictions, computer models can be used to model known complex situations from the past to see if they are able to predict what happened next. But predicting the past is always much easier than predicting the future.

Questions

1 a Make any one confident scientific/technological prediction about the world in 2020.
  b Explain why you are confident about your prediction.
  c Compare your prediction with other students' predictions.

2 Does a computer just use the human knowledge that has been input into it, or can it increase human knowledge? Explain your answer.

3 Is it reasonable to think that a computer program which accurately predicted today's weather one week ago, can accurately predict next week's weather today? Explain your answer.

73 **a** If the surface heat capacity of a large, deep lake is $4.2 \times 10^8\,\mathrm{J\,m^{-2}\,K^{-1}}$, predict the maximum possible temperature rise at the end of a 12 hour day in which the average intensity of the radiation falling on the water was $280\,\mathrm{W\,m^{-2}}$. Assume that the energy radiated away is negligible.
b Comment on your answer.

74 Use the fact that the specific heat capacity of water is $4180\,\mathrm{J\,kg^{-1}\,°C^{-1}}$ and its density is $1000\,\mathrm{kg\,m^{-3}}$ to show that the value quoted for surface heat capacity in question **73** assumes that the thermal energy spreads down evenly to a depth of 100 m.

75 Use a spreadsheet similar to that shown in Figure 8.56 to determine the equilibrium temperature that will be reached by a planet which has a surface temperature of 295 K if the incoming intensity is $418\,\mathrm{W\,m^{-2}}$ and the outgoing intensity is originally $395\,\mathrm{W\,m^{-2}}$.

8.6 Global warming

8.6.1 **Describe** some possible models of global warming.
8.6.2 **State** what is meant by the enhanced greenhouse effect.
8.6.3 **Identify** the increased combustion of fossil fuels as the likely major cause of the enhanced greenhouse effect.

The Earth has changed temperature in the past, but the recent increases appear to be relatively sudden. Figure 8.59 shows data representing global average temperatures over the last 130 years. Although there has been some doubt over the accuracy of some of this information (especially the older data), scientists generally agree that the planet is getting warmer, that the average temperature rose by 0.7 °C (± 0.2 °C) between 1906 and 2005, and that the rise in temperature in the 21st century will be almost certainly be even greater.

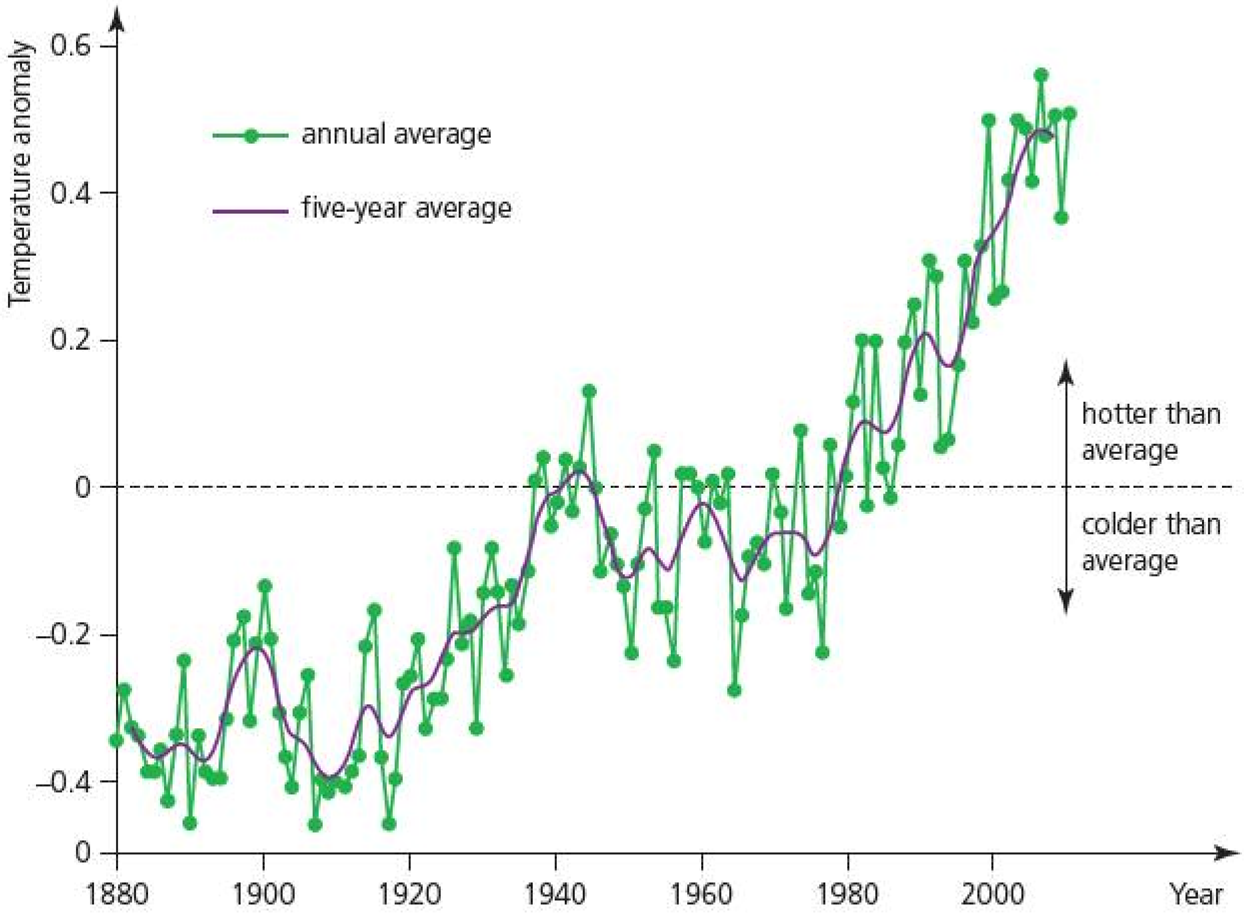


Figure 8.59 Changes in global temperatures. The temperature anomaly is the difference between the measured temperatures and the expected long-term average

In general, there are several possible explanations for any changes in mean global temperatures.

- An **enhanced greenhouse effect** due to the increasing concentration of greenhouse gases (especially carbon dioxide) caused by human activities. This reduces the intensity of radiation emitted from the Earth. The Intergovernmental Panel on Climate Change (IPCC) concluded that recent rises in global temperature were 'very likely due to the observed increase in anthropogenic greenhouse gas concentrations'. (Anthropogenic means caused by humans.) This means that the increased combustion of fossil fuels has released extra carbon dioxide into the atmosphere and is probably the most important cause of the current global warming.
- **Variation in the inclination of the Earth's axis** (to its plane of rotation around the Sun), which gradually take place over thousands of years and is believed to be responsible for 'ice ages'.
- Changes in the intensity of the radiation received from the Sun, especially **solar flare activity**. The radiation from the Sun varies over an 11 year cycle and there has been long-term evidence that this has an influence, periodically changing the Earth's temperature slightly. However, the recent rise in temperature has not been clearly linked to changes in the Sun's activities.

- **Geological processes, including volcanic activity** which releases ash and gases, including carbon dioxide, into the atmosphere. As a greenhouse gas, the carbon dioxide will have a slight effect on increasing temperatures, but the ash will also reflect or scatter more radiation and stop it reaching the Earth's surface, and could produce global cooling.
- Changes in the **Earth's orbital** distance from the Sun (although these natural cycles take many thousands of years).

Evidence linking global warming to the levels of greenhouse gases

8.6.4 **Describe** the evidence that links global warming to increased levels of greenhouse gases.

Ice extracted from deep below the Antarctic surface, at the Russian base at Vostok, has been studied in great detail (isotopic analysis) for evidence of atmospheric carbon dioxide concentrations. These concentrations have been linked to global temperatures going back hundreds of thousands of years.

The results are presented in Figure 8.60 and it seems clear that there is a definite *correlation* between average global temperatures and the amount of carbon dioxide in the atmosphere. This, in itself, does not prove that changing levels of carbon dioxide *caused* the changes in temperature.

More recently, and over a much shorter time scale, significant increases in the concentrations of greenhouse gases in the atmosphere have occurred at a time of global warming, as indicated approximately in Figure 8.61.

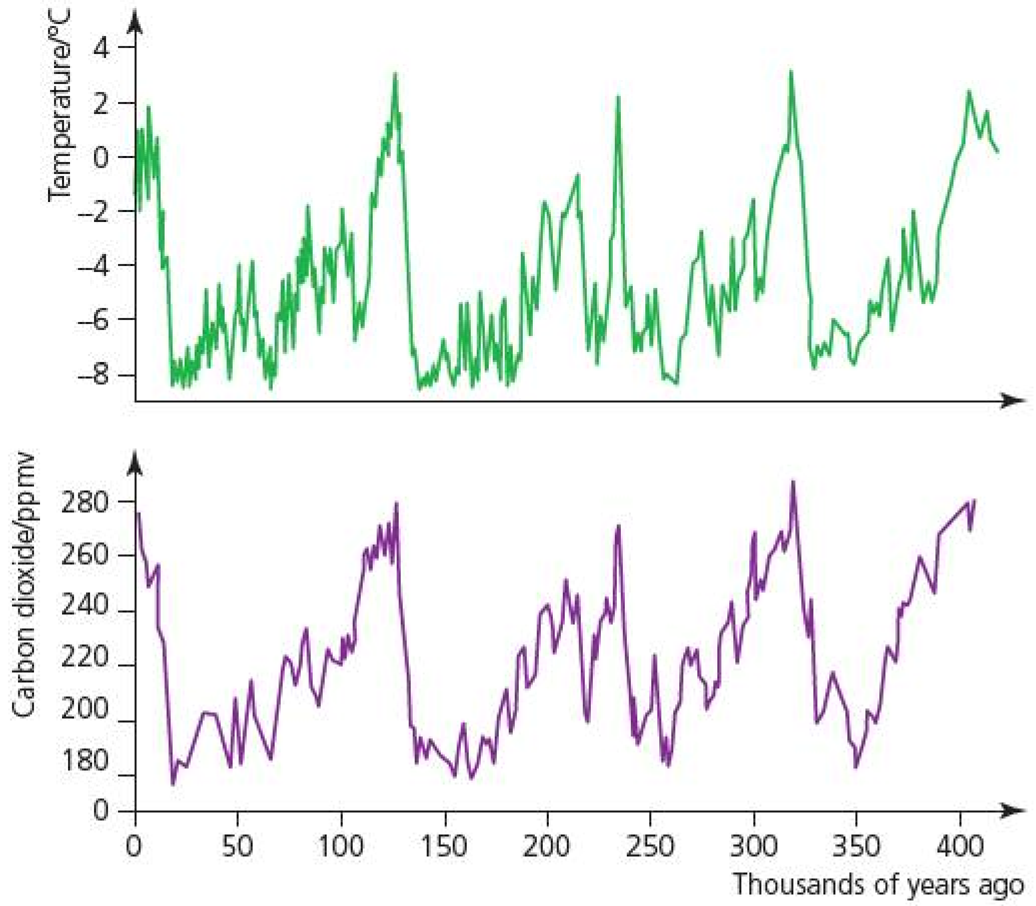


Figure 8.60 There is a correlation between average global temperatures and the amount of carbon dioxide in the atmosphere (ppmv = parts per million by volume)

TOK Link: Correlation and cause

The **correlation** (link) between increasing concentrations of greenhouse gases and rising global temperatures is well established and accepted by (almost) everybody. However, that does not mean that we can be sure that global warming is **caused** by the release of more greenhouse gases. Obviously, controlled experiments to test such a theory cannot be carried out and we must rely on limited statistical evidence, computer modelling and scientific reasoning. In such cases, 100% certainty is never possible and societies must make informed judgements based on the best possible scientific evidence. Of course, some people will always choose to disagree with, or ignore, the opinions of the majority.

Question

1 Suggest reasons why some people (rightly or wrongly) claim that there is no correlation between greenhouse gas emissions and global temperature rises.

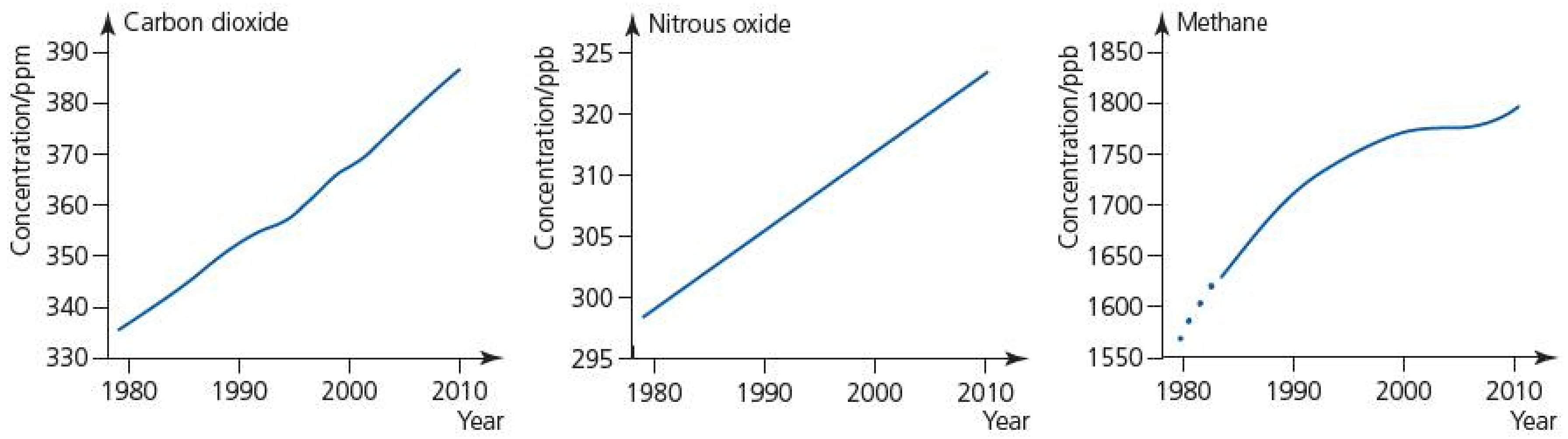


Figure 8.61 The increasing concentrations of greenhouse gases over the recent years (ppb = parts per billion)

Consequences of global warming

Climate change and rising sea-levels are the main direct effects on the human population of global warming but some other consequences will also have a feedback effect on the rate of change of global temperatures.

Effects which increase the rate of global warming

8.6.5 **Outline** some of the mechanisms that may increase the rate of global warming.

There are a number of mechanisms that may increase the rate of global warming.

1 Snow and ice have a high albedo and global warming will result in less snow and ice covering the Earth's surface, therefore lowering the Earth's average albedo and increasing the absorption of solar radiation in those places.
2 If the temperatures of the oceans increase, then less carbon dioxide will be dissolved in the water (the solubility of carbon dioxide reduces as the temperature rises) meaning that the gas is released into the atmosphere to further increase the greenhouse effect.
3 Changing land use has an effect on the concentration of greenhouse gases, especially carbon dioxide. Trees and other plants remove carbon dioxide from the air in the process of photosynthesis. This is an example of what is known as **carbon fixation** – the removal of carbon dioxide from the air and the formation of solid carbon compounds. Any widespread changes of land use (particularly deforestation) may lead to an increase in the atmospheric concentration of carbon dioxide. Changes of land use may also lead to changes in albedo and, if the land is cleared by burning, more carbon dioxide will be released in the process.

Even a small rise in the Earth's temperature may produce the effects detailed in the three points above, affecting the temperature balance of the planet even more. This makes predicting the future temperature and climate of the Earth very complicated. Figure 8.62 illustrates and simplifies just *some* of the many influences.

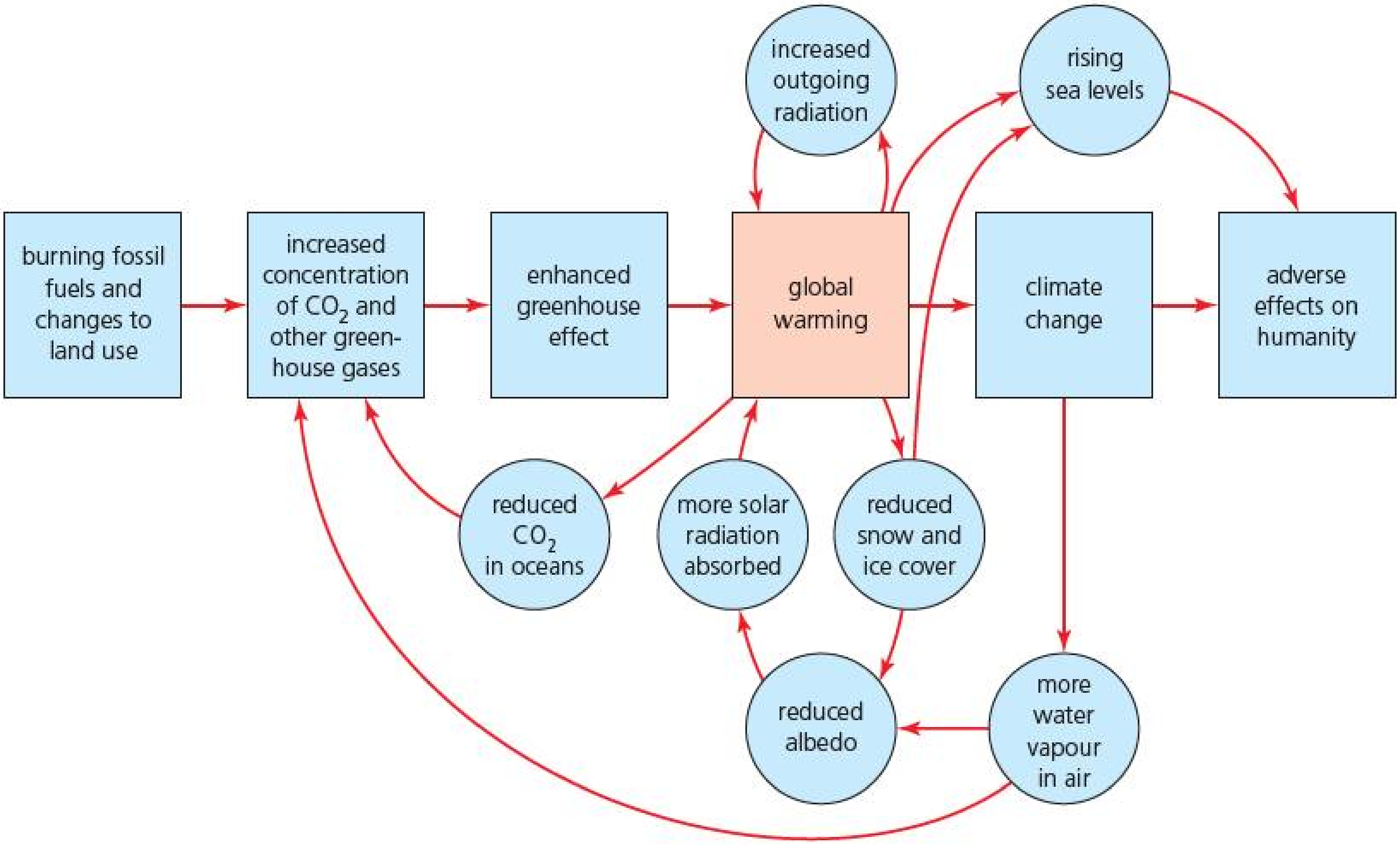


Figure 8.62 Some of the many interconnected factors affecting global warming

Rising sea-levels

8.6.6 **Define** *coefficient of volume expansion.*
8.6.7 **State** that one possible effect of the enhanced greenhouse effect is a rise in mean sea-level.
8.6.8 **Outline** possible reasons for a predicted rise in mean sea-level.

Global warming will result in rising sea-levels. There are two reasons for this:

- thermal expansion of water, and
- any ice and snow that melts on land will result in extra water flowing into the seas.

Note that when *floating* ice melts (for example the ice in the Arctic), there is no significant change in sea-level.

Seventy-one per cent of the Earth's surface is covered with water and the average depth is about 3800 m. This is a very large amount of water, so understanding the physical properties of water is very important when discussing the planet's climate. When water which is warmer than 4 °C rises in temperature, its molecules move very slightly further apart. The water therefore increases in size (expands) and its density will decrease very slightly (just like other solids and liquids when they are heated). This is shown in Figure 8.63.

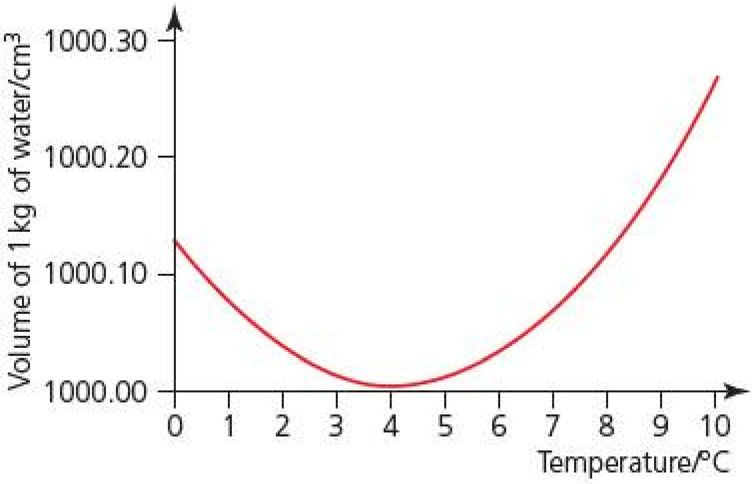


Figure 8.63 Water has its maximum density at 4°C

The expansion of water is very unusual (anomalous) because, although it expands with most increases in temperatures, for temperatures rising between 0 °C and 4 °C it contracts (gets smaller).

In order to make calculations on the expansion of water we need to use its coefficient of volume expansion.

The **coefficient of volume expansion**, γ, is defined as the fractional change in volume, $\Delta V/V_0$, per unit change in temperature, ΔT. Its unit is per kelvin, K^{-1} (or $°C^{-1}$).

$$\gamma = \frac{\Delta V}{V_0 \Delta T}$$

As can be confirmed from Figure 8.63, the expansion of water is not linear, so this coefficient is not constant for different temperatures. At 20 °C it has the value $2.07 \times 10^{-4}\,K^{-1}$. If it was free to do so, heated water would expand in all directions but water in the oceans can only expand upwards. This means that the rise in sea-levels caused by expansion (due to a small increase in temperature) can be surprisingly large.

Accurate predictions of rising sea-levels are difficult because of variations in water temperature with location and depth, the anomalous expansion of water and the differing effects of snow and ice melting on water and land.

Worked example

7 a Estimate the total volume of water in the world's oceans and seas.
b Calculate the approximate increase in volume (at 20 °C) which would result from expansion if the average temperature of the water rose by (only) 1.0 °C.
c Predict the rise in sea-level which would result from this increase in temperature.

a Surface area of Earth = $4\pi r^2 = 4 \times \pi \times (6.37 \times 10^6)^2 = 5.10 \times 10^{14}\,m^2$
Surface area of water is 71% of this, or $3.6 \times 10^{14}\,m^2$
Volume of water, V_0 = area × average depth = $3.6 \times 10^{14} \times 3800$
$V_0 = 1.4 \times 10^{18}\,m^3$

b $\gamma = \dfrac{\Delta V}{V_0 \Delta T}$
$\Delta V = \gamma V_0 \Delta T$
$\Delta V = 2.07 \times 10^{-4} \times 1.4 \times 10^{18} \times 1.0$
$\Delta V = 2.9 \times 10^{14}\,m^3$

c Increase in depth = $\dfrac{\text{increase in volume}}{\text{area}} = \dfrac{2.9 \times 10^{14}}{3.6 \times 10^{14}}$
$= 0.81\,m$

The figure calculated in Worked example 7 is both interesting and alarming, but to use it to suggest that every degree rise in temperature would result in a sea-level rise of about 80 cm would be unwise for a number of reasons:


Figure 8.64 The Maldives are a low-lying country threatened by rising sea-levels

- any increase due to melted snow and ice has not been included
- the calculation incorrectly assumes that all the oceans are at the same temperature, have the same temperature rise and the same coefficient of volume expansion
- increasing sea-levels will result (unfortunately) in flooding of low-lying land and a small increase in the surface area of the seas and oceans.

Rising water levels will be a very considerable problem for many countries, especially those which have large areas of land that are now only a few metres above current mean sea-levels, for example the Maldives (see Figure 8.64).

Climate change

8.6.9 **Identify** climate change as an outcome of the enhanced greenhouse effect.


Figure 8.65 Are hurricanes (typhoons) getting worse because of global warming?

As the mean temperatures of air and sea water rise due to the enhanced greenhouse effect, the climate at different locations around the world will also change. Exact details are very hard to predict, but these effects may be considerable, especially if the temperature rises are close to the upper end of the range of predictions. Apart from temperature changes, there may be changes to rainfall (amounts and its annual distribution), as well as to the amounts of cloud coverage.

These many and varied possible changes to the Earth's climate make it very difficult to predict what may be their overall influence on possible further global warming.

The changes will be more significant in some locations than in others. In some countries the changes may even be considered beneficial (for example, more rain could improve the conditions for growing crops), but many other changes will be less welcome. There is some evidence that extreme weather events (like typhoons, see Figure 8.65) may already be becoming more common, but this is not confirmed.

8.6.10 Solve problems related to the enhanced greenhouse effect.

76 Use data from Figure 8.61 to estimate by what percentage the concentrations in the atmosphere of the three greenhouse gases has increased between the years 1980 and 2010.

77 **a** Estimate the mass and total surface area of a large iceberg.
b How much energy would be needed to melt it? The specific latent heat of fusion of ice is $330\,\mathrm{kJ\,kg^{-1}}$.
c What assumption did you make?
d If the average thermal intensity flowing into the iceberg is $50\,\mathrm{W\,m^{-2}}$, how long will it take to melt?

78 **a** A lake has an average depth of 23 m. If the average temperature of the lake rises by 15 °C between winter and summer, use the equation on page 308 to estimate a possible rise in the level of water in the lake.
b Give reasons why the actual change in water level may be very different from your answer.

79 **a** Calculate the amount of energy needed to raise the average temperature of the world's oceans by 1 °C.
b How long would it take the Sun to radiate that amount of energy to Earth?

80 **a** The specific latent heat of vaporization of water is $2.26 \times 10^6\,\mathrm{J\,kg^{-1}}$. Explain what this means.
b Use this figure to calculate the thermal energy that must have been supplied to a 500 000 kg cloud to evaporate it from water on the surface of an ocean.
c If the cloud is 5 km above the ocean, compare your answer in **b** to the gravitational potential energy that has also been gained by the water in the cloud.
d Assuming that the thermal energy for evaporation was supplied by radiation from the Sun at an average rate of $25\,\mathrm{W\,m^{-2}}$ over an area of $10\,\mathrm{km^2}$, how long would it have taken for the water in the cloud to evaporate from the ocean (give your answer to the nearest hour)?

81 Use the Internet to find out about the West Antarctic Ice Sheet and why some scientists are worried about it.

82 Research into the origins and uses of the gas SF_6, which is possibly the most potent of all greenhouse gases.

Trying to reduce global warming

Reducing global warming has become a major world-wide issue and many people claim that this is the biggest problem facing the modern world. It receives an enormous amount of media attention and considerable efforts are being made at all levels of society and technology to try to solve the problems.

However, the increasing demand for energy in the developing nations and a world population which is still rising, mean that the overall demand for electrical energy (and the need to burn fossil fuels) is still rising. Many people would also claim that there is a fundamental conflict between the need to reduce energy demands and modern lifestyles in consumer-driven developed countries. Whilst it is true that most (but not all) informed scientists believe in the enhanced greenhouse effect, there is still an element of doubt and so there are a large number of individuals who prefer to believe, for a variety of reasons, that the problem has been exaggerated (or even that the scientists are wrong).

Possible solutions to reducing the enhanced greenhouse effect

8.6.11 Identify some possible solutions to reduce the enhanced greenhouse effect.

There are a number of possible solutions for reducing the enhanced greenhouse effect.

- Encouraging organizations and individuals to reduce their energy needs. It is never easy to make people change their habits and any government which tried to introduce large price rises or new laws in order to reduce energy consumption would be very unpopular. Many people (especially in the developed countries) claim that there should be trends in society towards the use of smaller and fewer cars and better, more energy-efficient public transport, travelling less (to work and for leisure), reduced use of heating or air-conditioning and generally a life of sufficiency rather than consumerism and excess. Reducing a person's or an organization's need for the energy transferred from burning fossil fuels has become known as reducing their **carbon footprint**.

- Developing and greatly increasing the use of renewable energy resources and nuclear power.
- Setting and enforcing strict limits and controls on greenhouse gas emissions, such as may be imposed by international agreements. Countries (or other large organizations) should then have no choice but to stick to the agreed total limits, although it may not be fair or sensible to set the same limits on different countries. To encourage enterprise and flexibility within these limits, **carbon trading** (also know as 'cap and trade') can be introduced. This is when one country or organization can exceed its limit by buying the right to extra emissions from another country or organization. People and organizations can be encouraged to become **carbon neutral**, so that any of their actions which result in the emission of greenhouse gases are balanced by other actions which are beneficial to the environment, for example, planting trees.
- Using the latest technology to ensure that power stations are as efficient as possible. This may mean closing older power stations and using natural gas as the fuel wherever possible. Using natural gas is more efficient and creates fewer emissions than coal or fuel oil. As discussed, a large and unavoidable percentage of the input energy to a power station is transferred to thermal energy, which is often just spread into the surroundings. In **combined heat and power systems (CHP)** some of that thermal energy is transferred using hot water to keep local homes and offices warm. Helsinki, in Finland, is a very good example of the widespread use of this idea (Figure 8.66).

Figure 8.66 Homes in the local community can be heated by thermal energy from this power station in Helsinki

- Encouraging the use of electric and **hybrid vehicles**. Vehicles which are powered by petrol (gasoline) are convenient and powerful, but they are very inefficient and emit greenhouse gases into the atmosphere. Electric vehicles have considerable advantages: they are quiet, efficient and emit no pollution or greenhouse gases while they are being driven. However, there are also disadvantages: (i) the energy that they use is stored in a battery, and even the very best batteries have a much lower energy density than petrol/gasoline. This means that, even with a large battery, it is not possible to store enough energy for long journeys, and a large battery also adds considerable mass to the car; (ii) electric vehicles are usually less powerful, and have less acceleration, than gasoline-driven vehicles; (iii) it takes hours to recharge the batteries; (iv) the electrical energy used to charge the battery had to be generated in a power station (probably burning a fossil fuel), and then transferred to the place where the vehicle was charged. The overall energy efficiency and greenhouse emissions of electric vehicles may be only slightly better than using gasoline engines, especially when the manufacture of the vehicles and batteries are also considered.

  Hybrid vehicles (Figure 8.67) have been developed to overcome the disadvantages of both electric and gasoline-powered vehicles, at the same time as combining their advantages. In a hybrid vehicle, an electric motor provides the power efficiently for driving at slower speeds and an internal combustion engine is used, at its greatest efficiency, when a greater power is need for higher speeds or accelerations. The use of a regenerative braking systems means that when brakes are used to reduce the speed of a vehicle, kinetic energy is transferred to chemical energy

Figure 8.67 A hybrid vehicle

in the battery using a generator (rather than thermal energy in the brakes, tyres and road). The whole system is computer-controlled and the vehicle has a smaller mass than a gasoline-powered vehicle.

- Carbon capture and storage. Carbon capture is the term used for the removal of carbon dioxide from the gases released when fossil fuels are burned at the power station. Various methods are available and effective, but they considerably reduce the efficiency of the power station, so that more fuel has to be burned to produce the same output power. This would mean a sharp rise in electricity prices. Once the carbon dioxide has been removed it has to be permanently stored (also known as sequestering). There are several possibilities, including storage as a gas and storage in the form of carbonates.

  Increasing the area of the Earth covered by trees and other plants (and stopping deforestation) captures carbon in the plant material and is an appealing way of increasing the amount of carbon dioxide removed from the atmosphere by photosynthesis.

International efforts to reduce global warming

8.6.12 Discuss international efforts to reduce the enhanced greenhouse effect.

Representatives from the governments of most countries of the world meet regularly for discussion and to decide policies and targets. The **Kyoto Protocol** of 1997 (now agreed by about 190 countries but not the USA or Canada) was an important step forward and set some targets for the emission of greenhouse gases for the years up to 2012.

Although progress is being made, many people claim that it is not enough, that the targets are not strict enough and only voluntary. Some countries think that they should be allowed more freedom because their economies are a lot less developed than richer countries, whereas some more developed countries believe that controls on them are too strict! The protocol was extended until 2017 at the United Nations climate talks in Durban in 2011, with the clear intention to work towards stricter and more binding controls by 2020. This, hopefully, will allow time for the many differences of opinion to be reconciled, although some people consider this to be too great a delay.

The Asia-Pacific Partnership on Clean Development and Climate (APPCDC or AP6) is a group of six major countries (Australia, China, India, Japan, South Korea and the USA) who have agreed to work together to reduce emissions of greenhouse gases.

For more than 20 years the IPCC (**Intergovernmental Panel on Climate Change**) has collected, analysed and organized the latest scientific information on the climate from around the world. It regularly publishes comprehensive and authoritative reports. It is well worth visiting the IPCC web site.

83 **a** Do you think that most of the people that you know are happy to change their habits in an effort to reduce global warming?
b If yes, explain how. If not, explain why not.

84 Hybrid cars are significantly more expensive to buy than conventional gasoline-powered cars, so not so many are sold. Explain why some people are still happy to pay the extra money for hybrid vehicles.

85 Use the Internet to research the latest information on carbon sequestering.

86 Find a suitable site on the Internet that will enable you to calculate the carbon footprint of your lifestyle.

87 Use the Internet to find out about the ways in which the city of Helsinki reduces the adverse effects of burning fossil fuels.

SUMMARY OF KNOWLEDGE

8.1 Energy degradation and power generation

- Heat engines can do useful work when significant amounts of thermal energy flow from hot to cold. Most commonly this involves an expanding gas. Heat engines are used in power stations to help generate electricity and in various means of transport.

- In a typical power station, thermal energy is transferred to internal energy and kinetic energy of steam at high pressure. The steam forces turbine blades to rotate. The turbine is connected to coils of wire that are made to rotate in a magnetic field, which then induces a voltage.
- The conversion of thermal energy to useful work can be very efficient, but only if it is in a single process. However, practical heat engines need to work in repeating cycles. During each cycle, some energy has to be transferred out of the system and into the surroundings. This greatly reduces the efficiency of all power stations and other heat engines.
- Energy which has dissipated into the surroundings cannot be recovered to do any useful work. It is called degraded energy.
- Flow (Sankey) diagrams can be used to represent the transfer of energy in different systems.

8.2 World energy sources

- Approximately 90% of the total energy used in the world is transferred from burning fossil fuels. Oil is the most widely used (about 40%), with coal and natural gas each about 25%.
- The remainder of the world's energy is transferred mostly from nuclear fuels and from various renewable sources (particularly biomass and hydroelectricity).
- Burning fossil fuels emits greenhouse gases, which almost certainly cause global warming.
- Fossil fuels are non-renewable. This means that we are using them faster than they are being formed, and we will run out of them in the future.
- Fossil fuels have the major advantage of being much more energy dense than renewable sources. The Sun is the prime source of most of the world's energy.
- The energy density of a fuel is defined as the energy transferred from unit mass ($J\,kg^{-1}$).
- Nuclear fuels have extremely high energy densities. Nuclear power stations have the very big advantage of not emitting greenhouse gases, but many people have worries about their safety.

8.3 Fossil fuel power production

- After the invention of the first heat engines which used the combustion of coal to do mechanical work, industries and communities developed close to coal mines.
- By the end of the 19th century large power stations were being built around the world to produce electrical energy from the combustion of energy dense fossil fuels. This, together with the development of transport (also using fossil fuels), rapidly transformed and improved the lifestyles of many millions of people.
- Many centres of population are not near sources of fossil fuels, so road and rail transport systems had to be developed, together with suitable and safe storage.
- Oil and natural gas can be conveniently transferred through pipelines, but there is a risk of pollution from leaks.
- The extraction of fossil fuels from underground is dangerous work which has caused the illnesses and deaths of many tens of thousands of people.
- The efficiencies of coal-, oil- and gas-powered stations are approximately 35%, 40% and 45%.
- The rate of fuel consumption in various power stations can be calculated from their efficiency and other relevant data.
- Fossil fuels provide a relatively cheap and very convenient source of large amounts of energy, but they are non-renewable and their supply will probably run out in the next hundred years.
- Combustion of fossil fuels also has the major disadvantage of almost certainly causing global warming due to the release of greenhouse gases.

8.4 Non-fossil fuel power production

- Nuclear fission fragments have considerable kinetic energy, and if the number of fissions can be sustained, the nuclear fuel will gain a large amount of internal energy. Thermal energy can then be transferred to steam, which can turn turbines and generate electricity.
- In order to sustain a continuous fission process, a controlled chain reaction is needed in which, on average, one neutron from each fission causes one more fission. Many neutrons escape from the material, so a critical mass is needed before a chain reaction is possible.

- Control rods are used to absorb neutrons and control the rate of reaction.
- A chain reaction is only possible if there is enough ^{235}U in the core, and the neutrons are travelling slow enough (about 1 eV). The neutrons released in fission are very energetic and they must be slowed down in collisions with atoms of a material known as a moderator.
- Uranium ore contains less than 1% ^{235}U and this percentage must be increased (to more than 3%) before the fuel can be used in a reactor. This process is called enrichment.
- The operation of a nuclear reactor can be represented in a schematic drawing which highlights the function of the heat exchanger in tranferring thermal energy from the reactor vessel.
- If, on average, more than one of the neutrons produced in fission goes on to cause further fission, then the reaction will very quickly become uncontrolled and an enormous amount of energy will be released. This is what happens in nuclear weapons.
- When ^{238}U nuclei in the core of a reactor capture neutrons, ^{239}U is produced, which then undergoes two beta decays into ^{239}Pu.
- ^{239}Pu is an alternative nuclear fuel to ^{235}U. ^{239}Pu undergoes fission more readily and releases more energy than ^{235}U.
- The risks associated with using nuclear power include (i) the leak of radioactive materials from a damaged power station; (ii) thermal meltdown due to loss of coolant or problems with the control rods; (iii) the disposal of waste products; (iv) the dangers of mining uranium and (v) the development and possible use of nuclear weapons.
- The development of nuclear weapons raises important moral and ethical issues for society.
- The fusion of light nuclei also releases large amounts of energy (see Chapter 7), but the technical difficulties of confining and sustaining a high temperature plasma continue to prevent the development of nuclear fusion reactors.
- When considering the possible use of any energy sources (including the renewable sources discussed below) the following factors should be considered: emission of greenhouse gases, availability and cost of fuel (if any), efficiency, energy density, cost of installation and maintenance, pollution of various kinds, dangers, the possible locations of suitable sites, and whether the energy supply is continuous.
- Solar heating panels are placed on the roofs of many homes in order to transfer radiant energy from the Sun directly to internal energy in water. The energy is then transferred using a heat exchanger to hot water which can be used in the home.
- When radiation from the Sun strikes the semiconducting material of a photovoltaic cell, electrons are released and a p.d. is produced, which can be used to transfer energy in an electrical circuit.
- Greater power can be produced by combining photovoltaic cells together in solar panels. Typical uses (usually involving rechargeable batteries) are in remote locations or in mobile equipment.
- Photo cells are also widely used to help provide energy to individual homes, and recent developments have seen them used in solar power stations connected to national grids.
- The intensity received by solar heating or photovoltaic panels varies during the day as the angle of incidence of radiation from the Sun changes (unless the device is able to rotate) and the length of atmosphere through which the radiation has to pass changes.
- For the same reasons, the power received varies during the year. Obviously, no radiation is received directly from the Sun at night and cloud cover can make the power unreliable.
- The kinetic energy gained by falling water is used to turn turbines and generate electricity in hydroelectric power stations.
- Water stored in lakes got its gravitational potential energy from the radiant energy of the Sun when water was evaporated to form clouds, which then fell as rain or snow.
- Tidal power stations work on similar principles, but use the gravitational potential energy transferred from the Earth–Moon–Sun system.
- Some hydroelectric power stations operate a pumped storage system, in which excess power is used to pump water back up to the reservoir at night.
- A wind generator transfers kinetic energy of air to the rotating blades of a turbine and then to electricity.

- Assuming that all the kinetic energy of the air is transferred to electricity, the output power can be calculated from $P = \frac{1}{2}A\rho v^3$, but some of the air must pass between the blades, so this equation gives an unrealistic maximum theoretical power. Actual efficiencies may be around 30%.
- Winds flowing across the oceans transfer an enormous amount of energy to water waves. This energy can then be transferred to electrical energy using one of a variety of different generator designs. For example, in an oscillating water column generator the changing water levels force air to flow backwards and forwards past a turbine.
- For water waves with a rectangular waveform, the power available per unit length can be calculated from $P = \frac{A^2\rho g v}{2}$.

8.5 Greenhouse effect

- The intensity of radiation at a distance r from the Sun is calculated by dividing the total emitted power by the surface area of a sphere of radius, r. $I = \frac{P}{4\pi r^2}$. Intensity is inversely proportional to distance squared.
- When radiation reaches the surface of a planet, some of it is reflected or scattered. Albedo is defined as total scattered or reflected power divided by total incident power.
- The albedo of the Earth in different places depends on the nature of the surface and whether there is any cloud cover, but it also varies daily, seasonally and with latitude. Oceans have a low albedo. Snow has a high albedo. The Earth's annual mean albedo is 0.3.
- The Earth's atmosphere keeps the planet warmer than it would be without an atmosphere. This is called the greenhouse effect. Some of the infrared radiation that is radiated away from the Earth's surface is absorbed by the atmosphere and then re-radiated in all directions, including back to the Earth's surface.
- The atmosphere also has other interconnected thermal transfer effects involving conduction, convection and evaporation.
- The atmosphere absorbs some of the infrared radiation emitted from the Earth because molecules of certain gases (mainly water vapour, carbon dioxide, methane and nitrous oxide) absorb radiation because they oscillate at the same frequencies as the radiation. This is a resonance effect. These gases occur naturally and they are known as the greenhouse gases.
- The incident radiation from the Sun is absorbed much less than that emitted from the Earth. This is because it has a different spectrum since it was emitted at a much higher temperature.
- Graphs showing the absorption or transmittance of infrared radiation through different gases help to explain the greenhouse effect.
- Human activities, especially burning fossil fuels, have increased the concentration of these gases in the atmosphere (although the percentage of water vapour has not changed significantly) and this has almost certainly led to an enhanced greenhouse effect and global warming.
- A black body is an idealized object which absorbs all the electromagnetic radiation that falls upon it. A 'perfect' absorber is also a 'perfect' emitter. The radiation emitted from a black body at different temperatures can be represented by a graph of its emission spectrum.
- The power radiated from any surface depends on its area, its (absolute) temperature, T, and the nature of its surface. $P = e\sigma AT^4$, this is called the Stefan–Boltzmann law and σ is called the Stefan–Boltzmann constant. e is the emissivity of the surface. A perfect black body has an emissivity of 1; all other surfaces have emissivities between 0 and 1.
- A planet will stay in thermal equilibrium if the average intensity of radiation that it receives, I_{in}, equals the average intensity it radiates into space, I_{out}. If these intensities are not balanced, the temperature rise ΔT in time Δt can be predicted as follows:

  $\Delta T = \frac{(I_{in} - I_{out})\Delta t}{C_s}$, where C_s is the surface heat capacity, defined as the amount of energy needed to raise the temperature of 1 m^2 of a surface by 1K. $C_s = \frac{Q}{A\Delta T}$, with units $J\,m^{-2}\,K^{-1}$.

- Computer spreadsheets (using the above equations) can be used to develop simple climate models and predict possible future outcomes.

8.6 Global warming

- The Earth is getting warmer and scientists mostly agree that it will continue to do so, although there is no agreement on how large the temperature rise will be (despite all the efforts that have gone into computer modelling possible climate changes).
- The Earth has changed temperature in the past (long before fossil fuels were ever used), but most scientists believe that the current relatively sudden global warming is due to an enhanced greenhouse effect.
- Scientists believe that this has occurred because of the increased concentrations of greenhouse gases in the atmosphere, due mainly to the combustion of fossil fuels, but also because deforestation and some other changes to land use have reduced carbon fixation (they also affect the Earth's albedo).
- Variations in the activity of the Sun (solar flares), the inclination of the Earth's axis, the orbital distance of the Earth and geological processes may all also be responsible for climate changes, but there is little convincing evidence linking these to the current trends.
- There is a reasonably close correlation between mean global temperatures and the amount of carbon dioxide in the atmosphere, (i) over hundreds of thousands of years (using ice core research) and (ii) in recent years, since detailed measurements have been made.
- After the Earth's temperature rises, there are many other interrelated factors that can have a feedback effect on global warming, including:
  - changes to climate and the amount of cloud cover
  - a reduction in albedo because of the melting of snow and ice
  - an increase in concentration of carbon dioxide in the air because of its reduced solubility in sea water
  - an increase in radiation from the surface of the Earth because it became hotter.
- Increased global temperatures will lead to a rise in mean sea-level because of (i) water flowing into the oceans from melting snow and ice (which is on land) and (ii) the expansion of water.
- The coefficient of volume expansion is defined by the following equation: $\gamma = \frac{\Delta V}{V_0 \Delta T}$. It is difficult to make reliable predictions about rising sea-levels because this coefficient varies with temperature, and any water between 0°C and 4°C (anomalously) contracts as it gets warmer.
- To reduce global warming, we should try to reduce our use of energy. Much greater use of renewable energy resources and nuclear power will also decrease the amount of fossil fuels that will need to be burned. The efficiency of power stations can be improved (using natural gas is the most efficient) and combined heat and power stations can use the thermal energy that is usually transferred wastefully to the surroundings for heating nearby buildings.
- Carbon dioxide capture and storage, and the increased use of hybrid vehicles can help to reduce the enhanced greenhouse effect.
- International efforts to reduce the enhanced greenhouse effect include the IPCC, the Kyoto Protocol and the APPCDC.

Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 It is agreed by most scientists that the cause of the enhanced greenhouse effect is

A the release of gases from erupting volcanoes.
B the combustion of fossil fuels.
C solar flares.
D changes in the Earth's orbit around the Sun.

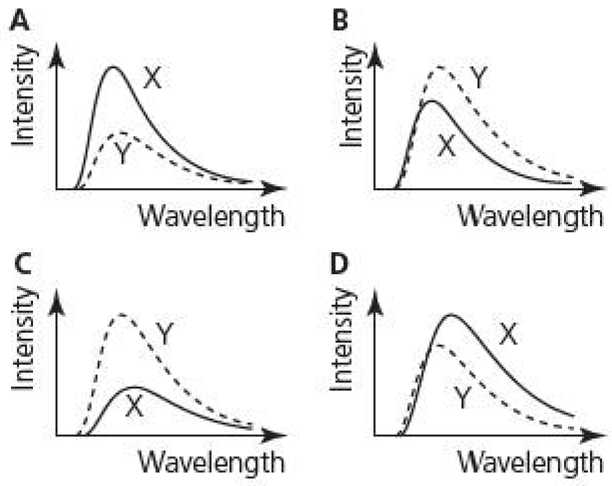
Q2 Which of the following does not affect the rate at which electromagnetic radiation is transferred away from a hot object kept at a constant temperature?

A the material from which it is made
B surface temperature
C surface area
D emissivity of the surface

Q3 Which parts of a nuclear reactor are responsible for ensuring that the rate of the reaction is kept constant?

A heat exchanger and control rods
B fuel rods and heat exchanger
C fuel rods and moderator
D moderator and control rods

Q4 Two black bodies X and Y are at different temperatures. The temperature of body Y is higher than that of body X. Which of the following shows the black-body spectra for the two bodies?



Standard Level Paper 1, Specimen Paper 09, Q29

Q5 The output power from a wind generator is 64 kW when the wind speed is 8 m s^{-1}. If the wind speed falls to 4 m s^{-1}, which of the following is the best estimate for the reduced power output?

A 32 kW
B 16 kW
C 8 kW
D 4 kW

Q6 In order to estimate the average surface heat capacity of a planet it is necessary to know

A the mass and average specific heat capacity of the planet.
B the thermal capacity and volume of the planet.
C the mass of the planet and its temperature rise in one year.
D the average temperature rise of one square metre caused by the transfer of a known amount of thermal energy.

Q7 Plutonium-239 is produced in nuclear power stations which use uranium as their fuel. Which of the following statements about this process is correct?

A The plutonium is fissile and can be used as a nuclear fuel.
B The plutonium is produced by nuclear fission.
C The plutonium is created in the moderator.
D Plutonium does not undergo radioactive decay.

Q8 Ocean waves of amplitude 1.2 m and speed 1.8 m s^{-1} generate a power of 8 kW in an oscillating water column (OWC) electricity generator. Which of the following is the best estimate for the power produced by the same generator if the wave amplitude is 2.0 m?

A 10 kW **B** 13 kW
C 22 kW **D** 37 kW

Q9 A planet with a surface temperature of 30 °C radiates power at a rate P per square metre. Another planet has a similar surface area but its temperature is 90 °C. Which of the following is the best estimate of the power radiated per square metre from the second planet?

A $1.2P$ **B** $2P$
C $3P$ **D** $9P$

Q10 The volume of a given mass of water at a temperature of T_1 is V_1. The volume increases to V_2 at temperature T_2. The coefficient of volume expansion of water may be calculated from

A $\frac{V_2}{T_2} - \frac{V_1}{T_1}$ **B** $\frac{V_2 - V_1}{T_2 - T_1}$

C $\frac{V_2 - V_1}{V_1(T_2 - T_1)}$ **D** $\frac{V_2 - V_1}{V_2(T_2 - T_1)}$

Standard Level Paper 1, Specimen Paper 1 09, Q25

Q11 Which of the following is the correct meaning of the enrichment of uranium?

A increasing the percentage of uranium-237 in the fuel
B increasing the percentage of uranium-235 in the fuel
C increasing the density of the fuel
D increasing the percentage of plutonium in the fuel

Q12 Which of the following is *not* considered to be an advantage of using hydroelectric power generated from water stored in lakes?

A The source of energy is renewable.
B Does not emit significant amounts of greenhouse gases.
C The power generation is more efficient than in fossil fuelled power stations.
D Negligible effect on the environment.

Q13 If global warming results in less snow and ice on the Earth's surface, how will this affect the Earth's albedo and energy absorption?

A The albedo and the rate of energy absorption will both increase.
B The albedo and the rate of energy absorption will both decrease.
C The albedo will decrease and the rate of energy absorption will increase.
D The albedo will increase and the rate of energy absorption will decrease.

Q14 Which form of renewable energy currently produces the greatest amount of electrical power around the world?

A hydroelectric power
B wind power
C wave power
D solar power

Q15 Which of the following is *not* a possible way to reduce the rate of global warming?

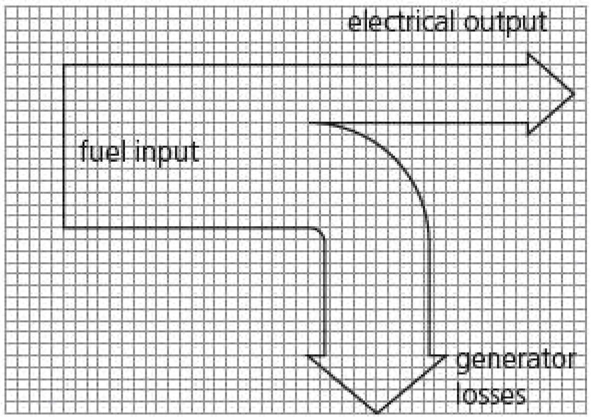
A replacing coal-fired power stations with the use of natural gas
B increasing the use of nuclear power
C using hybrid vehicles
D reducing the amount of carbon fixation

Paper 2 IB questions and IB style questions

Q1 **a** State two examples of fossil fuels. [2]

b Explain why fossil fuels are said to be non-renewable. [2]

c A Sankey diagram for the generation of electrical energy using fossil fuel as the primary energy source is shown.



Use the Sankey diagram to estimate the efficiency of production of electrical energy and explain your answer. [2]

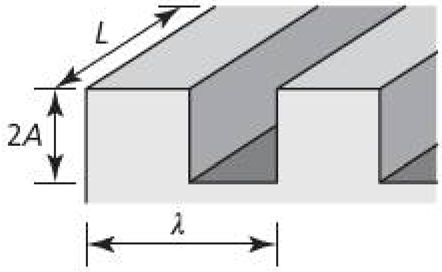
d Despite the fact that fossil fuels are non-renewable and contribute to atmospheric pollution there is widespread use of such fuels. Suggest **three** reasons for this widespread use. [3]

Standard Level Paper 2, May 09 TZ2, QB1 (Part 2)

Q2 This question is about wave power.

a Outline how the energy of a wave can be converted to electrical energy. [2]

b A wave on the surface of water is assumed to be a square wave of height $2A$, as shown.



The wave has wavelength λ, speed v and has a wavefront of length L. For this wave,

i show that the gravitational potential energy E_p stored in one wavelength of the wave is given by

$$E_p = \frac{1}{2}A^2\lambda g\rho L$$

where ρ is the density of the water and g is the acceleration of freefall. [3]

ii Deduce that the gravitational wave power P per unit length of the wavefront is given by

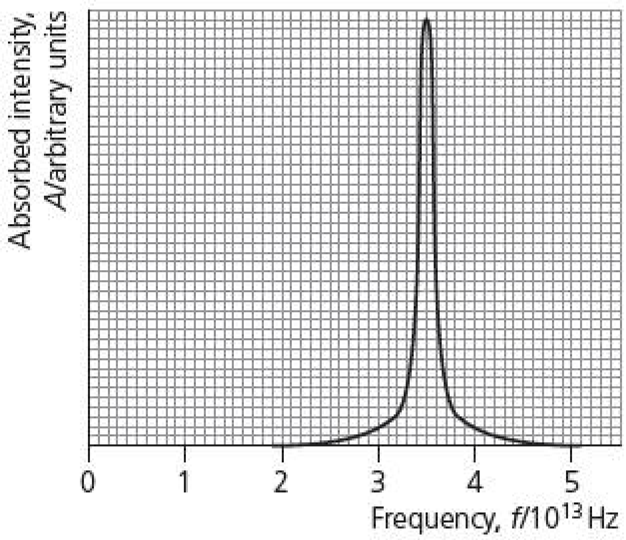
$P = \frac{1}{2}A^2 vg\rho$ [2]

c The density of sea water is $1.2 \times 10^3\,\text{kg}\,\text{m}^{-3}$. Using the expression in **b ii**, estimate the gravitational power per metre length available in a wave of height 0.60 m. [2]

d In practice a water wave is approximately sinusoidal in cross-section. Outline whether a sine wave of the same height as in **b** transfers a greater or a smaller amount of power than that derived from **b ii**. [2]

Standard Level Paper 2, May 09 TZ2, QB3 (Part 1)

Q3 a The graph shows part of the absorption spectrum of nitrogen oxide (N_2O) in which the intensity of absorbed radiation A is plotted against frequency f.

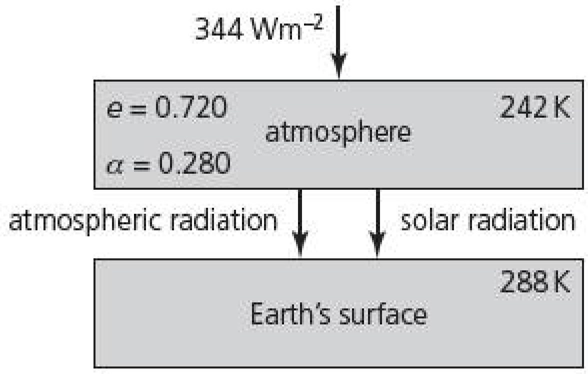


i State the region of the electromagnetic spectrum to which the resonant frequency of nitrogen oxide belongs. [1]

ii Using your answer to **a i**, explain why nitrogen oxide is classified as a greenhouse gas. [2]

b Define *emissivity* and *albedo*. [3]

c The diagram shows a simple energy balance climate model in which the atmosphere and the surface of the Earth are two bodies each at a constant temperature. The surface of the Earth receives both solar radiation and radiation emitted from the atmosphere. Assume that the Earth's surface behaves as a black body.



The following data are available for this model.

average temperature of the atmosphere of the Earth	= 242 K
emissivity, e of the atmosphere of Earth	= 0.720
average albedo, α of the atmosphere of the Earth	= 0.280
solar intensity at top of atmosphere	= $344\,\text{W}\,\text{m}^{-2}$
average temperature of the surface of the Earth	= 288 K

Use the data to show that the

i power radiated per unit area of the atmosphere is $140\,\text{W}\,\text{m}^{-2}$ [2]

ii solar power absorbed per unit area at the surface of the Earth is $248\,\text{W}\,\text{m}^{-2}$. [1]

d It is hypothesized that, if the production of greenhouse gases were to stay at its present level, then the temperature of the Earth's atmosphere would eventually rise by 6.0 K. Calculate the power per unit area that would then be

i radiated by the atmosphere [1]

ii absorbed by the Earth's surface. [1]

e Estimate, using your answer to **d ii**, the increase in temperature of Earth's surface. [3]

Standard Level Paper 2, May 09 TZ1, QB1 (Part 1)

9 Motion in fields

STARTING POINTS

- Newton's laws of motion and the equations of motion can be used mathematically to describe and predict the motion of objects moving with constant accelerations.
- Air resistance is a force which opposes the motion of objects moving through the air.
- If air resistance is negligible, the force of gravity makes all masses moving through the air accelerate towards the Earth at a rate, $g = 9.81\,\text{m s}^{-2}$.
- Changes of gravitational potential energy in a uniform gravitational field can be calculated from $\Delta E_P = mg\Delta h$. Kinetic energy, E_K, is calculated from $\frac{1}{2}mv^2$.
- When a mass moves freely upwards or downwards, there is an exchange of energy between gravitational potential energy and kinetic energy.
- If the resultant force acting on an object moving at constant speed is always perpendicular to its motion, it will undergo a centripetal acceleration and move in a circular path. The centripetal force can be calculated from $F = mv^2/r$.
- It is often useful to consider that a single vector quantity has the same effect as two independent components at right angles to each other.
- Work is done when an object is moved by a force, or a component of a force, acting in the same direction as the motion. Work done can be determined from the area under a force–distance graph.
- Newton's law of gravitation and Coulomb's law describe how forces vary with the distance between masses and charges.
- Gravitational fields exist wherever masses experience forces, g = force/mass.
- Electric fields exist wherever charges experience forces, E = force/charge.
- Fields can be represented on paper by field lines, with arrows showing the direction of the force on a mass or a positive charge.
- The electric potential difference between two points in a circuit is the ratio of energy transferred to the charge flowing between those points.

9.1 Projectile motion

In Chapter 2 we considered the motion of objects moving vertically up or down. Now we will extend that work to cover objects moving in any direction. A **projectile** is an object that has been projected through the air (fired, launched, thrown, kicked or hit, for example) and which then moves under the action of the forces of gravity and air resistance, if significant. A projectile has no ability to power or control its own motion.

Components of a projectile's velocity

9.1.1 State the independence of the vertical and the horizontal components of velocity for a projectile in a uniform field.

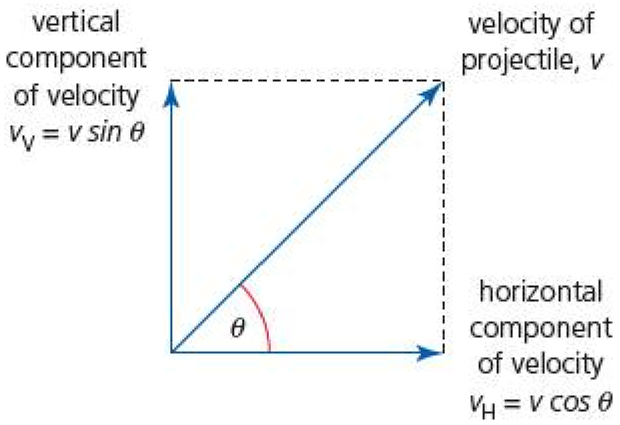


Figure 9.1 Vertical and horizontal components of velocity

The instantaneous velocity of any projectile at any time can be resolved into vertical and horizontal components, v_V, and v_H, as shown in Figure 9.1.

Because these components are perpendicular to each other, they can be treated independently (separately) in calculations.

When there is no air resistance, all objects moving through the air in the uniform gravitational field close to the Earth's surface will accelerate vertically downwards at $9.81\,\text{m s}^{-2}$ because of the force of gravity. This is true for *all* masses and for *all* directions of motion (including moving upwards). In other words, any object which is projected at any angle will always accelerate vertically downwards at the same rate as an object dropped vertically (in the absence of air resistance).

Because of the acceleration due to gravity, the values of the vertical component and the resultant velocity of a projectile will change continuously during the motion, but it is important to realize that the horizontal component will remain the same *if* the air resistance is negligible, because there are no horizontal forces acting on the projectile.

1 At one particular moment a tennis ball is moving upwards with velocity of $28.4\,m\,s^{-1}$ at an angle of 15.7° to the horizontal. Calculate the vertical and horizontal components of this velocity.

2 An aircraft is descending with a constant velocity of $480\,km\,h^{-1}$ at an angle of 2.0° to the horizontal.
  a What is the vertical component of the plane's velocity?
  b How long will it take to descend by 500 m on this flight path? (Give the answer to the nearest minute.)

3 A stone is projected upwards at an angle of 22° to the vertical. At that moment it has a vertical component of velocity of $38\,m\,s^{-1}$.
  a What is the horizontal component of velocity at this time?
  b After another second will **i** the horizontal component and **ii** the vertical component, be greater, less, or the same as before? (Ignore the effects of air resistance.)

Parabolic trajectory

9.1.2 **Describe** and **sketch** the trajectory of projectile motion as parabolic in the absence of air resistance.

Figure 9.2 shows a stroboscopic photograph of a bouncing ball. In a stroboscopic photograph the time intervals between the different positions of the ball are always the same.

The typical **trajectory** of a projectile is **parabolic** (shaped like a parabola or part of a parabola) when air resistance is negligible. For example, Figure 9.3 shows the trajectory of an object projected horizontally compared to that of an object dropped vertically at the same time. Note that both objects fall the same vertical distance in the same time.

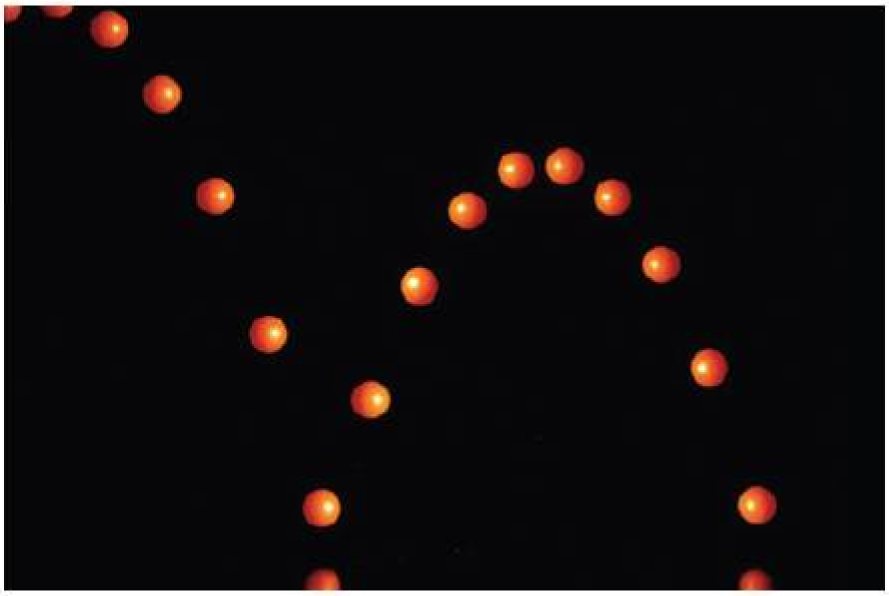


Figure 9.2 Parabolic trajectory of a bouncing ball

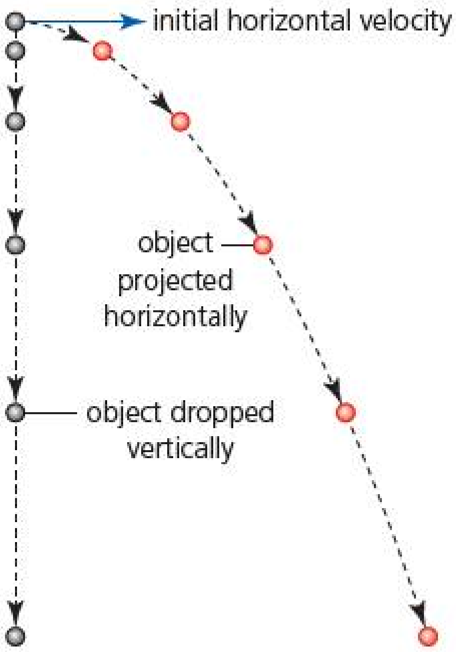


Figure 9.3 The parabolic trajectory of an object projected horizontally compared to an object dropped vertically

Effects of air resistance

9.1.3 **Describe** qualitatively the effect of air resistance on the trajectory of a projectile.

In practice, ignoring the effects of air resistance can be unrealistic, especially for smaller and/or faster moving objects. So, it is important to understand, in general terms, how air resistance affects the motion of projectiles.

Air resistance (drag) provides a force which opposes motion. Without air resistance, we assume that the horizontal component of a projectile's velocity is constant, but with air resistance it actually decreases. Without air resistance, the vertical motion always has a downwards acceleration of 9.81 m s^{-2}, but with air resistance the acceleration will be reduced for falling objects and increased for objects moving upwards.

Figure 9.4 shows typical trajectories with and without air resistance (for the same initial velocity). Note that with air resistance the path is no longer parabolic or symmetrical.

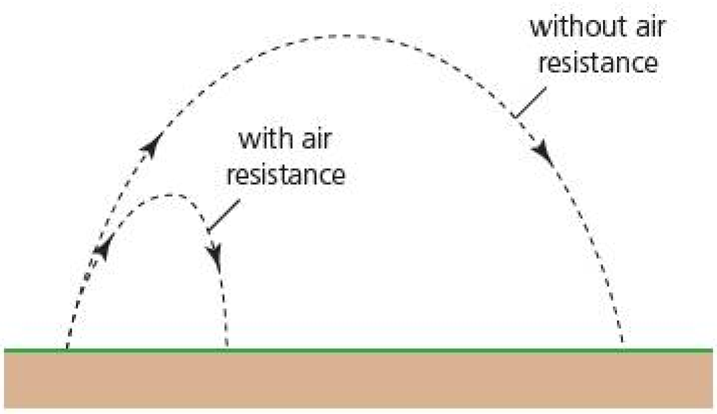


Figure 9.4 Effect of air resistance on the trajectory of a projectile

Calculations on projectile motion

9.1.4 Solve problems on projectile motion.

If the velocity (speed and direction) of any projectile moving through the air is known at any moment, then the equations of motion (see Chapter 2) can be used to determine the object's velocity at any time during its trajectory. To carry out any of these calculations, we must assume that there is no air resistance and that the downwards acceleration due to gravity is always 9.81 m s^{-2}.

The transfer of gravitational potential energy (mgh) to kinetic energy $\left(\frac{1}{2}mv^2\right)$ can sometimes provide an alternative solution to a problem – by equating these two energies we see that, for a mass falling from rest through a vertical height close to the Earth's surface:

$$gh = \frac{1}{2}v^2$$

or

$$v = \sqrt{2gh}$$

Objects projected horizontally

Worked example

1 A bullet is fired horizontally with a speed of 524 m s^{-1} from a height of 22.0 m above the ground. Calculate where it will hit the ground.

First we need to calculate how long the bullet is in the air. We can do this by finding the time that the same bullet would have taken to fall to the ground if it had been dropped vertically from rest (so $u = 0$):

$$s = ut + \frac{1}{2}at^2$$

$$22.0 = \frac{1}{2} \times 9.81 \times t^2$$

$$t = 2.12\text{ s}$$

Without air resistance the bullet will continue to travel with the same horizontal component of velocity (524 m s^{-1}) until it hits the ground 2.12 s later. Therefore:

horizontal distance travelled = horizontal velocity × time
horizontal distance = 524 × 2.12 = 1110 m

4 Make a copy of Figure 9.3 and add to it the trajectories of an object projected horizontally with:
a lower initial velocity
b higher initial velocity.

5 a Use a spreadsheet to calculate the vertical and horizontal displacements (every 0.2 s) of a stone thrown horizontally off a cliff (from a height of 48 m) with an initial velocity of 25 m s^{-1}. Continue calculations until it hits the ground.
b Plot a graph of the stone's trajectory.

6 A rifle is aimed horizontally and directly at the centre of a target which is 52.0 m away.
  a If the bullet had an initial velocity of 312 m s^{-1}, how long would it take to reach the target?
  b How far below the centre would the bullet hit the target?

Objects projected at other angles

The physics is the same for all projectiles at all angles: trajectories are still parabolic and the vertical and horizontal components remain independent of each other, but the mathematics is more complicated if the initial motion is not vertical or horizontal.

The most common problems involve finding the maximum height and the maximum horizontal distance (**range**) of the projectile.

If we know the velocity and position of a projectile, we can always use its vertical component of velocity to determine:

- the time taken before it reaches its maximum height, and the time before it hits the ground
- the maximum height reached (assuming its velocity has an upwards component).

The horizontal component can then be used to determine the range.

If the velocity at any time is needed, for example when the projectile hits the ground, then the vertical and horizontal components have to be combined to determine the resultant.

Worked example

2 A stone was thrown upwards from a height 1.60 m above the ground with a speed of 18.0 m s^{-1} at an angle of 52.0° to the horizontal. Assuming that air resistance is negligible, calculate:
  a its maximum height
  b the vertical component of velocity when it hits the ground
  c the time taken to reach the ground
  d the horizontal distance to the point where it hits the ground
  e the velocity of impact.

First we need to know the two components of the initial velocity:

$$u_V = u \sin\theta = 18.0 \sin 52.0° = 14.2 \text{ m s}^{-1}$$
$$u_H = u \cos\theta = 18.0 \cos 52.0° = 11.1 \text{ m s}^{-1}$$

a Using $v^2 = u^2 + 2as$ for the upwards vertical motion (with directions upwards considered to be positive), and remembering that at the maximum height $v = 0$, we get:

$$0 = 14.2^2 + [2 \times (-9.81) \times s]$$
$s = +10.3$ m above the point from which it was released; a total height of 11.9 m.

(Using $\frac{1}{2}v^2 = gh$ is an alternative way of performing the same calculation.)

b Using $v^2 = u^2 + 2as$ for the complete motion gives:

$$v^2 = 14.2^2 + [2 \times (-9.81) \times (-1.60)]$$
$v = 15.3 \text{ m s}^{-1}$ downwards

c Using $v = u + at$ gives:

$$-15.3 = 14.2 + (-9.81)t$$
$$t = 3.0 \text{ s}$$

d Using $s = vt$ with the horizontal component of velocity gives:

$$s = 11.1 \times 3.0 = 33.3 \text{ m}$$

e Figure 9.5 illustrates the information we have so far and the unknown angle and velocity.

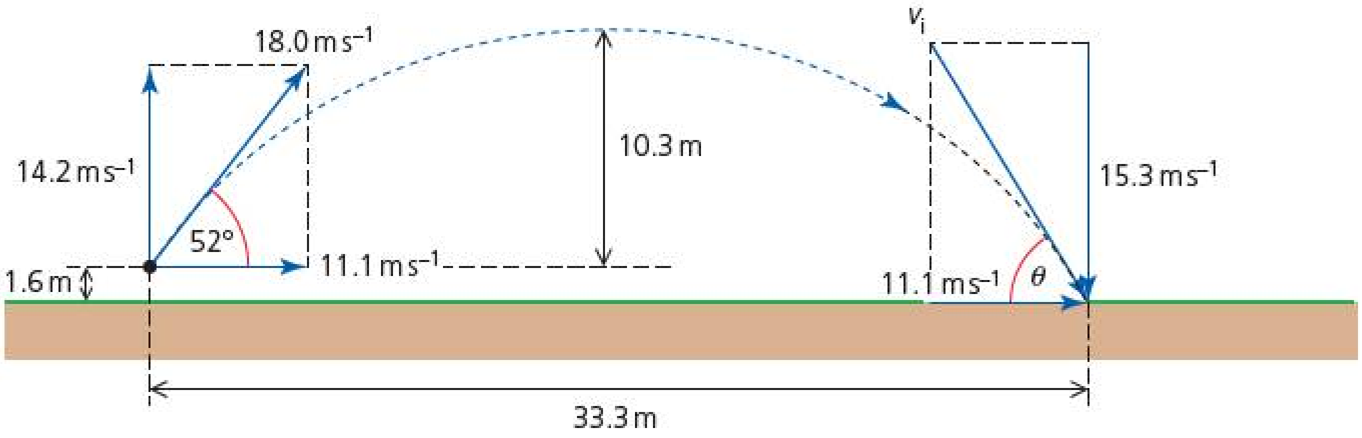


Figure 9.5

From looking at the diagram, we can use Pythagoras' theorem to calculate the velocity of impact.

(velocity of impact)^2 = (horizontal component)^2 + (vertical component)^2

$v_i^2 = 11.1^2 + 15.3^2$

$v_i = 18.9\,\text{m s}^{-1}$

The angle of impact with the horizontal, θ, can be found using trigonometry:

$$\tan\theta = \frac{15.3}{11.1}$$

$$\theta = 54.0°$$

7 Repeat Worked example 2 for a stone thrown with a velocity of 26 m s⁻¹ at an angle of 38° to the horizontal from a cliff top. The point of release was 33 m vertically above the sea.

8 The maximum theoretical range for a projectile occurs when it is projected at an angle of 45° to the ground (once again, ignoring the effects of air resistance). Calculate the maximum distance a golf ball will travel before hitting the ground, if its initial velocity is 72 m s⁻¹. (Your answer will be much greater than the actual ranges achieved by top class golfers.)

9 A jet of water from a hose is aimed directly at the base of a flower as shown in Figure 9.6. The water emerges from the hose with a speed of 3.8 m s⁻¹.

a Calculate the angle θ and the vertical and horizontal components of the initial velocity of the water.

b How far away from the base of the plant does the water hit the ground?

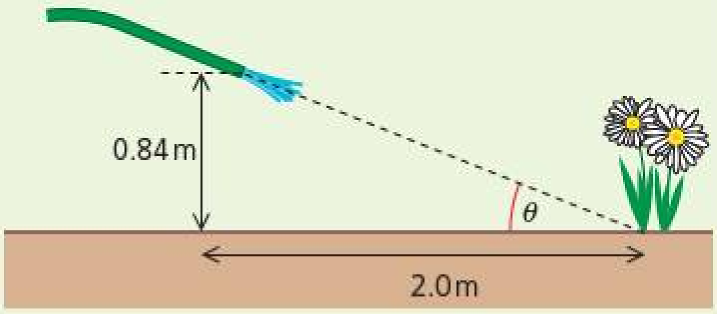


Figure 9.6

10 A ball rolls down the slope shown in Figure 9.7 without friction and is then projected horizontally off the table top at point P.

a Show that the maximum range of the ball is given by $R = 2\sqrt{h_1 h_2}$

b What assumption(s) did you make?

c Explain why your answer to **a** did not depend on the mass of the ball.

11 If the maximum distance a man can throw a ball is 78 m, what is the minimum speed of release of the ball? (Assume that the ball lands at the same height from which it was thrown and that the greatest range for a given speed is when the angle is 45°.)

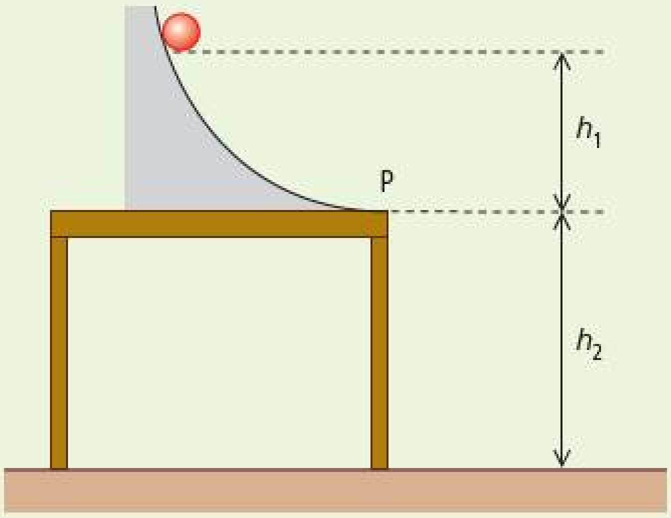


Figure 9.7

Additional Perspectives

Projectiles in sport

Many sports and games involve some kind of object (often a ball) being thrown, kicked or hit through the air. Obvious examples are basketball (see Figure 9.8), tennis, football, badminton, archery, cricket and golf. The skill of the players is to make the ball, or other object, move with the right speed and trajectory, and often to also be able to judge correctly the trajectory of an object moving towards them.

Figure 9.8 A basketball moves in approximately parabolic flight

Figure 9.9 Badminton trajectories are not parabolic

The law of conservation of momentum can be very useful for predicting the initial velocity of the object but, in a ball game, the design of the ball is the all-important factor. The mass, shape, diameter and the nature of its surface will all affect the way the ball moves through the air after it has been 'projected'. Although in most sports it can be assumed that the ball will follow an approximately parabolic path, if the ball always had a perfectly parabolic trajectory, the game would be predictable and less skilful. The effect of the air moving over the surface of the ball plays an important part in many sports and good players can use this to their advantage by putting spin on the ball. There is a difference in air pressure on opposite sides of a spinning ball, producing a force that affects the direction of motion.

Part of the fun of playing or watching sports is to see a ball being struck or thrown with such skill that it travels with great accuracy and speed, or goes a large distance. It is interesting to consider how the design of the balls in different ball games has evolved.

Badminton is an unusual sport because the design of the shuttlecock produces non-parabolic trajectories (see Figure 9.9). A shuttlecock has a small mass for its cross-sectional area, which means that it can travel very fast when it is first hit; after that, air resistance has a significant effect, considerably reducing its range. Most of a shuttlecock's mass is concentrated in the 'cork' at the opposite end to the feathers, so that it always moves in flight such that the cork leads the motion.

Questions

1 In which sport can the struck ball travel the greatest distance? Find out if there are any regulations in that game which try to limit how far the ball can travel.
2 If the balls from a variety of different sports were all dropped from the same height onto the same hard surface, which one would bounce up to the greatest height? Discuss possible reasons why that ball loses the smallest fraction of its energy when colliding with the surface and why that is important for the sport in which it is used.
3 Research into an explanation of how spin can cause a ball to change direction.

9.2 Gravitational field, potential and energy

9.2.1 **Define** gravitational potential and gravitational potential energy.

We say that there is **gravitational potential energy** stored in any system of masses because, at some time in the past, energy was transferred (work was done) when the masses were moved to their present positions.

If gravitational forces exist between two or more masses and they are moved so that the distances between them change, then work has to be done. We refer to this as a *change* in gravitational potential energy. For example, we may say that when a certain book was raised away from the Earth, it gained 2 J of gravitational potential energy. We often talk about the (changes in) gravitational potential energy of an object, but the gravitational energy we refer to is really stored, not in a single object, but in the *system* that is the object and the Earth combined.

We are usually concerned with changes in gravitational potential energy, but if we wish to define the total gravitational potential energy in a system, we first need to define our starting point.

The total gravitational potential energy of a system is defined as the work done when bringing all the masses of the system to their present positions, assuming that they were originally at infinity.

In this definition, infinity is chosen to be the place where gravitational potential energy is zero. The work that would be done in moving a mass from infinity is *not* infinite, because the gravitational forces involved become very, very small (negligible) when the distances become large.

In this course we will not be concerned with the total gravitational energy of complex systems of many masses, and our discussions will be mainly limited to the gravitational potential energy of single objects located in the gravitational field of the Earth, or another planet.

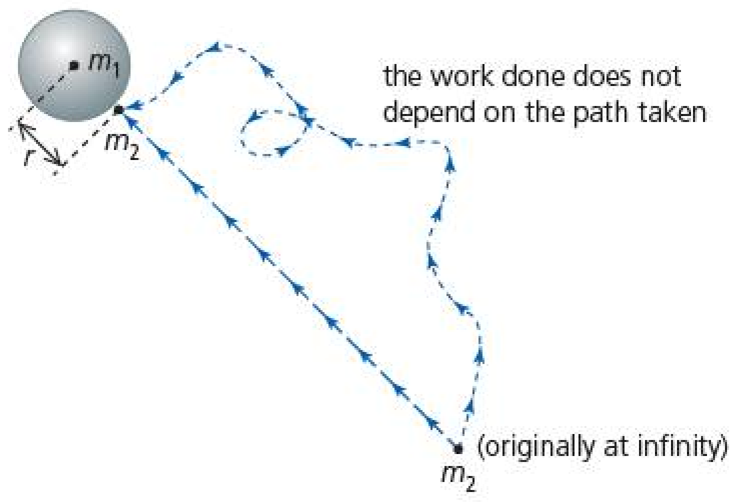


Figure 9.10 Work done in moving a mass from infinity

Consider Figure 9.10 which represents a smaller mass, m_2, being moved from infinity close to a larger mass, m_1.

The gravitational potential energy of the system can be determined by calculating the work done in bringing m_2 to its current position from infinity (assuming that any movement of m_1 is insignificant). From Chapter 2, we know that

work done = force × distance moved in the direction of the force

This is not a straightforward calculation because the magnitude of the force varies with the distance between the centres of the masses (Newton's universal law of gravitation), and also because the distance is infinite. However, the work done can be found from the area under a force–distance graph, as shown in Figure 9.11.

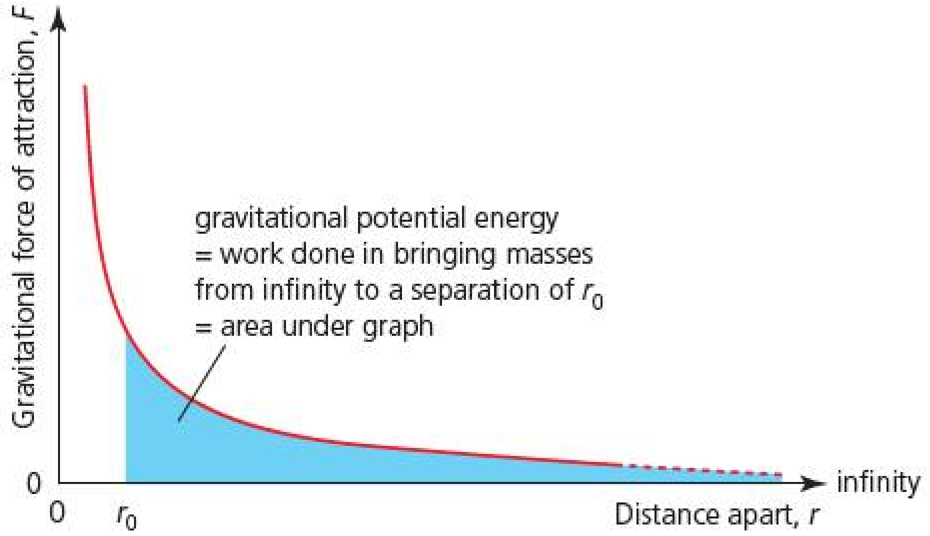


Figure 9.11 Finding the work done in bringing a mass from infinity to a separation r_0

The total amount of work done is independent of the path taken. This means that, whatever path is taken between two points, the work done will be the same.

Using equations is easier than graphically determining areas and it can be shown that the gravitational potential energy, E_p, stored between two masses, m_1 and m_2, separated by a distance r, can be calculated from the following equation:

$$E_p = \frac{-Gm_1m_2}{r}$$

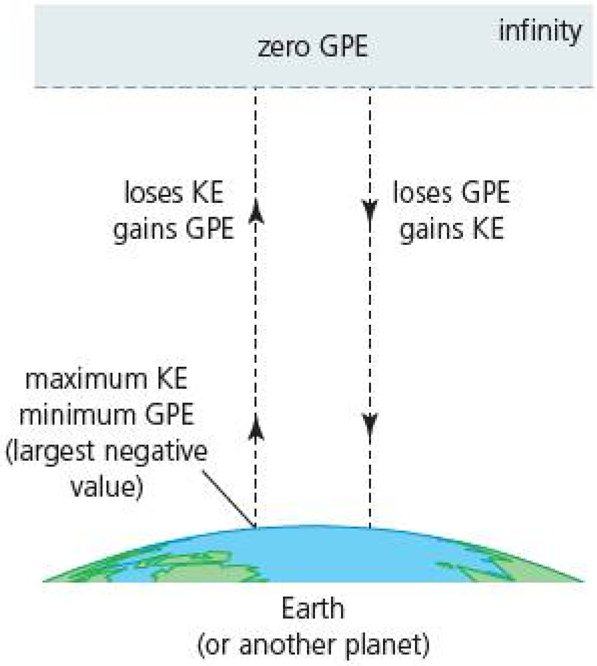


Figure 9.12 Changes of energy when moving between a planet and infinity

The negative sign is important here; it represents the fact that infinity is chosen to be the place where gravitational potential energy is zero. *All* gravitational potential energies are negative, but *changes* in gravitational potential energy can be positive or negative.

If any mass, a ball for example, is projected upwards from the Earth's surface, it loses kinetic energy (KE) and gains gravitational potential energy (GPE). This is also true for an unpowered spacecraft and is even true for the theoretical extreme in which a mass travels all the way 'up' to infinity. For the same reasons, a theoretical mass 'falling' from infinity would gain kinetic energy and lose gravitational potential energy (as shown in Figure 9.12), so that its minimum gravitational energy (on the Earth's surface) corresponds to the largest negative value.

Worked example

3 Calculate the gravitational potential energy of a 1.00 kg mass on the Earth's surface.

$$E_p = -\frac{Gm_1m_2}{r}$$

$$E_p = -\frac{\text{G} \times \text{mass of Earth} \times 1}{\text{radius of Earth}}$$

$$E_p = -\frac{(6.67 \times 10^{-11}) \times (5.97 \times 10^{24})}{6.36 \times 10^6}$$

$$E_p = -6.26 \times 10^7\,\text{J}$$

This means that a 1.00 kg mass would need to be given 6.26×10^7 J to move it to infinity (ignoring the effects of air resistance in the Earth's atmosphere).

12 Calculate the gravitational potential energy of a 69 kg astronaut on the Moon's surface (mass of Moon = 7.3×10^{22} kg, radius of Moon = 1.7×10^6 m).

13 Consider Figure 9.10 again. Explain why the work done in moving the mass along the curved path is the same as moving it in a straight line between the same two points.

Calculating changes in gravitational potential energy

To calculate a change of gravitational potential energy *close* to the surface of the Earth, we can use the equation $\Delta E_p = mg\Delta h$ because we can assume that the value of the gravitational field strength, g, is constant ($9.81\,\text{N kg}^{-1}$). However, when much greater distances from the Earth are involved, this equation can no longer be used because g varies. Instead, the equation $E_p = -Gm_1m_2/r$ should be used for the two separate locations and the difference calculated, as shown in Worked example 4.

Worked example

4 a Use $\Delta E_p = mg\Delta h$ to calculate the change in gravitational potential energy when a 100 kg mass is raised 500 km above the Earth's surface.

b Calculate the gravitational potential energy of a 100 kg mass on the Earth's surface and 500 km above it.

c What is the change in gravitational potential energy if the mass is moved between these places?

d Compare your answers to **a** and **b**.

a $\Delta E_p = mg\Delta h = 100 \times 9.81 \times 500\,000 = 4.91 \times 10^8$ J

b Using $E_p = -\frac{Gm_1m_2}{r}$

On the Earth's surface,

$$E_p = -\frac{(6.67 \times 10^{-11}) \times (5.97 \times 10^{24}) \times (100)}{6.36 \times 10^6}$$

$$E_p = -6.26 \times 10^9 \text{ J}$$

At a height of 500 km,

$$E_p = -\frac{(6.67 \times 10^{-11}) \times (5.97 \times 10^{24}) \times (100)}{(6.36 \times 10^6) + (5 \times 10^5)}$$

$$E_p = -5.80 \times 10^9 \text{ J}$$

c The change in gravitational potential energy $= (-5.80 \times 10^9) - (-6.26 \times 10^9)$ J
$= +4.6 \times 10^8$ J

d Using the equation $\Delta E_p = mg\Delta h$ with a constant value for g has given a value for the change of gravitational potential energy which is too high. It is an overestimate of about 7%. Of course, for larger distances the errors involved with using $\Delta E_p = mg\Delta h$ become much greater.

14 a Calculate how much gravitational potential energy must be transferred to a 2500 kg mass to put it in an orbit which is 3.2×10^7 m above the Earth's surface.

b Give at least two reasons why the actual amount of energy transferred when putting such a satellite in orbit is much greater than the answer to **a**.

15 Explain why there is no change in gravitational potential energy for a mass which is moved horizontally on the Earth's surface, or for a satellite moving in a circular orbit around the Earth.

Additional Perspectives

Launching satellites

To send something up into space so that it stays up and does not come down again was a dream of scientists for centuries. The first artificial satellite was sent into orbit by Russia on 4 October 1957 and this truly historic event has rightly been called the start of the 'space age'. The satellite had a mass of 84 kg and it orbited the Earth at an approximate average height of 570 km and speed of 29 000 km h^{-1}. It was called 'Sputnik 1' (Figure 9.13). Three months after launching, Sputnik 1 burned up as it re-entered the Earth's atmosphere. Soon afterwards, the launch of Sputnik 2 carried the first animal into space, a dog called Laika (see Figure 9.14). She survived the launch but died in orbit a few hours after launch.

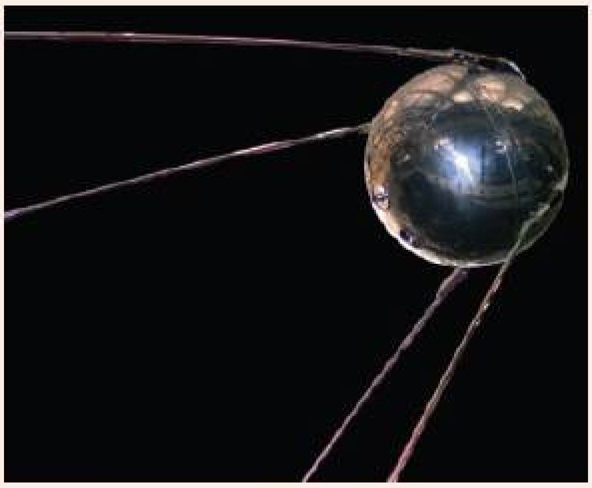


Figure 9.13 Sputnik 1, the first satellite

Figure 9.14 Laika: the first animal in space

In simple terms, the major challenge was to provide sufficient energy to the Sputnik satellite to get it up above the Earth's atmosphere and moving in the right direction at the right speed.

Questions

1 Calculate the gravitational potential energy of the Sputnik satellite on the Earth's surface and in orbit.

2 a Calculate the gravitational potential energy that has to be gained by the satellite for it to reach a height of 570 km.

b Calculate the satellite's kinetic energy in orbit, and therefore its total increase in energy compared to being stationary on the Earth's surface.

Your final answer to question **b** should be about 4×10^9 J. Much more energy than this, however, needs to be transferred since a large amount of energy will be dissipated due to air resistance and because the engines are inefficient. It is not possible to simply project a satellite into space, so the launch has to be powered by rocket engines. One of the requirements of rocket fuel is that a large amount of energy can be transferred from a relatively small mass of fuel. For example, if 35 MJ is the theorectical maximum amount of energy that can be released by the combustion of 1 kg of fuel and oxidant, then more than 100 kg of fuel and oxidant would be needed to supply 4×10^9 J. Of course, more energy would then be needed to raise the fuel itself. These calculations totally ignore the mass of the engine and the tanks needed for the fuel and oxidant. The total mass of the rocket reduces considerably as the fuel and oxidant are burned and, when they are empty, the tanks can be discarded.

A typical useful 'lifetime' of a satellite is about ten years. Currently there are about 3000 satellites in use around the Earth. Russia and the USA have the most, but many other countries, including China (Figure 9.15), have launched their own satellites. Since 1957 about 25 000 objects have been recorded in orbit. Most have burned up after a few years as their orbits re-entered the Earth's upper atmosphere, but many are still in orbit and commonly described as 'space junk'.

Figure 9.15 The launch of a satellite in China

Questions

3 Find out what fuel is used for rocket launches and the reasons why it is chosen.

4 Find out which countries currently have working satellites in orbit.

5 Find out if there is any country or organization which controls the use of the space around the Earth.

Gravitational potential

When calculating gravitational potential energies we refer to a particular mass in a particular place, but if we want to ask the question 'how much energy would be needed to put a mass in a certain location?' we need to use the important concept of **gravitational potential**. (Potential energy and potential are not the same thing, although they are closely connected.)

The gravitational potential at a point is defined as the work done per unit mass in bringing a small test mass from infinity to that position.

In other words, gravitational potential is the gravitational potential energy per unit mass. It is given the symbol V and has the units of joules per kilogram, J kg^{-1}. Potential is a scalar quantity.

$$V = \frac{E_p}{\text{mass}}$$

Gravitational potential around a point mass

9.2.2 State and **apply** the expression for gravitational potential due to a point mass.

Consider a test mass m_2 being brought near to a mass m_1, as in Figure 9.10. We know that the gravitational potential energy is given by the expression:

$$E_p = -\frac{Gm_1m_2}{r}$$

So, the potential can be calculated from:

$$V = \frac{E_p}{\text{mass}} = -\frac{Gm_1}{r}$$

Replacing m_1 with m, we get the general expression for the gravitational potential around a point mass, m:

$$V = -\frac{Gm}{r}$$

This equation is given in the IB *Physics data booklet*.

This expression can also be used for the potential around spherical masses, such as planets. As with gravitational potential energies, gravitational potentials are always negative because moving to infinity requires an increase of gravitational potential energy and potential, since the gravitational potential energy and potential at infinity are defined to be zero.

From Chapter 6 we know that $g = \frac{Gm}{r^2}$, so we can also write:

$$V = -gr$$

Worked example

5 Calculate the gravitational potential on the Earth's surface.

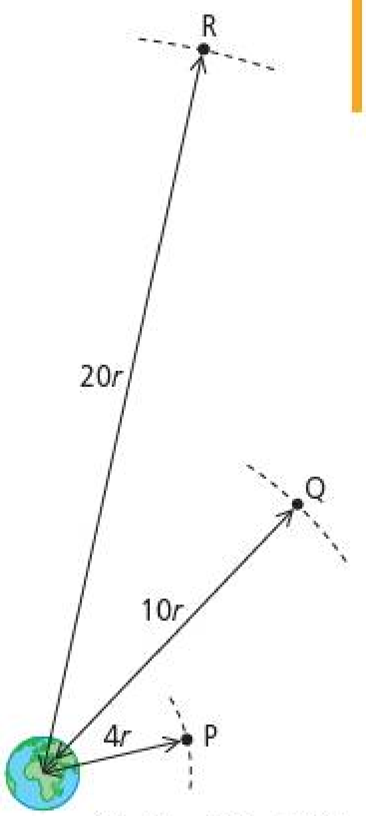
$$V = -\frac{Gm}{r}$$

$$V = -\frac{(6.67\times10^{-11})\times(5.97\times10^{24})}{6.37\times10^{6}}$$

$$V = -6.25\times10^{7}\,\text{J kg}^{-1}$$

This means that 6.25×10^7 J would have to be given to 1 kg to move it to infinity from the Earth's surface.

(Alternatively: $V = -gr = -9.81\times(6.37\times10^6) = -6.25\times10^7\,\text{J kg}^{-1}$)



mass of Earth = 6.0×10^{24} kg
radius of Earth, $r = 6.4\times10^6$ m

Figure 9.16

9.2.9 Solve problems involving gravitational potential energy and gravitational potential.

16 a Calculate the gravitational potential 5000 km above a planet of mass 1.1×10^{25} kg and radius 7500 km.

b How much energy would be needed to move a spacecraft of 12 000 kg from that position to a place which was a very, very long way away from the planet?

17 a Explain exactly what we mean when we say that gravitational potential on the Earth's surface is $6.25\times10^7\,\text{J kg}^{-1}$ (at a distance r from its centre).

b Calculate the gravitational potential at the points P, Q and R in Figure 9.16.

c How much gravitational potential energy would have to be given to a 2420 kg spacecraft to move it from Q to R?

18 What is the gravitational field strength on the surface of a planet of radius 2.89×10^7 m if the gravitational potential there is $-1.45\times10^8\,\text{J kg}^{-1}$?

19 Use a spreadsheet to calculate the gravitational potential at different distances from the centre of the Earth (every 10 000 km from zero to 100 000 km). Use the computer program to generate a graph of the results.

Gravitational potential differences and potential gradients

9.2.3 **State** and **apply** the formula relating gravitational field strength to gravitational potential gradient.

We are usually much more concerned with *difference* in potential (**potential difference**, ΔV) between two places, than the actual potential at any location.

The gravitational potential difference, ΔV, between two points is the work done when moving a unit mass between those two points.

Assuming that g is constant over the distance Δr, the equations $V = E_p/m$ and $V = -gr$, can be written in the following related forms, which are both given in the IB *Physics data booklet*:

$$\Delta V = \frac{\Delta E_p}{m}$$

and

$$g = -\frac{\Delta V}{\Delta r}$$

The second of these equations can be expressed as:

gravitational field strength is equal to (the negative of) the **potential gradient.**

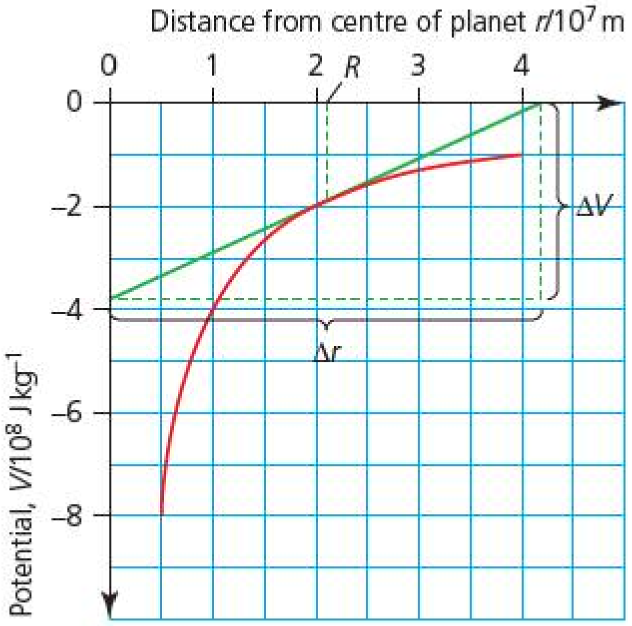


Figure 9.17 The graph shows the variation of gravitational potential around a planet

Figure 9.17 shows an example of how the gravitational potential varies around a planet. The field can be found at any distance (for example at R) by determining the gradient of the graph at that distance, as shown.

9.2.9 **Solve** problems involving gravitational potential energy and gravitational potential.

20 The change in gravitational potential energy when a 1200 kg satellite moved between different circular orbits was -7.5×10^9 J.
a Did the satellite move closer to, or further away from the Earth?
b What was the potential difference between these two orbits?

21 Consider Figure 9.17. Determine the gravitational field strength at distance of 1.5×10^7 m from the centre of the planet.

Combining potentials

9.2.4 **Determine** the potential due to one or more point masses.

We have already seen how to calculate the potential around a single point mass or a sphere, but sometimes two or more large masses, like planets and moons, are close enough together that they create significant gravitational fields around each other. We know from Chapter 6 that determining the resultant gravitational field at any location is done by adding the individual fields (taking their directions into account because field is a vector quantity). Similarly, the overall gravitational potential can also be determined by adding individual potentials. But, since potential is a scalar quantity, this is just an addition without any reference to direction.

Worked example

6 Figure 9.18 shows a planet, P, and its moon, M (not to scale).

a Calculate the potentials at N due to both of these masses.

b What is the overall potential at N?

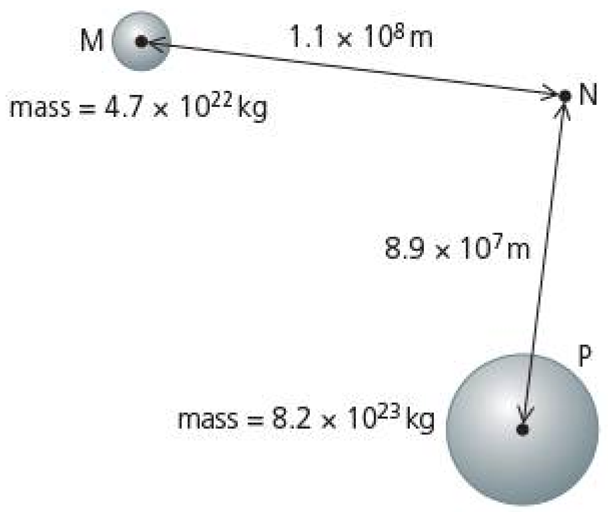


Figure 9.18

a Using:

$$V = -\frac{Gm}{r}$$

The gravitational potential due to

$$\text{M at point N} = -\frac{(6.67\times10^{-11})\times(4.7\times10^{22})}{1.1\times10^{8}}$$
$$= -2.8 \times 10^4\,\text{J kg}^{-1}$$

The gravitational potential due to

$$\text{P at point N} = -\frac{(6.67\times10^{-11})\times(8.2\times10^{23})}{8.9\times10^{7}} = -6.1\times10^{5}\,\text{J kg}^{-1}$$

b Overall gravitational potential at N = $(-2.8 \times 10^4) + (-6.1 \times 10^5) = -6.4 \times 10^5\,\text{J kg}^{-1}$
The combined gravitational potential must be greater than the individual gravitational potentials because it would take more energy to move a mass a long way away from both the planet and the moon, compared to moving the mass away from either on its own.

9.2.9 **Solve** problems involving gravitational potential energy and gravitational potential.

22 a Use a spreadsheet to calculate the combined gravitational potential every 40 000 km on a straight line between the Earth and the Moon.
b Where does the gravitational potential have its lowest value?
c What can you say about the gravitational field at that point?

Equipotential surfaces and lines

9.2.5 **Describe** and **sketch** the pattern of equipotential surfaces due to one and two point masses.
9.2.6 **State** the relation between equipotential surfaces and gravitational field lines.

We would not normally expect the gravitational potential (or field) around a large mass, like a planet or star, to change, and that is one reason why the concept of potential is so useful. We can draw 'energy maps' to represent how the gravitational potential varies around planets and similar bodies. Figure 9.19 shows **equipotential** lines around a spherical planet, which acts like a point mass. An equipotential line is a line connecting places with the same potential.

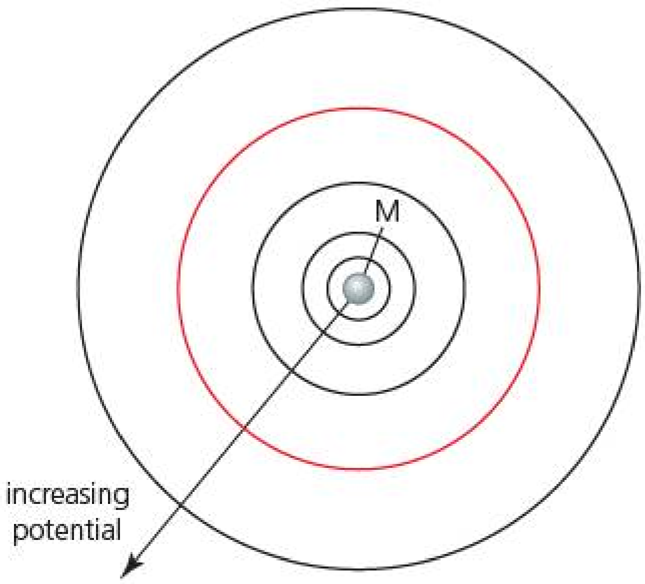


Figure 9.19 Equipotential lines around a point mass

The same amount of energy would be needed to take a certain mass located at *any* point on the red equipotential line to infinity, and no *net* energy would be needed to move the mass from any place on the red line to any other place on the same equipotential line. Since potential is a scalar quantity, there are no arrows on the lines.

The difference in gravitational potential between *successive equipotential lines* is always kept the same, which means that the lines get further apart as the distance from the mass increases (because the gravitational field is getting weaker). Figure 9.19 represents a three-dimensional situation on a flat piece of paper; in three dimensions the equipotential *surfaces* in this example would be spherical. In Chapter 6 we discussed the use of field lines to represent gravitational fields. Drawings of equipotential lines of the same field do not provide any different information, but they do provide another way of visualizing the same field.

The force on any small mass placed near to the big mass, M, in Figure 9.19 will act *radially* in towards M, perpendicular to the equipotential lines.

Field lines are always perpendicular to equipotential lines, pointing from higher potential to lower potential, as shown in Figure 9.20.

Figure 9.20a shows a radial gravitational field and Figure 9.20b shows a uniform gravitational field, such as in a room. (We usually consider that the gravitational field we experience on the Earth's surface is uniform, but in reality what we experience is just a tiny part of an enormous radial field.) The field is always strongest where the field lines and the equipotential lines are closest together.

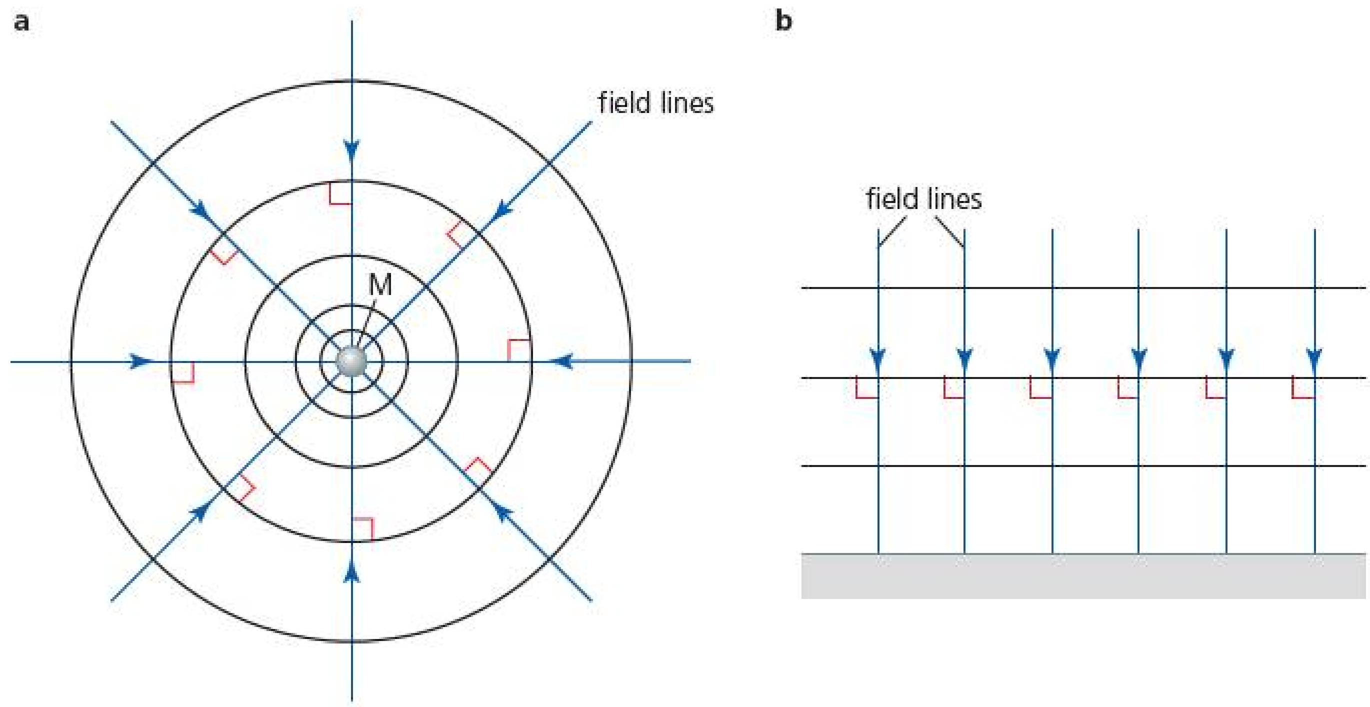


Figure 9.20 Equipotential lines and field lines are perpend

The shape of the equipotential lines around two equal masses is shown in Figure 9.21.

Contour lines drawn on a geographical map, such as shown in Figure 9.22, are useful for the same reason that equipotential lines are useful. In fact contour lines are a type of equipotential line. Contour lines join up places at the same height, so that, by looking at the map, we know in which direction to move to go up (to gain gravitational energy), to go down, or to stay at the same level. We also know that the ground is steepest where the lines are closest together and that anything free to move, like a river, will go downwards, perpendicular to the contour (equipotential) lines.

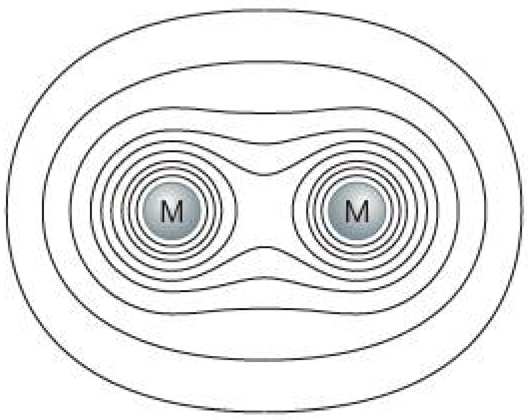


Figure 9.21 Equipotential lines around two equal masses

Figure 9.22 Contour lines on a map

23 **a** Make a copy of Figure 9.21 and add to it gravitational field lines.
b Mark where the gravitational field strength is zero.

24 Draw a sketch of the equipotential lines around two masses which are not equal.

Additional Perspectives

The Moon

We can see the Moon because it reflects light from the Sun. As the Moon orbits the Earth its appearance (phase) changes because, depending on the relative positions of the Sun, Earth and Moon, the light from the Sun falls on a changing proportion of the side of the Moon which is facing the Earth. When we can see a complete circular moon, or 'full moon', the whole of the side that is facing the Earth is illuminated; this occurs every 29.53 days. In fact, the Moon orbits the Earth every 27.32 days, but the repeating lunar cycle is longer because it depends on the relative positions of the Sun, Earth and Moon, and not just the rotation of the Moon.

The regular phases of the Moon have long been the basis for calendars and the timing of annual events, festivals and holidays.

Questions

1 Draw a diagram to illustrate the different phases of the Moon.
2 Name three different events in different cultures or religions that occur at a particular time depending on the phase of the Moon. Explain how the exact date of one of these events is decided.
3 Do you believe that a full moon can have an effect on human or animal behaviour? Give a scientific reason for your answer.
4 Calculate the gravitational attraction between the Moon and the Earth.

Gravitational attraction keeps the Moon in orbit around the Earth and this same force (of the order of 10^{20} N) also affects the Earth. The most noticeable effect is the tides in the Earth's oceans. As the Earth spins on its axis, water in the oceans is pulled closer to the Moon, causing a high tide. At the same time, a high tide also occurs on the opposite side of the Earth, resulting in high tides (and low tides) about twice every day in any particular location. Boats that are in the water at high tide can become stranded when the water level goes down at low tide (Figure 9.23).

The Sun also affects the tides. When the Earth, Sun and Moon are in an approximate straight line (at a new moon or full moon), the gravitational effects are greatest and the tides highest.

Figure 9.23 Boats stranded at low tide in Vietnam

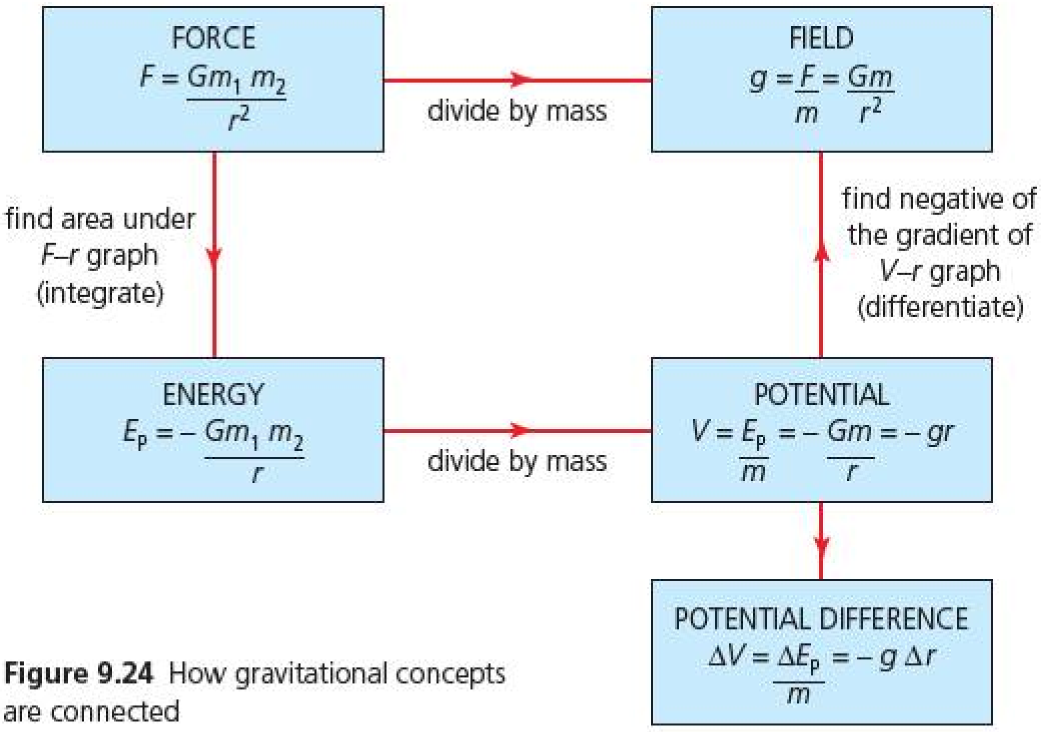


Figure 9.24 How gravitational concepts are connected

Figure 9.24 summarizes the gravitational concepts we have met in this section.

Escape speed from a planet

9.2.7 **Explain** the concept of escape speed from a planet.
9.2.8 **Derive** an expression for the escape speed of an object from the surface of a planet.

In theory an object can be *projected* (not powered) upwards in such a way that it could continue to move away from Earth for ever. For this to be possible the object would need to be given a very high speed. To calculate that speed we need to consider energies. In general, the initial kinetic energy given to an object will be transferred to gravitational energy and dissipated due to air resistance in the Earth's atmosphere.

If we assume that the effects of air resistance are negligible, we can calculate the *minimum* theoretical speed that a projectile of mass m_2 needs to 'escape' from a planet of mass m_1. This is called its **escape speed**, v_{esc}. It can be calculated as follows:

initial kinetic energy = change in gravitational potential energy between location and infinity

$$\frac{1}{2}m_2 v_{esc}^2 = 0 - \left(-\frac{Gm_1m_2}{r}\right)$$

$$v_{esc} = \sqrt{\frac{2Gm_1}{r}}$$

Note that the theoretical value for the escape speed does not depend on the mass or the direction of motion.

Worked example

7 Calculate the escape speed of a projectile launched from the Earth's surface. (Ignore any effects due to the Moon.)

$$v_{esc} = \sqrt{\frac{2Gm_1}{r}}$$

$$v_{esc} = \sqrt{\frac{2 \times 6.67 \times 10^{-11} \times 5.97 \times 10^{24}}{6.37 \times 10^{6}}}$$

$$v_{esc} = 1.12 \times 10^{4}\,\text{m s}^{-1}$$

This means that, ignoring the effects of air resistance, any mass projected in any direction from the Earth's surface with a speed of $11.2\,\text{km s}^{-1}$ will never come back down to Earth. This speed is theoretically enough to take it to infinity, where its speed would reduce to zero. If it was launched with a greater speed, the mass would still be moving when it reached infinity.

Of course, it is not sensible to assume that there is no air resistance for an object moving at high speed through the Earth's atmosphere, so this calculated value is just a theoretical minimum. However, on a planet or a moon which has no atmosphere, or if the starting point was already above the Earth's atmosphere, the calculations are valid. Such calculations also ignore any possible effects of other gravitational fields from other planets, moons and stars.

It is possible to give an object an initial speed of $11.2\,\text{km s}^{-1}$, but it is not possible to project any object with the much greater speed and kinetic energy needed to overcome all the resistive forces of the atmosphere. Spacecraft put into orbit around the Earth are *not* projectiles; they are transported away from the Earth's surface by continuous forces from rocket engines which act over extended periods of time.

9.2.9 Solve problems involving gravitational potential energy and gravitational potential.

25 **a** Calculate the escape speed from the Moon's surface (mass = 7.4×10^{22} kg and radius = 1.7×10^{6} m).
b Explain why the actual escape speed is a little higher than your answer.
c Explain why a space vehicle launched from the Moon could reach the Earth if it was given a smaller initial speed.

26 Calculate the escape speed from an orbit 350 km above the Earth's surface.

27 How far away from Earth does the escape velocity reduce to 1% of its value on the Earth's surface?

28 **a** The mass of the Sun is 2.0×10^{30} kg. Calculate how small its radius would have to be if the escape speed from its 'surface' was to be equal to the speed of light (3×10^{8} m s^{-1}), so that it might become a black hole.
b Compare your answer to the actual radius of the Sun.

29 Consider the combined effect of the Earth's and the Moon's gravitational fields on the Earth's surface: how much greater than 1.12×10^{4} m s^{-1} is the escape speed from the Earth? Can the effect of the Moon be ignored?

30 Calculate the escape speed from a planet of radius 2.2×10^{3} km if the gravitational field strength on its surface is 3.9 N kg^{-1}.

Additional Perspectives

Escape from the atmosphere

In a gas in a container on the Earth's surface we would normally assume that the molecules are distributed randomly and move with an average speed represented by the absolute temperature (see Chapter 3). For example, at 300 K nitrogen molecules have an average speed of about 500 m s^{-1}. However, if the container is really large we would also need to consider the effect of gravity, which acts on gas molecules in the same way as it acts on everything else.

As the molecules move upwards they slow down and, because of the range of molecular speeds in all gases, there will be fewer molecules per unit volume at greater heights. This results in a decrease in the macroscopic properties of the gas: pressure, density and temperature. Remember that molecules continually collide with each other, resulting in random changes of individual kinetic energies and speeds.

The Earth's atmosphere is not an isolated system and, although the principles described above are valid, the situation is actually much more complicated because energy is continually being transferred to the atmosphere from the Sun and the Earth. Energy is mainly transferred to the air from the Earth's surface, so this is where we would expect the air to be warmest and is why it gets colder when we go up a mountain. Such temperature changes cause changes in air pressure and density, setting up local and worldwide convection currents.

Figure 9.25 The Moon has no significant atmosphere, so the sky is black

A very small percentage of molecules that reach the upper atmosphere still have a speed greater than the escape speed, so they leave the atmosphere permanently. However, the geological and biological processes that created the Earth's atmosphere are still occurring and the gases are being replenished.

Because of its smaller mass, the escape speed from the Moon is much lower than from the Earth. This means that, over a long period of time, almost all gas molecules that were present have escaped and there are no processes to replace them. As a result, the Moon has no significant atmosphere and that is why the sky seen from the Moon is black – there are no gas molecules to scatter the light (Figure 9.25).

Questions

1 a Show that molecules moving vertically upwards at 500 m s^{-1} would rise to a height of about 12 km if they did not collide with other molecules.
  b Give reasons why many air molecules are found at much greater heights.
2 Find out where the Earth's atmosphere came from, and whether there could be an atmosphere on the Moon if it was larger.
3 Do Mercury, Venus and Mars have atmospheres? Are there any other planets or moons in the solar system which have an atmosphere?

9.3 Electric field, potential and energy

9.3.1 **Define** electric potential and electric potential energy.

The mathematics of gravitational and electric fields is very similar and most of the ideas discussed in this section should already be familiar from the work done on gravitation. One important difference is that there is only one kind of mass and so only *attractive* gravitational forces, but there are two kinds of charge which result in both *attractive* and *repulsive* forces. The other important difference is that the study of gravitational fields usually concerns the behaviour of a relatively small mass in the field of a much greater mass, like a planet, but the study of electric fields often involves charges which are similar in magnitude to each other.

Electric potential energy

In any arrangement of charges, **electric potential energy** is stored because, at some time in the past, work was done in moving the charges to their present position. Work had to be done because of the electric forces between the charges.

The total electric potential energy of a system is defined as the work done when bringing all the charges of the system to their present positions, assuming that they were originally at infinity.

Infinity is chosen to be the place where electric potential energy is zero. As with gravitational fields, the net work done to produce a certain arrangement is always the same, it is independent of the paths taken by the charges.

The simplest possible system in which electrical potential energy is stored is two point charges, q_1 and q_2, separated by a distance r, as shown in Figure 9.26.

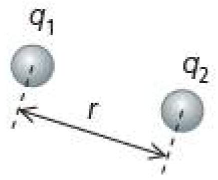


Figure 9.26 Two point charges separated by a distance r

Coulomb's law (Chapter 6) enables us to calculate the force between charges at different separation distances. The work done in bringing the charges to their present positions can then be calculated from the area under a force–distance graph, as shown in Figure 9.27. This graph represents the variation of the repulsive force between two positive charges, q_1 and q_2 (a test charge), when q_2 is moved from a long way away to a distance r_0 from q_1.

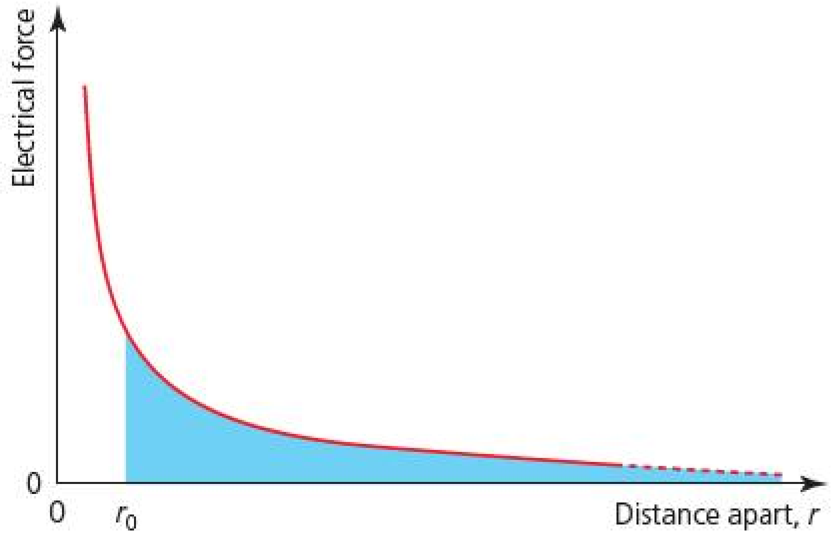


Figure 9.27 The area under a force–distance graph represents the work done in bringing the charges from infinity to a separation of r_0

Alternatively, it can be shown that the electric potential energy can be calculated from the following equation:

$$E_P = \frac{kq_1q_2}{r}$$

where $k = \frac{1}{4\pi\varepsilon_0}$ is the Coulomb constant.

The same equation can be used for charged conducting spheres, which behave as if all of their charge is concentrated at their centres.

If the charges have opposite signs, the mathematics shows that the energy will be negative. This simply means that the forces between opposite charges are attractive and work has to be done to move them apart, increasing the energy so that they have zero energy when they are an infinite distance apart (like two masses in a gravitational field). The energy stored between two like charges is positive. This is because the forces are repulsive and no work has to be done from outside to separate them. As they move apart the potential energy decreases, falling to zero when they are an infinite distance apart.

Electric potential

The **electric potential** at a point is defined as the work done per unit charge in bringing a small *positive* test charge from infinity to that position.

In other words, it is the electric potential energy per unit charge:

$$V = \frac{E_P}{q}$$

Electric potential is given the symbol V and has the units $\mathrm{J\,C^{-1}}$, which is called the **volt, V**; ($1\,\mathrm{J\,C^{-1}} = 1\,\mathrm{V}$).

Potential is a scalar quantity and, although it can be either positive or negative, the sign does not represent a direction. Note also that we had to choose between positive and negative charge in the definition of potential. This affects the sign of the potential, but not the magnitude.

When dealing with gravitation, the choice of a very long way away (infinity) for zero gravitational potential energy seems sensible; when looking at electric fields, large distances are rarely involved and for practical purposes the Earth itself is assumed to be at zero potential. The phrase 'to earth' means to connect to the Earth (0 V).

Electric potential around a point charge

9.3.2 **State** and **apply** the expression for electric potential due to a point charge.

The simplest situation involves the potential around a point charge (or a charged sphere). Consider Figure 9.28, which shows a test charge q_2 brought from infinity to be near the point charge q_1. The electric potential at point P is given by:

$$V = \frac{E_P}{q} = \frac{kq_1q_2/r}{q_2}$$

This has the same value, whatever the route taken by charge q_2.

Replacing q_1 with q, we get the general equation for the variation of potential around a point charge, q:

$$V = \frac{kq}{r}$$

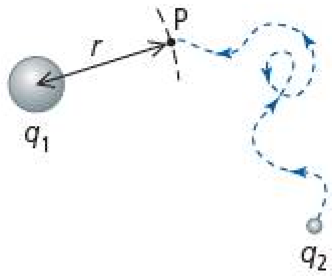


Figure 9.28 Moving a test charge, q_2, by any route from infinity to a distance r from q_1

This equation can also be expressed as:

$$V = \left(\frac{1}{4\pi\varepsilon_0}\right)\frac{q}{r}$$

Both forms of this equation are in the IB *Physics data booklet*.

Figure 9.29 shows how the electric potential varies around positive and negative charges.

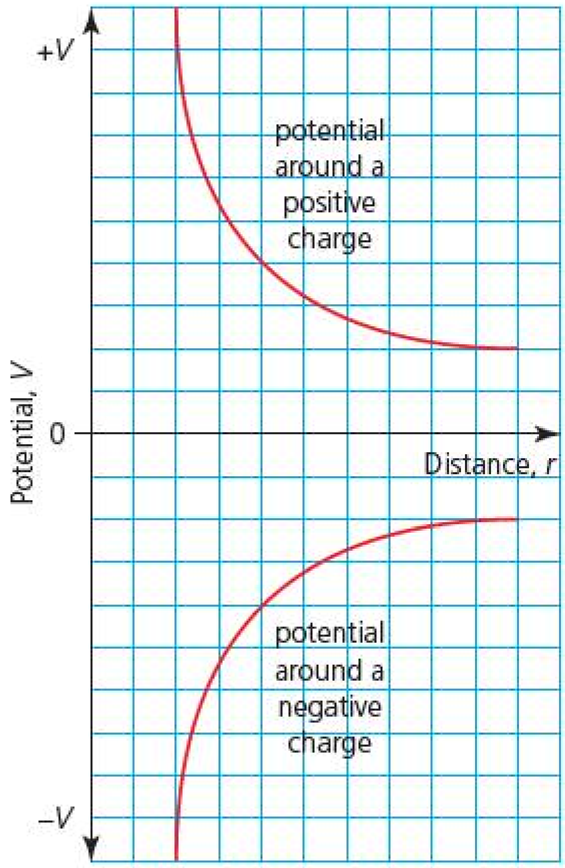


Figure 9.29 A graph to show the variation of potential around a point charge

Worked example

8 Calculate the electric potential 40.0 cm from the centre of an isolated conducting sphere which has a charge of –480 nC.

$$V = \frac{kq}{r}$$

$$V = \frac{(8.99 \times 10^9) \times (-480 \times 10^{-9})}{0.400}$$

$$V = -1.08 \times 10^4\,\text{V}$$

9.3.7 Solve problems involving electric potential energy and electric potential.

31 What is the electric potential at a distance of 12.6 cm from a point charge of 60 μC?

32 What is the electric potential 10^{-10} m away from a positive charge of 1.6×10^{-19} C? (This is the approximate distance of the nucleus from the electron in a hydrogen atom.)

33 A conducting sphere of radius 2.5 cm has a charge of –76 nC. How far from its surface will the potential be –1000 V?

34 What point charge will produce a potential of 5000 V at a distance of 1.0 m?

TOK Link: Fundamental concepts?

The concept of potential is considered by many students to be difficult, and one way of dealing with this is to simply replace the word 'potential' with its meaning: 'potential energy per unit charge (or mass)'. Seen in this way, it is clear that 'potential' is not a fundamental idea in itself, but just a convenient label for a very useful concept – potential energy per unit charge, or mass. Potential is useful in a similar way to which price per unit mass is useful when discussing, for example, the cost of buying or selling goods like rice or gold.

Similarly, speed is just a way of describing distance travelled per unit time. However, most people understand the meaning of the word speed from quite an early age because, unlike potential, there are numerous examples in everyday life.

Scientists have produced a range of interlinking concepts which help to explain and communicate ideas and information, but which of them are absolutely essential? Force and energy are perhaps the key concepts in physics, but even they are defined in terms of the fundamental quantities of mass, length and time.

Questions

1 Was the concept of potential discovered or invented?

2 Explain why understanding the concept of potential is considered difficult. Give another example of a concept that you have found difficult in physics for the same reason.

Electric potential differences and potential gradients

9.3.3 State and **apply** the formula relating electric field strength to electric potential gradient.

Because we are usually concerned with differences in electric potential (potential differences, ΔV) between two places, rather than electric potentials at a point, the equation $V = E_p/q$ can be written in the form:

$$\Delta V = \frac{\Delta E_p}{q}$$

This equation is given in the IB *Physics data booklet*.

The concept of electric potential difference has been met before in Chapter 5, where it was applied to electrical circuits. Now, we are looking for a much wider understanding.

The electric potential difference, ΔV, between two points is the work done when moving unit charge between those two points.

Earlier in this topic we learned that (gravitational) field strength is equal to the negative of the potential gradient, $g = -\Delta V/\Delta r$, and similarly we can write:

$$E = -\frac{\Delta V}{\Delta r}$$

This equation is given in the IB *Physics data booklet*.

This can be expressed as electric field strength is equal to (the negative of) the potential gradient.

Since the electric field inside a hollow conductor is zero, there can be no potential difference inside it, which means that everywhere is at the same potential (and equal to the potential on the surface).

9.3.7 Solve problems involving electric potential energy and electric potential.

35 Figure 9.30 shows how the potential varies around a charged sphere.
a Estimate the charge on the sphere.
b What is the electric field strength 0.5 m from the centre of the sphere?

36 Look again at Figure 9.24, which shows the connections between four gravitational concepts. Draw a similar diagram to represent the same four concepts for electric forces.

37 Draw a graph to show how potential varies around a point charge which produces an electric field of 12000 N C^{-1} at a distance of 40 cm.

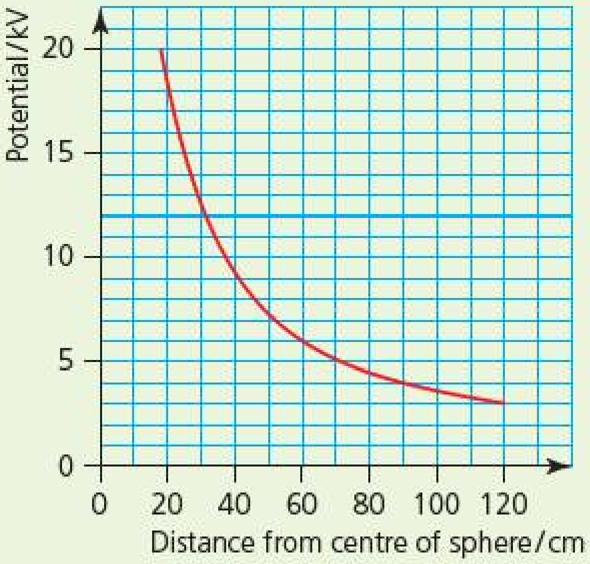


Figure 9.30

Combining potentials

9.3.4 Determine the potential due to one or more point charges.

The overall electric potential at a point can be determined by simply adding individual potentials.

Although electric potential is a scalar quantity, it may be positive or negative depending on the signs of the charges involved. So, it is certainly possible for potentials to combine at a certain place to give zero potential, although the electric field will not be zero. (This is different from gravitational fields, in which the forces are always attractive and the only place to have zero potential is at infinity.)

Worked example

9 Two charges (A and B) of 5.2 μC and −4.4 μC are separated by a distance of 18.0 cm. What is the potential midway between them?

total potential $= V_A + V_B$

$$V = \frac{(8.99 \times 10^9) \times (5.2 \times 10^{-6})}{0.09} + \frac{(8.99 \times 10^9) \times (-4.4 \times 10^{-6})}{0.09}$$

$V = 8.0 \times 10^4\,\text{V}$

9.3.7 Solve problems involving electric potential energy and electric potential.

38 A conducting sphere has a radius of 25 cm. The electric potential on its surface is −1500 V.
a Calculate the charge on the sphere.
b Make a sketch of the 750 V, 500 V and 250 V equipotential lines.
c Add electric field lines to your diagram.

39 What are the potentials at points R, S and T as shown in Figure 9.31? (Assume that the drawing is full size.)

40 Four equal positive charges were located at the corners of a square of sides 10 cm. If the potential at the centre of the square was 1000 V, what was the size of each of the charges?

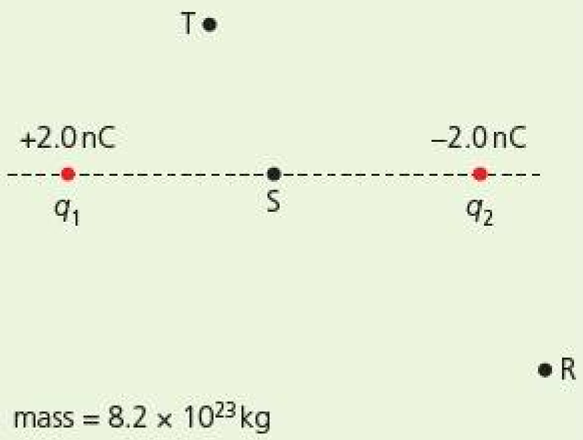


Figure 9.31

Equipotential surfaces and lines

9.3.5 **Describe** and **sketch** the pattern of equipotential surfaces due to one and two point charges.
9.3.6 **State** the relation between equipotential surfaces and electric field lines.

Variations of electric potential around any particular arrangement of charges are best shown on drawings of equipotential lines or surfaces. As with gravitational fields, drawing equipotential lines provides us with an excellent way of visualizing electric fields. The same information is also represented (in another way) by field lines.

Field lines are always perpendicular to equipotential lines and point from higher to lower potential.

Figure 9.32 shows the equipotential lines and field lines around individual point charges and Figure 9.33 represents the uniform field between charged parallel plates. The numerical difference between neighbouring equipotential lines is always kept the same. Where the lines are getting further apart, the potential gradient is decreasing and the field is getting weaker. Field lines and equipotential lines drawn on a flat surface are a simplified way of representing a three dimensional situation and more generally we should be referring to equipotential surfaces.

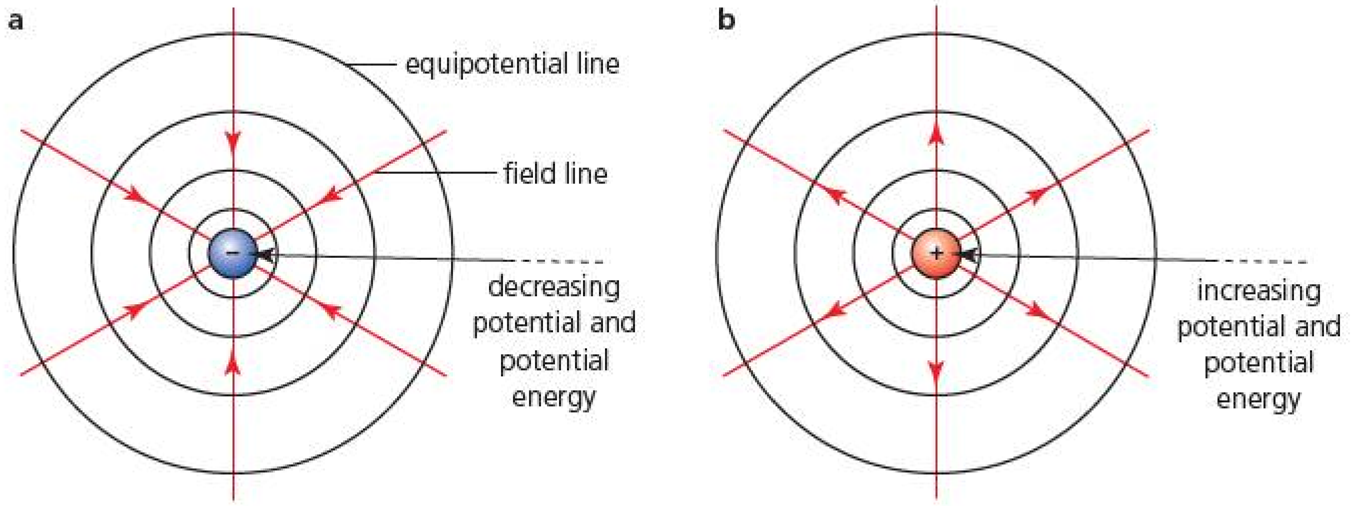


Figure 9.32 Electric field and potential around point charges

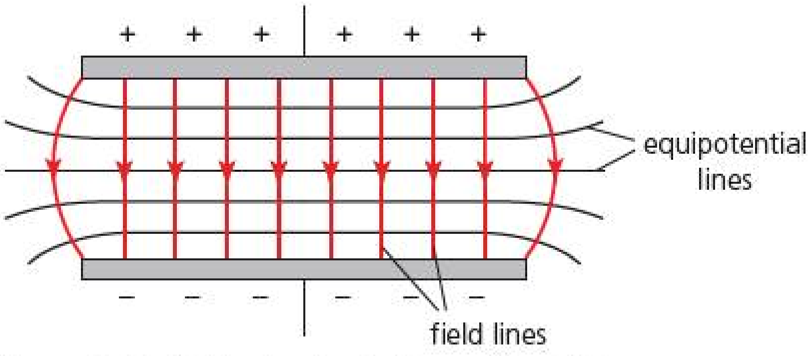


Figure 9.33 Field and potential in a uniform field

By definition, no net energy is transferred when a charge is moved between different places on the same equipotential line.

Figure 9.34 shows the equipotential lines around two equally sized positive charges. Some field lines have also been included.

Figure 9.35 shows the equipotential lines around two equally sized charges of opposite sign. (A pair of equal and opposite charges separated by a small distance is called a **dipole**.)

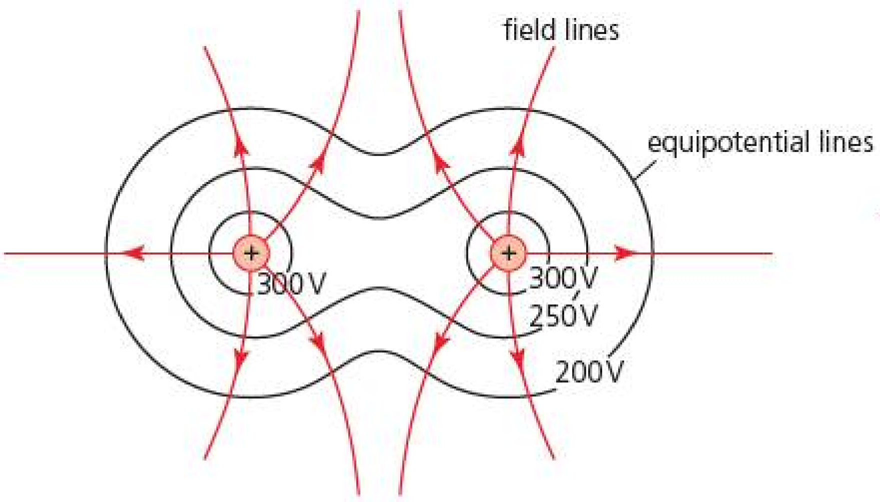


Figure 9.34 Equipotentials around equal point charges

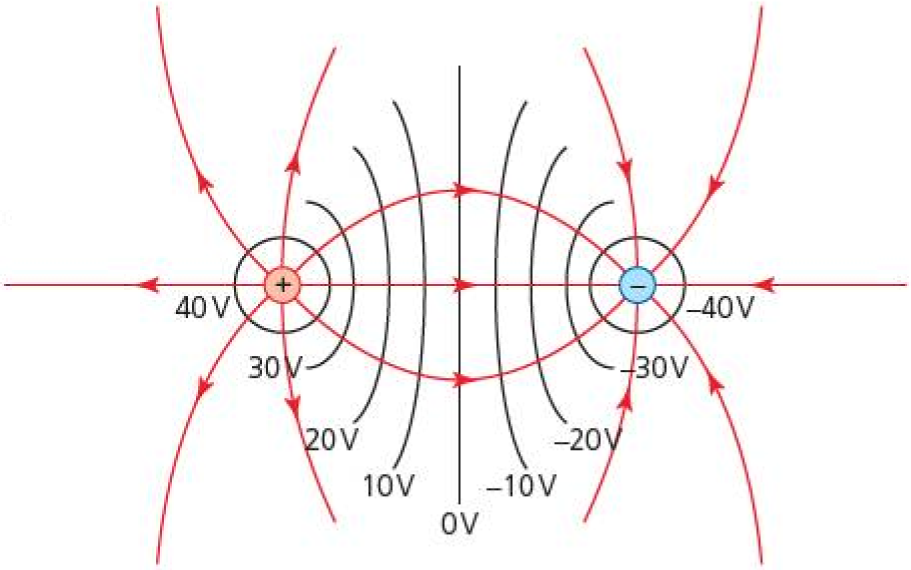


Figure 9.35 Equipotentials around a dipole

41 Sketch the equipotential and field lines around two point charges of different magnitude if
- **a** they have similar signs
- **b** they have opposite signs.

42 Consider the equipotential lines in the uniform electric field between the two parallel plates shown in Figure 9.36.
- **a** Use $E = V/d$ to determine the electric field strength.
- **b** If a charge of $+7.8 \times 10^{-7}$ C was placed at point A, what size force would it experience?
- **c** What is the direction of that force?
- **d** What is the potential difference between points A and B?
- **e** How much work is done if the charge moves from A to B?
- **f** What happens to the charge as it moves from A to B?

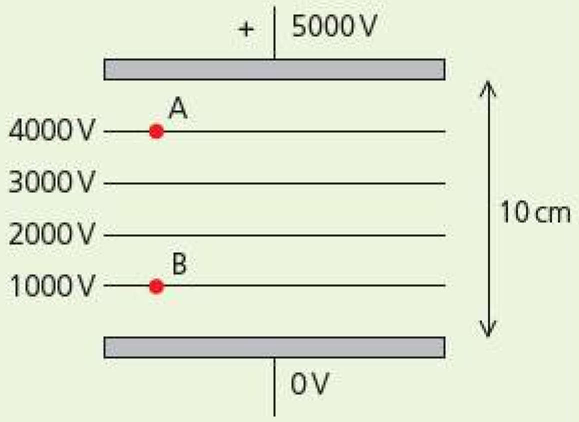


Figure 9.36

9.4 Orbital motion

9.4.1 State that gravitation provides the centripetal force for circular orbital motion.

The gravitational forces between two masses are equal in size, but opposite in direction. However, if one of the masses is very much bigger than the other, we often assume that the force on the larger mass has negligible effect, while the same force acting on the much smaller mass produces a significant acceleration. If the smaller mass is already moving, then the gravitational force can provide the centripetal force to make it **orbit** (move continuously around) the larger mass. It is then described as a **satellite** of the larger mass. The Earth and the other planets orbiting the Sun, and moons orbiting planets, are all examples of **natural satellites**. In the modern world we are becoming more and more dependent on the use of the **artificial satellites** which orbit around the Earth.

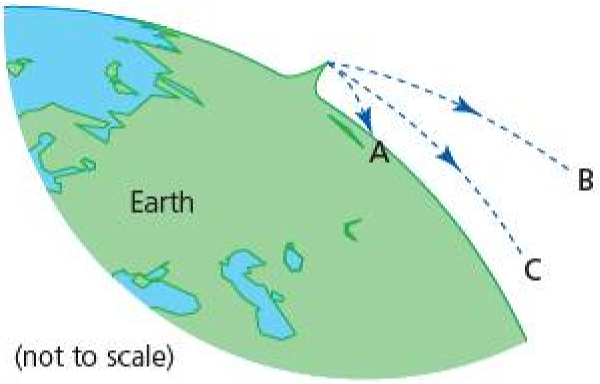


Figure 9.37 The path of objects projected at different speeds from a mountain top

Artificial satellites around the Earth

If we assume that there is no air resistance, a bullet which is fired horizontally from the top of a mountain would move in a parabolic path and hit the ground some distance away, as shown by path A in Figure 9.37. If the bullet had a speed greater than the escape speed, it could move as shown in path B, never coming down. Path C shows the path of an object moving with exactly the right speed and direction so that it remains at the same distance above the Earth's surface (remember that we are assuming that that there is no air resistance); that is, it is in orbit around the Earth.

Gravity is the only force acting on the bullet and it is acting continuously and perpendicular to its instantaneous velocity. This is a necessary condition for circular motion (see Chapter 2). The force of gravity (weight) is providing the centripetal force. Remembering the equation for centripetal acceleration, we can write:

$$\frac{v^2}{r} = g \qquad \left(\text{or } \frac{mv^2}{r} = mg\right)$$

This enables us to calculate the theoretical speed necessary for an orbit very close to the Earth's surface (radius = 6.37×10^6 m):

$$v^2 = gr = 9.81 \times (6.37 \times 10^6)$$
$$v = 7910\,\text{m}\,\text{s}^{-1}$$

Without air resistance, an object moving horizontally with a speed of $7910\,\text{m}\,\text{s}^{-1}$ close to the Earth's surface would orbit the Earth.

To avoid air resistance, a satellite needs to be above the Earth's atmosphere, which means it must be at a height of about 250 km or more. In order to maintain a satellite in a circular orbit around the Earth, it needs to be given the correct speed for its particular height. If there is no air resistance, a satellite in a circular orbit will continue to orbit the Earth without the need for any engine. The force of gravity acts perpendicularly to motion so that no work is done by that force.

Knowing the value of g at any particular height enables us to calculate the speed necessary for a circular orbit at that height. The speed does not depend on the mass. All satellites at the same height move with the same speed.

If the speed of a satellite is greater than the speed necessary for a circular orbit (but less than the escape speed) it will move in an **elliptical** path. The orbits of planets, moons and satellites are not perfectly circular, but the difference is often insignificant.

9.4.6 **Solve** problems involving orbital motion.

43 **a** Calculate values for the gravitational field strengths at heights of 250 km, 1000 km, 10 000 km and 30 000 km.

b Calculate the necessary speeds for circular orbits at these heights.

c Use a compass to draw a scale diagram of the Earth with these orbits around it.

d Use $v = 2\pi r/T$ to determine the times for complete orbits (periods), T, at these heights and mark them on your diagram.

44 Two satellites of equal mass orbit the same planet as shown in Figure 9.38. Satellite B is twice as far away from the centre of the planet as satellite A. Copy the table and complete it to show the properties of the orbit of satellite B.

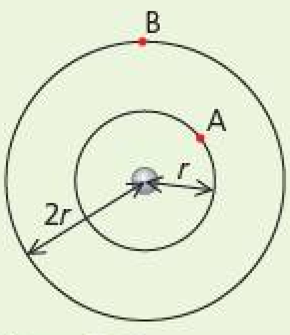


Figure 9.38

	Satellite A	Satellite B
Distance from planet's centre	r	$2r$
Gravitational field strength	g	
Gravitational force	F	
Circumference of orbit	c	
Speed	v	
Time period	T	

Kepler's third law

9.4.2 **Derive** Kepler's third law.

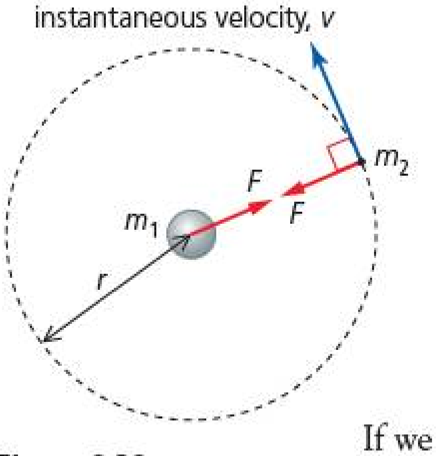


Figure 9.39

The centripetal force (m_2v^2/r) required to keep any mass, m_2, in a circular orbit around a larger mass, m_1, is provided by the force of gravity (Gm_1m_2/r^2), which always acts perpendicularly to the motion of the orbiting mass (see Figure 9.39).

$$\frac{m_2v^2}{r} = \frac{Gm_1m_2}{r^2}$$

$$v^2 = \frac{Gm_1}{r}$$

(This equation is equivalent to $v^2 = gr$ as used on page 343.)

If we replace v with $2\pi r/T$ (circumference/period), we get an important equation which shows us directly how the time period of an orbit depends on its radius. This equation can be applied to elliptical orbits if the average radius is used.

$$\left(\frac{2\pi r}{T}\right)^2 = \frac{Gm_1}{r}$$

Replacing m_1 with m, and rearranging, we get a general equation linking the radius to the period for all satellites orbiting the same mass, m.

$$\frac{r^3}{T^2} = \frac{Gm}{4\pi^2}$$

Figure 9.40 Johannes Kepler

$Gm/4\pi^2$ is a constant for all masses in orbit around the same mass, m. This means that r^3/T^2 must also be a constant. This was first discovered by the German mathematician Johannes Kepler (1571–1630, Figure 9.40). Kepler was using observations and data on the planets of our solar system, but his law can be applied wherever different smaller masses orbit the same larger mass, m. (For example, the moons around Jupiter, or the artificial satellites around Earth). His law is empirical (based only on observation, rather than theory) and the derivation shown above was first put forward by Isaac Newton many years later. Kepler's first two laws are not needed for this course.

Kepler's third law states that for all the planets orbiting the Sun, the average radius cubed is proportional to the period squared (that is, r^3/T^2 is a constant).

Worked examples

10 Io is a moon of Jupiter. It is an average distance of 422 000 km from the centre of Jupiter and takes 1.77 days to complete an orbit. Calculate the mass of Jupiter.

$$\frac{r^3}{T^2} = \frac{Gm}{4\pi^2}$$

$$\frac{(4.22 \times 10^8)^3}{(1.77 \times 24 \times 3600)^2} = \frac{(6.67 \times 10^{-11})m}{4\pi^2}$$

$$m = 1.90 \times 10^{27}\,\text{kg}$$

11a Calculate the orbital radius needed for a satellite which has a period of exactly 24 hours. (Satellites orbiting in the same plane as the equator at this height will remain above the same place on the Earth's surface. They are described as being **geostationary** and they are very useful for global communications, like the transmission of live TV pictures.)

b Use a diagram to help to explain why a geostationary satellite must be orbiting above the equator.

a $\frac{r^3}{T^2} = \frac{Gm}{4\pi^2}$

$$\frac{r^3}{(24 \times 3600)^2} = \frac{(6.67 \times 10^{-11})(5.97 \times 10^{24})}{4\pi^2}$$

$r = 4.22 \times 10^7\,\text{m}$

b See Figure 9.41. In **a**, the satellite orbits in the same plane as the equator and a line joining the satellite to the centre of the Earth would always pass through the same point on the Earth's surface as they both rotate at the same rate. In **b**, the Earth and the satellite rotate at the same rate, but not with the same axis.

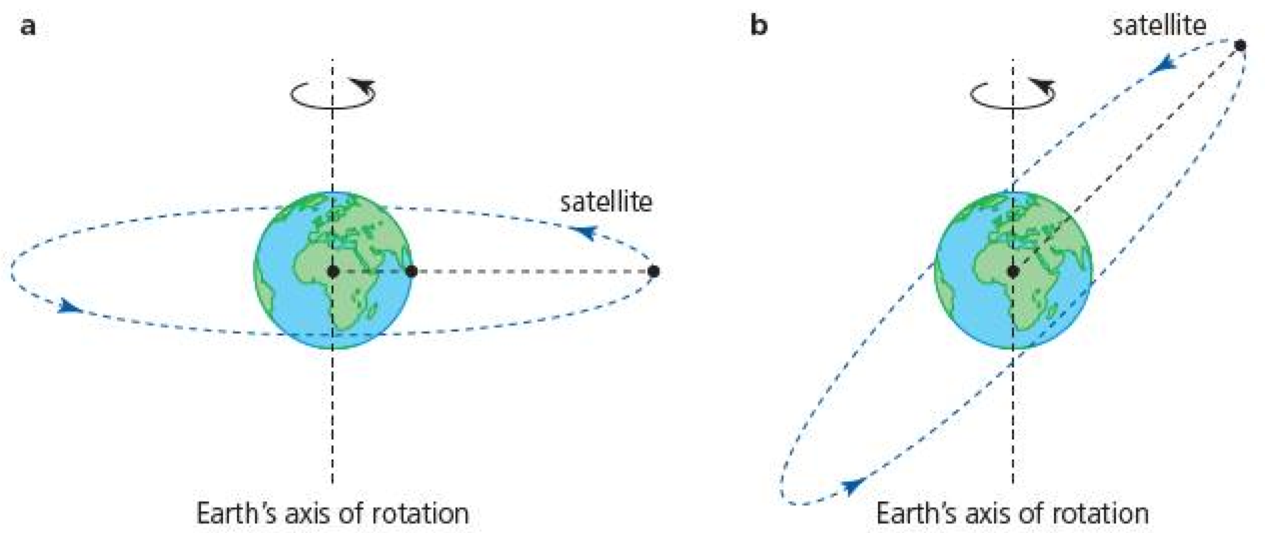


Figure 9.41 Geostationary orbit must be in the same plane as the equator

9.4.6 Solve problems involving orbital motion.

45 **a** Calculate the time period and orbital speed of a satellite in a low orbit 300 km above the Earth's surface.
b Suggest one advantage and one disadvantage of placing satellites at this height.

46 The mass of the Sun is 1.99×10^{30} kg. Use Kepler's third law and the time period of the Earth to determine the average distance to the Sun.

47 **a** Make a spreadsheet of the planets of the solar system, their average distances from the Sun, r, and their periods, T.
b Use the spreadsheet to calculate r^3, T^2 and r^3/T^2.
c Use the program to plot a graph of r^3 against T^2.
d Do your results confirm Kepler's third law? Explain.
e Alternatively, a graph of $\log T$ could be plotted against $\log r$. What could be determined from the gradient of that graph?

48 Two moons, A and B, orbit a planet at distances of 5.4×10^7 m and 8.1×10^7 m from the planet's centre. If moon A has a period of 24 days, calculate the period of moon B.

Energy of an orbiting satellite

9.4.3 **Derive** expressions for the kinetic energy, potential energy and total energy of an orbiting satellite.
9.4.4 **Sketch** graphs showing the variation with orbital radius of the kinetic energy, gravitational potential energy and total energy of a satellite.

We know that when a ball is thrown through the air, it will gain gravitational energy and lose kinetic energy as it moves away from the Earth. The process is reversed when it moves closer to the Earth. If there were no air resistance, the *total* energy of the ball would remain constant. The same principle applies to satellites, moons, planets and comets in their orbits.

To determine the total energy, E_T, of a satellite, we add its kinetic energy, E_K, to its gravitational potential energy, E_P. Remember that the total energy and the gravitational potential energy must both be negative because an orbiting satellite does not have enough energy to escape from the Earth's gravitational field. (The gravitational potential energy is defined to be zero at an infinite distance from Earth.)

$$E_T = E_K + E_P$$

For a satellite of mass m_2, orbiting a planet of mass m_1, with a speed v, at a distance r from its centre, we know that

$$E_P = -\frac{Gm_1m_2}{r}$$

so that

$$E_T = \frac{1}{2}m_2v^2 + \left(-\frac{Gm_1m_2}{r}\right)$$

But we also know that $v^2 = \frac{Gm_1}{r}$, so $E_K = \frac{1}{2}m_2v^2 = \frac{1}{2}m_2\left(\frac{Gm_1}{r}\right)$ which gives

$$E_K = \frac{1}{2}\frac{Gm_1m_2}{r}$$

This leads to $E_T = \frac{1}{2}\frac{Gm_1m_2}{r} + \left(-\frac{Gm_1m_2}{r}\right)$

$$E_T = -\frac{1}{2}\frac{Gm_1m_2}{r}$$

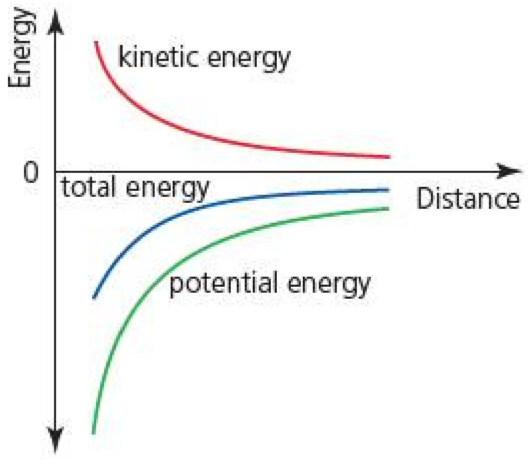


Figure 9.42 Energies of a satellite

The total energy of a satellite in a stable orbit is always equal in magnitude to its kinetic energy, but the kinetic energy is positive, while the total energy is negative. The total energy is equal to half of the gravitational energy. This is shown graphically in Figure 9.42.

If a satellite moving in a circular orbit (at a relatively low height) experiences any air resistance, its speed at that height will decrease and the gravitational force on it will then be greater than the centripetal force needed at that height and speed. If the satellite has no way to control its motion, this will cause a change of height, moving it closer to Earth and therefore it will gain speed and kinetic energy. The total energy and the gravitational potential energy will decrease as energy is dissipated as thermal energy. As the satellite moves closer to Earth, the air resistance will increase further and its downwards trajectory will become steeper, resulting in greater speed and greater energy dissipation. Because the air resistance will be very small at great heights, this process may take a long time, but low height satellites have occasionally 'crashed' to Earth, although most of them 'burn up' in the atmosphere.

9.4.6 Solve problems involving orbital motion.

49 a What happens to the total energy of a satellite if it encounters some air resistance?
b Use the equation for total energy of a satellite to explain what must happen to a satellite which is losing energy.
c Explain why the speed of the satellite will increase and why these effects get greater as the satellite gets lower.

50 a Explain what you think 'burn up' means in the paragraph above.
b Research an occasion when a satellite actually crashed on the Earth's surface and find out what happened.

51 A satellite of mass 820 kg is orbiting at a height of 320 km above the Earth's surface. Calculate
a its gravitational potential energy
b its kinetic energy
c its total energy (Earth's radius = 6.4×10^6 m, Earth's mass = 6.0×10^{24} kg).

■ Additional Perspectives

Global positioning systems (GPS)

There are more than 30 GPS satellites orbiting the Earth at an approximate height of 20000 km. The most used system is maintained by the United States government, but its use is freely available to anyone. Each GPS satellite continually transmits microwave messages of frequency 1.6 GHz (wavelength 19 cm), with the exact time at which each is transmitted. Each satellite orbits the Earth about twice a day.

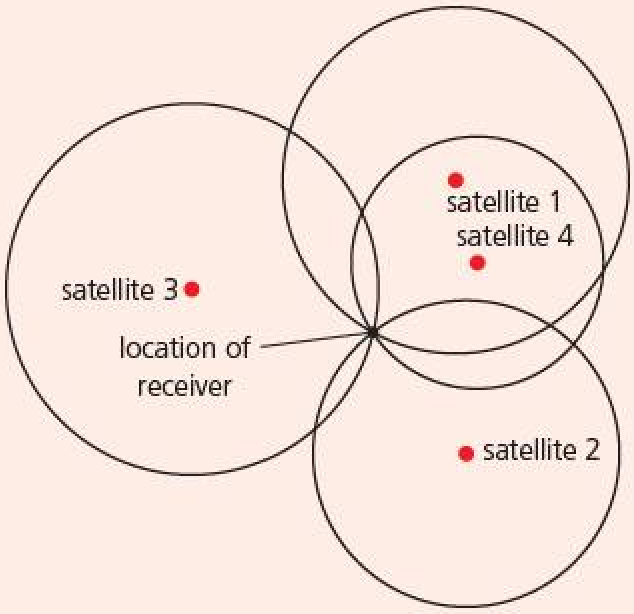


Figure 9.43 Intersecting spheres around GPS satellites

When a GPS receiver detects a message it compares the time it was received to the time it was transmitted, and then calculates the time delay (transit time) and the distance to the satellite (using the speed of electromagnetic waves). Obviously, the receiver must contain a very accurate clock. If similar calculations are made using the signals from another two, or more, satellites, the position of the receiver can be determined. Figure 9.43 shows a two-dimensional representation of the three-dimensional intersection of four spheres. All the satellites are at similar heights above the Earth's surface but their distances from any given receiver are constantly changing.

With the very best (military) equipment it is possible to locate a receiver in terms of latitude, longitude and altitude to within a few centimetres. However, international regulations limit the accuracy of civilian receivers to several metres.

If a moving receiver re-calculates its position after a known short interval of time, it can calculate its speed and direction of motion. All this information is commonly displayed on a two-dimensional electronic map of the area, but GPS can also be used to determine altitudes.

Figure 9.44 A GPS system in a car

Questions

1 Suggest why using the signals from four or more satellites is better than just using three.
2 If a system is to be accurate to within 5 cm, approximately what level of accuracy is needed for the clock in the receiver?
3 Why do you think that governments want to limit the accuracy of the GPS systems that are available to the general public?
4 Why is it important that the signal travels from the satellite to the receiver in a straight line?

Weightlessness

9.4.5 **Discuss** the concept of 'weightlessness' in orbital motion, in free fall and in deep space.

Since weight is the force of gravity on a mass, to be truly weightless a mass would have to be in a place where there was no significant gravitational field, such as in deep space – far, far away from the gravitational effects of any stars or planets.

Alternatively, a mass could appear weightless because it was at a point where the gravitational forces from two or more fields cancelled each other out, such as a point somewhere between the Earth and the Moon.

Figure 9.45 'Weightless' in orbit around the Earth

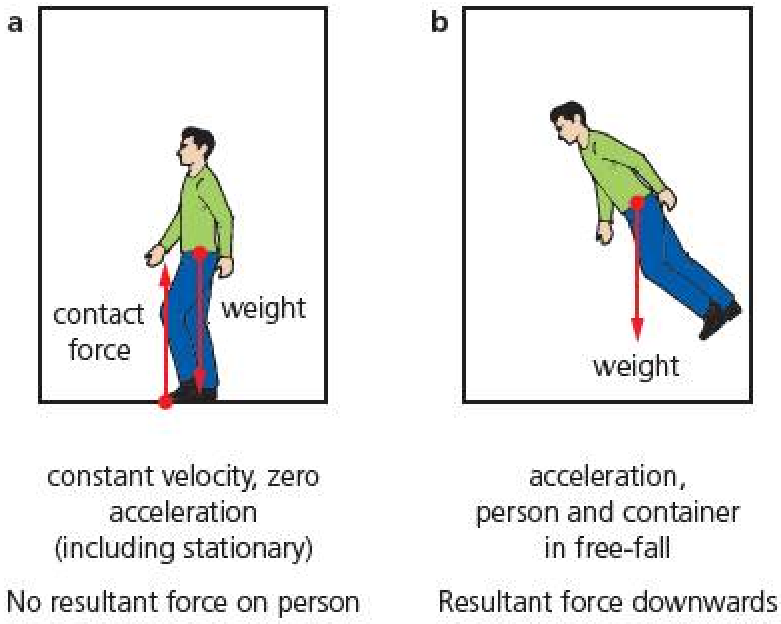


Figure 9.46 Losing contact with the floor

We are all familiar with videos of astronauts moving around freely inside, or outside of, satellites orbiting the Earth (see Figure 9.45). They are often described as being 'weightless', but this cannot be true because the gravitational field strength at the height of a satellite's orbit is not much less than it is on the Earth's surface. We need to understand how and why an object can seem to be weightless when it is in a gravitational field.

We cannot feel or sense our own weight directly, but we are aware of our weight because of the contact force from the surface underneath us pushing upwards. If that contact is removed, we feel weightless, although we may be more concerned with the fact that we are free falling! In an elevator, on a fairground ride, or in a vehicle on a bumpy ride, the contact force between us and the supporting surface changes during any vertical change of speed. As a consequence, we may have a sense of increased or decreased weight; we may even feel brief 'weightlessness' if, for a short time, there is no contact at all between us and the surface.

A satellite is effectively accelerating towards the centre of the Earth (while moving sideways at the same time). Astronauts in the satellite will be accelerating in exactly the same way as the satellite and there will be no need for contact between them and the satellite, so they will feel 'weightless' (see Figure 9.46).

Training for astronauts includes simulating weightlessness aboard flights in aircraft which are following the parabolic trajectory that would be followed by a freely moving projectile (without air resistance).

52 Determine the location of the point between the Earth and the Moon where an object would experience equal and opposite gravitational fields. (Use the Internet to determine the relevant data.)

53 Choose a type of fairground ride and explain the sensations felt by the passengers in terms of forces and accelerations.

SUMMARY OF KNOWLEDGE

9.1 Projectile motion

- The velocity of a projectile can be resolved into vertical and horizontal components ($v_V = v \sin\theta$, $v_H = v \cos\theta$). Because these components are perpendicular to each other, they can be considered independently.
- Projectiles move in parabolic paths if air resistance is negligible.
- The equations of motion and the conservation of energy (gravitational potential energy to or from kinetic energy) can be used with the vertical and horizontal components to predict the exact motion of a freely moving projectile.
- But air resistance is not usually negligible, it reduces speeds, heights and ranges, and the trajectory will not be perfectly parabolic.

9.2 and 9.3 Field, potential and energy

- The mathematics of gravitational and electric fields is very similar. Electric fields may be understood by analogy with gravitational fields, although there are two kinds of charge and only one kind of mass.

- The gravitational (or electric) potential energy of a system, E_p, is defined as the work done when bringing all the masses (or charges) of the system to their present positions, assuming that they were originally (a very long way apart) at infinity. That is, an infinite separation is defined to have zero potential energy.
- The potential energy between two masses (or charges) is inversely proportional to their separation and can be determined from the area under a force–separation graph.
- The gravitational potential energy stored between two masses can be determined from $E_p = -Gm_1m_2/r$. Because gravitational forces are always attractive, all gravitational potential energies (rather than changes of gravitational energy) are negative.
- The electric potential energy between two charges can be determined from $E_p = kq_1q_2/r = (1/4\pi\varepsilon_0)q_1q_2/r$. Because there are both positive and negative charges, electrical potential energy can be either positive or negative.
- Potential, V, is the potential energy per unit mass (E_p/m), or per unit charge (E_p/q). It is defined as the work done per unit mass (or charge) in bringing a small test mass (or charge) from infinity to that position.
- Potential is inversely proportional to separation. The gravitational potential due to a point mass (or a spherical mass) can be calculated from $V = -Gm/r$. The electric potential due to a point charge (or charged sphere) can be calculated from $V = kq/r$.
- Potential is a scalar quantity and the potential at a point due to more than one mass (or charge) can be found by simple addition.
- The potential difference, ΔV, between two points is the work done when moving unit mass (or charge) between those two points. It is independent of the path taken.
- The potential gradient ($\Delta V/\Delta r$) can be determined from a potential–separation graph and it is equal in magnitude to the field strength, g (or E).
- Energies in fields can be represented in drawings by equipotential lines which join places with the same potential. Equipotentials are perpendicular to field lines, which point from higher potential to lower potential.
- For a projectile to 'escape' from a planet it needs to have enough kinetic energy to reach infinity. If the air resistance is negligible, this is equal to the projectile's gravitational potential energy. $v_{esc} = \sqrt{2Gm/r}$

9.4 Orbital motion

- A satellite is the word we use to describe any mass which orbits a much larger mass. Moons are natural satellites of planets, but there are also thousands of artificial satellites around the Earth. The Earth and the other planets are all satellites of the Sun. The force of gravity provides the force which keeps satellites in orbit.
- If the gravitational force is always perpendicular to the motion, the satellite will move in a circular orbit with a constant speed. The speed depends on the distance from the centre of the circle and the strength of the gravitational field: centripetal acceleration $= v^2/r = g$. The speed of a satellite also equals $2\pi r/T$.
- Centripetal force can be calculated from $F = mv^2/r$. If this formula is combined with the equation for the gravitational force ($F = Gm_1m_2/r^2$), an equation linking the period to the radius of a satellite can be derived: $r^3/T^2 = Gm_1/4\pi^2$, which is a constant for all the satellites orbiting the same mass, m_1. This is known as Kepler's third law (when applied to the planets in our solar system).
- The kinetic energy of a satellite, $E_K = \frac{1}{2}m_2v^2 = \frac{1}{2}m_2(Gm_1/r)$. If this is added to its gravitational potential energy ($-Gm_1m_2/r$), an equation for the total energy of the satellite is obtained: $E_T = -\frac{1}{2}(Gm_1m_2/r)$.
- These equations show us that the kinetic energy of a satellite is equal to its total energy, but opposite in sign. Graphs can be used to compare potential, kinetic and total energies.
- A mass can only be truly weightless if it is in 'deep space' a very long distance away from any large masses. But a person can appear to be weightless if they have no surface underneath them to provide a contact force pushing up on them, such as for astronauts in a satellite.

Examination questions – a selection

Paper 1 IB questions and IB style questions

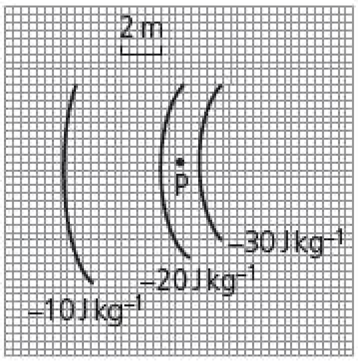
Q1 A satellite is in a circular orbit around the Earth. When it is moved to an orbit that is closer to the Earth's surface, which of the following describes the energy changes that take place?

A gravitational potential energy increases, kinetic energy increases
B gravitational potential energy increases, kinetic energy decreases
C gravitational potential energy decreases, kinetic energy increases
D gravitational potential energy decreases, kinetic energy decreases

Q2 The diagram represents equipotential lines of a gravitational field.

Which of the following is the direction and strength of the field at point P?

	Direction	Strength
A	←	$5.0\,\text{N kg}^{-1}$
B	→	$5.0\,\text{N kg}^{-1}$
C	←	$13\,\text{N kg}^{-1}$
D	→	$13\,\text{N kg}^{-1}$



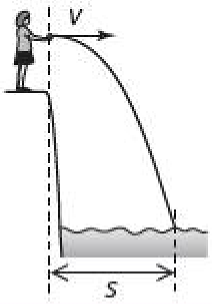
Higher Level Paper 1, May 09 TZ2, Q8

Q3 Which of the following is a correct description of the electric field strength and electric potential around a single point charge?

A Field strength is a vector quantity; potential is a vector quantity.
B Field strength is a vector quantity; potential is a scalar quantity.
C Field strength is a scalar quantity; potential is a vector quantity.
D Field strength is a scalar quantity; potential is a scalar quantity.

Q4 A person standing at the top of a cliff throws a stone horizontally with a speed v. After a time, t, it hits the surface of the sea a distance, s, away.

If a second stone is thrown horizontally with a speed $2v$, which of the following describes the time it takes to reach the sea and the distance from the bottom of the cliff? Assume that air resistance is negligible.



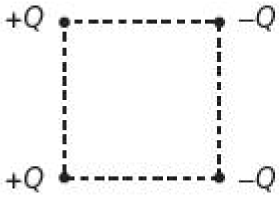
A time t, distance s
B time t, distance $2s$
C time $2t$, distance s
D time $2t$, distance $2s$

Q5 Which of the following is the correct expression for the minimum kinetic energy needed by a mass m in order to escape from the gravitational attraction of a planet of mass M and radius R?

A $\dfrac{GM}{R}$ **B** $\dfrac{GM}{2R}$

C $\dfrac{GMm}{R}$ **D** $\dfrac{GM}{R^2}$

Q6 Four charges are arranged at the four corners of a square as shown in the diagram. At which position is the total electric potential zero?



A nowhere
B at the centre of the square
C midway between the two negative charges
D midway between the two positive charges

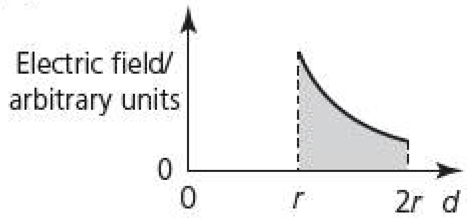
Q7 The gravitational potential on the surface of a planet is V. What is the potential on the surface of another planet of the same average density, but with half the radius?

A $\dfrac{V}{4}$ **B** $\dfrac{V}{2}$

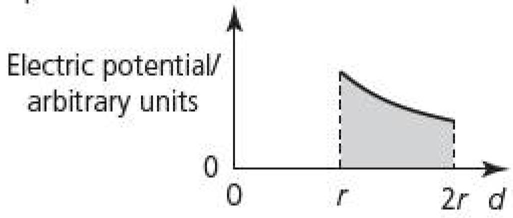
C $2V$ **D** $4V$

Q8 The two graphs below represent the variation with distance, d, for $d = r$ to $d = 2r$, of the electric field and the electric potential around an isolated point charge.

Graph 1:



Graph 2:



The work done by an external force in moving a test charge $+q$ from $d = 2r$ to $d = r$ is equal to q multiplied by the:

A shaded area under graph 1.
B shaded area under graph 2.
C average value of the electric field.
D average value of the electric potential.

Higher Level Paper 1, May 09 TZ1, Q8

Q9 Two satellites, P and Q, have been placed in circular orbits around the Earth. The time period of satellite P is 2 hours and the period of satellite Q is 16 hours. Which of the following is the correct ratio of radius of orbit of satellite Q to radius of orbit of satellite P?

A 64
B 16
C 4
D 2

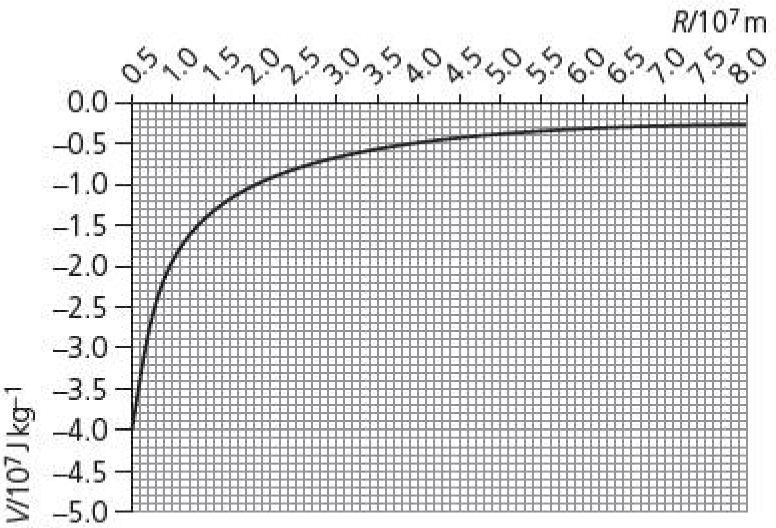
Q10 A stone is thrown up into the air at an angle to the horizontal. Assuming that air resistance is negligible, which of the following is *not* constant while the stone is moving through the air?

A horizontal component of velocity
B vertical component of velocity
C total energy of the stone
D acceleration of the stone

Paper 2 IB questions and IB style questions

Q1 a Define *gravitational potential* at a point in a gravitational field. [3]

b The graph below shows the variation with distance R from the centre of a planet of gravitational potential V. The radius R_0 of the planet = 5.0×10^6 m. Values of V are not shown for $R < R_0$.

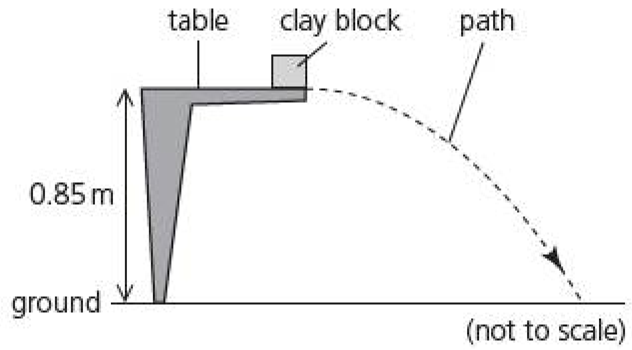


Use the graph to determine the magnitude of the gravitational field strength at the surface of the planet. [3]

c A satellite of mass 3.2×10^3 kg is launched from the surface of the planet. Use the graph to determine the minimum launch speed that the satellite must have in order to reach a height of 2.0×10^7 m above the surface of the planet. (You may assume that it reaches its maximum speed immediately after launch.) [4]

Higher Level Paper 2, Specimen Paper 09, QB3 (Part 2)

Q2 A clay block initially on the edge of a table is fired away from the table as shown in the diagram. The initial speed of the clay block is $4.3\,\text{m s}^{-1}$ horizontally. The table surface is 0.85 m above the ground.



a Ignoring air resistance, calculate the horizontal distance travelled by the clay block before it strikes the ground. [4]

b The diagram shows the path of the clay block neglecting air resistance. Make a copy of the diagram and show on it the approximate shape of the path that the clay block will take assuming that air resistance acts on the clay block. [3]

Higher Level Paper 2, Nov 10, QB4 (Part 1)

10 Thermal physics

STARTING POINTS

- Temperature is a measure of the average kinetic energy of the molecules in a substance.
- The Kelvin and Celsius temperature scales are related by $T/\text{K} = t/°\text{C} + 273$.
- The internal energy of a substance is the total potential energy and random kinetic energy of the molecules of the substance.
- For objects in contact, the flow of energy is always from hotter to colder. This net flow of energy is called thermal energy (sometimes called heat).
- The kinetic model of an ideal gas assumes that all the molecules are identical and have negligible size, the molecules move around randomly because there are negligible forces between them, and collisions between molecules are elastic.
- The pressure of a gas is the force per unit area exerted as a result of collisions of the molecules of the gas with the surface of the container.
- One mole is the amount of a substance that contains the same number of particles as in exactly 12 g of carbon 12. (This number is called the Avogadro constant, N_A, and equals $6.02 \times 10^{23}\,\text{mol}^{-1}$.) The molar mass is the mass of a substance that contains one mole.
- Efficiency of a machine $= \dfrac{\text{useful power out}}{\text{total power in}}$

Introduction

The invention of devices that could continuously use the thermal energy (heat) transferred from a burning fuel to do useful mechanical work completely changed the world. No longer did people and animals have to do such hard work – they could get engines to do it for them, and much more quickly than they could do the same work themselves.

The idea that burning a fuel to heat water to make steam, which could then be used to make something move, had been understood for a very long time. Using this in a practical way was much more difficult, and it was not until the early 18th century that the first commercial steam engines were produced. It was about 100 years afterwards that George Stephenson built the *Locomotion* for the first public steam railway, opened in Britain in 1825.

Nearly 200 years later, things are very different. We live in a world that is dominated by **heat engines** (devices that get useful mechanical work from heat). We are surrounded by all sorts of different engines – cars, boats, trains, planes, factory engines and power stations producing electricity (see Figure 10.1). All these need a transfer of thermal energy from fuels in order to work. It is difficult to overstate the importance of these devices in modern life, because without them our lives would be very different. Of course, we are now also very much aware of the problems associated with the use of heat engines: limited fossil fuel resources, inefficient devices, pollution and global warming.

This chapter is about studying the process of using thermal energy to do useful mechanical work. This branch of physics is known as **thermodynamics**. Although thermodynamics grew out of a need to understand heat engines, it has much wider applications. The study of thermodynamics leads to a better understanding of key scientific concepts like internal energy, heat, temperature, work, pressure and how they are all connected to each other and to the microscopic behaviour of particles.

Figure 10.1 Using heat engines

■ Additional Perspectives

Engine, motor or machine?

Most people would find it difficult to distinguish between a machine, a motor and an engine, because in everyday life the three terms can sometimes be used for the same thing (e.g. a car). The definitions of words often evolve. In practice, the words being discussed here have overlapping meanings. Scientific words can also get absorbed more generally into the language and phrases like 'search engine' and a 'motor for change' only further blur the meaning of the words.

A 'machine' is a general term meaning anything that can do something useful by transferring energy (whether it is a simple, single device, or a complicated mechanism involving many parts). However, in everyday life the word is most commonly used when referring to large, complex machines that involve movement, that is, doing work. But devices like knives and scissors are also machines in the precise sense of the word, and most physics courses teach important and basic physics ideas using simple machines like ramps, pulleys, gears and levers. The word 'device' normally means a fairly simple machine that has been specially made for a particular purpose, like a bottle opener or a hair drier. Parts of larger machines are often called 'components'.

'Engine' is a more specific term than machine. It means a type of machine designed to make something move, but the use of the word is almost always limited to those machines that use the thermal energy (heat) obtained by burning a fuel to do work, for example, an internal combustion engine in a car (see Figure 10.2). To make that use clear, they are often called *heat* engines.

Figure 10.2 Internal combustion engine

Figure 10.3 Rocket being launched

A 'jet engine' is a particular kind of heat engine in which the forward force (thrust) is produced by ejecting a gas, or liquid, in the opposite direction with high speed. The law of conservation of momentum or Newton's third law of motion are usually used to explain jet propulsion (see Chapter 2).

Most heat engines use the combustion of a fuel when it is combined in a chemical reaction with oxygen from the air outside, but if there is no air available, then an oxidizing agent (oxidant) must be provided from within the device, and it is then called a 'rocket' (Figure 10.3).

A 'motor' is another name for a machine that is designed to produce motion, but its most common specific use is with electric motors. However, the word has become closely linked with transport, such as motor cars and motor buses.

Question

1 The following are all heat engines: jet, rocket, internal combustion (car), diesel, steam. Write some short notes to explain the principal differences between them.

10.1 Thermodynamics

Gas laws

The equation of state for an ideal gas

10.1.1 **State** the equation of state for an ideal gas.

In Chapter 3 we briefly discussed the macroscopic physical properties of (real) gases: pressure, P, volume, V, and temperature, T. The gas laws show how they are interconnected for a fixed amount of gas. If these quantities are known for a gas, we know all its important physical properties, for example, its density. Together they define the **state** of the gas. Real gas behaviour was then compared to, and explained by, the predictions of the microscopic **kinetic model of an ideal gas**.

Further developments of the kinetic theory lead to the prediction that, for a fixed amount of an ideal gas, the ratio PV/T is always a constant. (Remember that the temperature must be measured in kelvin) If the amount of gas in a container is increased, the pressure will increase in proportion (if the temperature and volume are kept the same), so the equation can be adapted to:

$$\frac{PV}{nT} = \text{constant}$$

In this equation n represents the amount of gas in moles, that is, it is a measure of the total number of molecules in the container. The value of the constant in this equation is the same for *all* ideal gases. This is because, at the same temperature, the molecules of all ideal gases have the same average kinetic energy: more massive gas molecules travel slower than lighter molecules. The result is that equal amounts of all ideal gases at the same temperature, exert the same pressure in the same volume. There is nothing in this equation that is used to represent the properties of a particular gas.

The constant in this equation is called the **universal (molar) gas constant**. (This is often simply reduced to 'the gas constant'.) It is given the symbol R and has the value of $8.31\,J\,K^{-1}\,mol^{-1}$. This value is given in the IB *Physics data booklet*.

The equation can be written as follows:

$$PV = nRT$$

This equation is given in the IB *Physics data booklet*.

This equation is known as the **equation of state for an ideal gas**. This important equation defines the **macroscopic** behaviour of an ideal gas. That is, the pressure or volume or temperature can be calculated exactly for a known amount of an ideal gas using the equation of state if the other two variables are known. The **microscopic** meaning of an ideal gas was given in Chapter 3. By arrangement, the equation also defines the meaning of the universal gas constant ($R = PV/nT$).

Differences between ideal gases and real gases

10.1.2 **Describe** the difference between an ideal gas and a real gas.

Of course, *ideal* gases are not the same as *real* gases. However, the differences are usually, but not always, insignificant, so that we often assume that the equation of state for an ideal gas can be used to make calculations for a sample of a real gas. But, real gases will *not* behave like ideal gases if the intermolecular forces or molecular sizes can no longer be considered negligible. This may happen at high pressures and densities, when molecular separations are smaller, or at low temperatures.

When a gas is cooled down the molecules move slower and slower and the pressure reduces (assuming that the mass and volume do not change). As the temperature falls, real gases turn into liquids and then solids, if the temperature becomes so low that the molecules do not have enough kinetic energy to overcome the intermolecular forces. However, an ideal gas will not liquefy.

Absolute zero and the Kelvin scale of temperature

10.1.3 **Describe** the concept of the absolute zero of temperature and the Kelvin scale of temperature.

In theory, the temperature and pressure of an ideal gas could be reduced continuously until (almost) all molecular motion has stopped. This is the lowest possible temperature, called **absolute zero** (previously mentioned in Chapter 3). Its value is –273 °C (more exactly, –273.15 °C) or 0 K (see Figure 10.4).

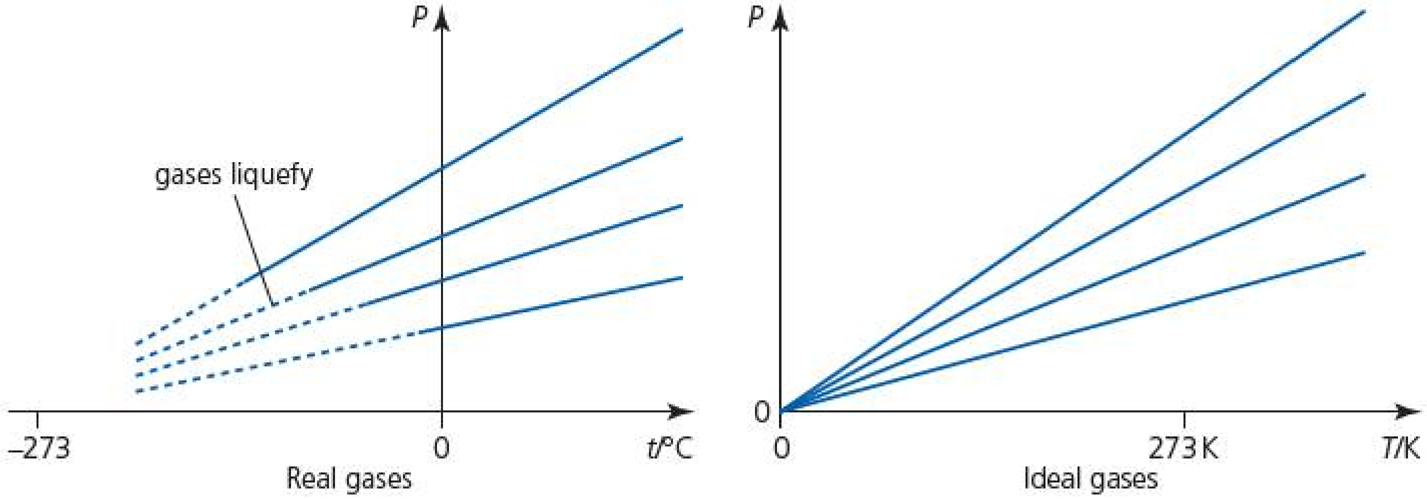


Figure 10.4 Variation of gas pressure with temperature

Figure 10.5 The absolute temperature scale is named after Lord Kelvin

The **Kelvin scale of temperature** is defined by the behaviour of an ideal gas, so that the pressure is proportional to the temperature (in kelvin). It is sometimes called the **absolute temperature scale**.

Worked examples

1 What is the pressure of 23 mol of a gas behaving ideally in a 0.25 m³ container at 310 K?

$$P = \frac{nRT}{V}$$

$$P = \frac{23 \times 8.31 \times 310}{0.25}$$

$$P = 2.4 \times 10^5 \text{ Pa}$$

2 a A fixed mass of an ideal gas has a volume of 23.7 cm³ at 301 K. If its temperature is increased to 365 K at the same pressure, what is its new volume, V_2?

b Explain why the volume has increased.

a $\frac{V}{T} = \frac{nR}{P} = \text{constant}$

$$\frac{23.7}{301} = \frac{V_2}{365}$$

$$V_2 = 28.7 \text{ cm}^3$$

b The volume has increased because the pressure has not changed. The molecules are moving faster and would collide with the walls more often (and with more force), creating greater pressure unless the volume increased.

10.1.4 Solve problems using the equation of state of an ideal gas.

1 What volume of gas (in cm³) contains 0.780 mol of a gas at 264 K at atmospheric pressure (1.00×10^5 Pa)?

2 At what temperature (°C) is 0.46 mol of gas if the pressure exerted in an 8520 cm³ container is 1.4×10^5 Pa?

Figure 10.6 Becoming a scuba diver involves learning about variations of gas pressure under water (see question **3**)

3 Approximately what amount of air would exert a pressure of 2.60×10^7 Pa in a 12 000 cm³ cylinder used by a scuba diver?

4 **a** What is the molar mass of oxygen?
b What volume will be occupied by 1.0 kg at 25 °C and 1.3×10^5 Pa?

5 What pressure will be exerted by putting 2.49 g of helium gas in a 600 cm³ container at −30.5 °C?

6 **a** A container has 4060 cm³ of a gas at twice atmospheric pressure (2.00×10^5 Pa). If the volume is reduced to 3240 cm³ at the same temperature, what is the new pressure?
b Explain why the pressure has increased.

7 **a** Some helium gas in a flask exerts a pressure of 2.12×10^6 Pa at −234 °C. If the temperature is increased by exactly 100 °C, calculate a value for the new pressure (assuming that the flask does not expand).
b Explain why the pressure increases.
c Explain why the use of the equation of state might lead to an inaccurate result in this example.

Figure 10.7 A hot air balloon

8 A gas at 287 K has a volume of 2.4 m³ and a pressure of 1.2×10^5 Pa. When it is compressed the pressure rises to 1.9×10^5 Pa while the temperature rises by 36 K. What is the new volume?

9 Container C has 4 moles of gas. Container D is at the same temperature but has 3 moles of gas in twice the volume. What is the value of the ratio P_C/P_D?

10 The volume of a gas is to be doubled, but the pressure must be kept the same. What must the final temperature (Celsius) be if it was originally at 17 °C?

11 Explain how the height of the hot air balloon shown in Figure 10.7 can be controlled.

■ **Additional Perspectives**

Molar gas constant

Inspection of the units for R ($JK^{-1} mol^{-1}$) suggests that there is a link with specific heat capacity, which has the units $JK^{-1} kg^{-1}$. It suggests that 8.31 J of thermal energy are needed to raise the temperature of one mole of an ideal gas by 1 K. However, unlike solids and liquids, the specific heat capacity of a gas depends on whether the gas expands or is restricted to the same volume (since more energy would be needed if the gas expanded as well as got hotter).

In fact, the molar heat capacity, at constant volume, for a gas that has molecules with two atoms (like nitrogen, N_2) equals approximately $\left(\frac{5}{2} \times 8.31\right)$, or about $20 JK^{-1} mol^{-1}$.

Questions

1 a Calculate the average amount of thermal energy that needs to be given to nitrogen molecules for the temperature of the gas to rise by 1 K.

b Find out what the connection is between the gas constant and the Boltzmann constant (mentioned in Chapter 3).

2 Estimate the amount of thermal energy that needs to be removed from the air in the room where you are located, in order for the temperature to fall by 5 °C.

10.2 Processes

A thermodynamic system can be as complex as a rocket engine, planet Earth or a human body, but in this chapter we will restrict ourselves to the behaviour of ideal gases in heat engines.

As explained in the introduction, heat engines of many kinds play a very important part in all of our lives, so we will concentrate our attention on understanding basic principles by considering those processes that involve the volume increase (**expansion**) of fixed masses of ideal gases.

In the rest of this chapter, and throughout physics, there are many references to (thermodynamic) 'systems and surroundings'. Before going any further we should make sure that these simple and widely used terms are clearly understood.

- A **system** is simply the thing we are studying or talking about. In this chapter it will be a gas.
- The **surroundings** are everything else – the gas container and the rest of the universe.

Sometimes we call the surroundings the **environment**. If we wish to suggest that a part of the surroundings was deliberately designed for thermal energy to flow into it or out of it, we may use the term (thermal) **reservoir**.

Work done when a gas changes state

10.2.1 Deduce an expression for the work involved in a volume change of a gas at constant pressure.

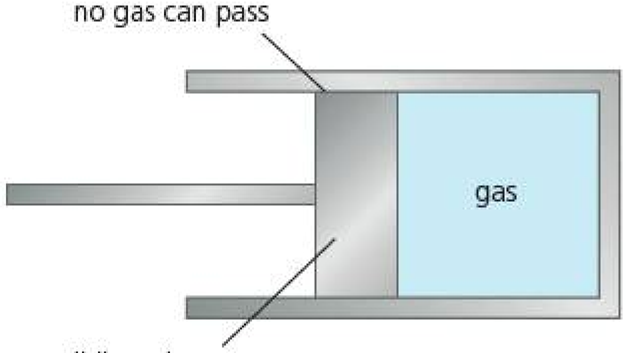


Figure 10.8 Gas in a cylinder with a movable piston

For simplicity, the system that we are considering is often shown as a gas in a regularly shaped cylindrical container, constrained by a gas-tight **piston** that can move without friction (Figure 10.8).

First consider the idealized, simplified example of a gas expanding slightly, so that there is no change in pressure or temperature.

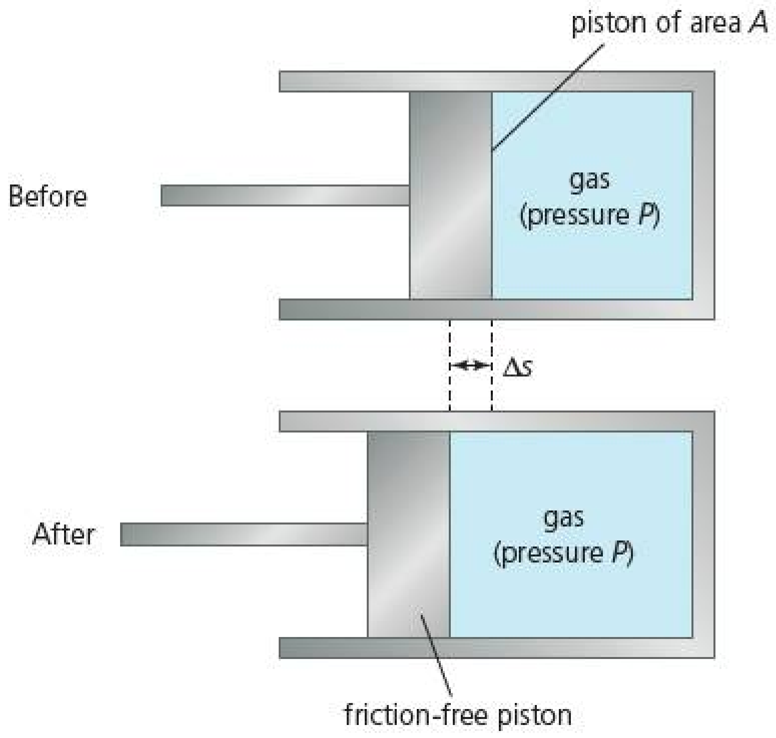


Figure 10.9 Gas expanding in a cylinder

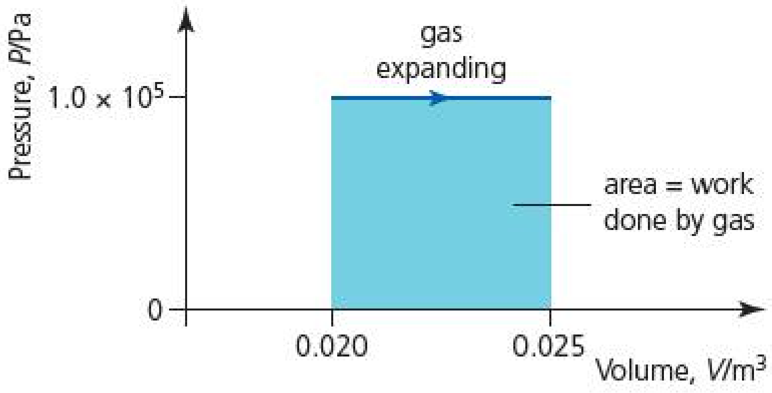


Figure 10.10 Work done during expansion of an ideal gas

If the gas trapped in the cylinder in Figure 10.9 exerts a resultant force on the piston (because the pressure in the cylinder is greater than the pressure in the surroundings), the piston will move outwards as the gas expands. We say that work has been done *by* the gas in pushing back the surrounding gas. Because the displacement, Δs, is small, there will be no significant change in the gas pressure, P. Assuming the temperature is constant, we can write:

work done by gas = force × distance moved in direction of force

or

work done by gas = $(PA)\Delta s$ $\quad\left(\text{since } P = \dfrac{F}{A}\right)$

$$W = P\Delta V \qquad (\text{since } \Delta V = A\Delta s)$$

This equation is given in the IB *Physics data booklet*.
(If the gas is compressed, then work is done *on* the gas (system) and W is given a negative sign.)

This could be represented on a P–V diagram (graph) as in the example shown in Figure 10.10, which shows the expansion of a gas from $0.02\,\text{m}^3$ to $0.05\,\text{m}^3$ at a constant pressure of $1.0 \times 10^5\,\text{Pa}$.

The work done by the gas in expansion in this example, $W = P\Delta V$, is therefore:

$$W = (1.0 \times 10^5) \times (0.050 - 0.025) = 5.0 \times 10^2\,\text{J}$$

Note that this calculation to determine the work done, $P\Delta V$, is numerically equal to calculating the area under the P–V diagram. This is true for *all* thermodynamic processes, regardless of the shape of the graph and this is one reason why P–V diagrams are so widely used in thermodynamics to represent various processes. Figure 10.11 shows two further examples.

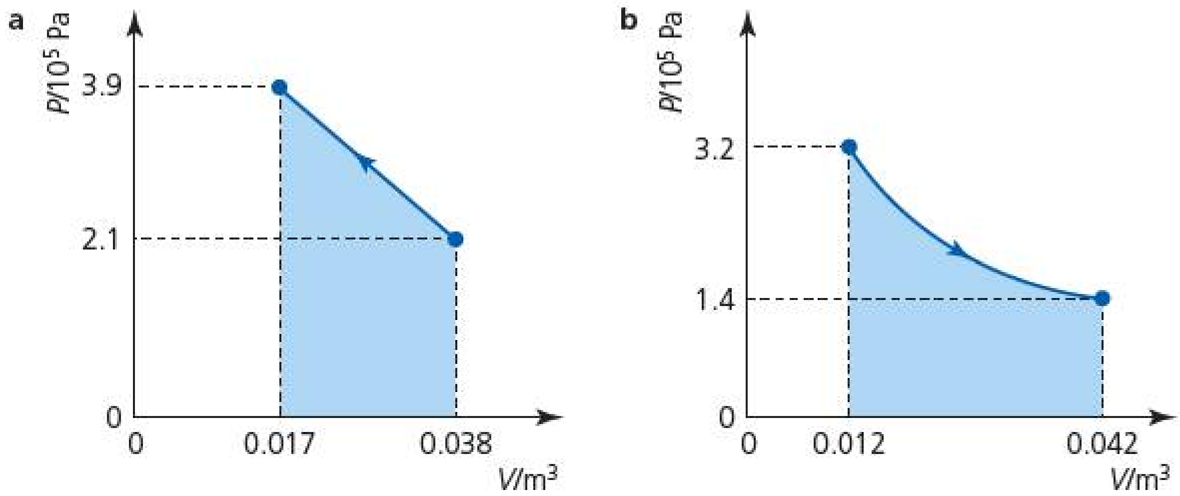


Figure 10.11 Determining areas under pressure–volume diagrams

Worked example

3 Determine values for the work done in the two changes of state represented in Figure 10.11.

a $W = P\Delta V$ = area under graph $= \frac{1}{2}(3.9 - 2.1) \times 10^5 \times (0.038 - 0.017) + (0.038 - 0.017) \times 2.1$ $= 6.3 \times 10^3\,\text{J}$. The work is done *on* the gas as it is compressed into a smaller volume.

b In this example work is done *by* the gas as the volume increases. Because the graph is curved, the area underneath it must be estimated. This could be done by drawing a rectangle which is judged by eye to have the same area as that shaded in the figure:
W = area under graph $= (2.2 \times 10^5) \times (0.042 - 0.012) = 6.6 \times 10^3\,\text{J}$.

First law of thermodynamics

10.2.2 **State** the first law of thermodynamics.
10.2.3 **Identify** the first law of thermodynamics as a statement of the principle of energy conservation.

Expansion, or any other change to the physical state of a gas, involves an energy transfer. Thermodynamic processes may involve doing mechanical work, W (as described on page 358), changing the internal energy of the gas (internal energy is given the symbol, U), or transferring thermal energy, Q. Thermal energy (heat) is the *non-mechanical* transfer of energy from hotter to colder.

We can use the principle of conservation of energy to describe how these quantities are connected: if an amount of thermal energy, $+\Delta Q$, is transferred to the gas/system, then the gas will gain internal energy, $+\Delta U$, and/or the gas will expand and do work on the surroundings, $+\Delta W$. That is:

$$\Delta Q = \Delta U + \Delta W$$

This equation is given in the IB *Physics data booklet*.

This important equation is known as the **first law of thermodynamics**. The first law of thermodynamics is a statement of the principle of conservation of energy as applied to thermodynamic systems.

The directions of energy transfers, as shown by positive or negative signs, can cause confusion. Remember that:

- $-\Delta Q$ refers to a flow of thermal energy *out* of the gas
- $-\Delta U$ refers to a *decrease* in internal energy of the gas
- $-\Delta W$ refers to work done *on* the gas by the surroundings during compression.

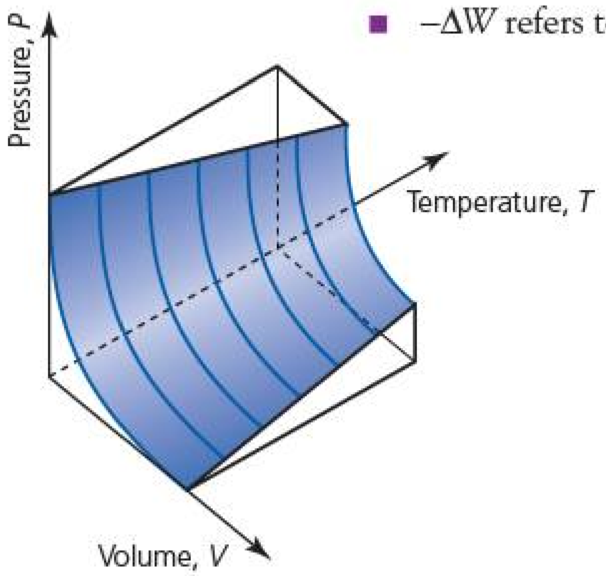


Figure 10.12 Three-dimensional curved surface representing the possible states of a fixed amount of an ideal gas

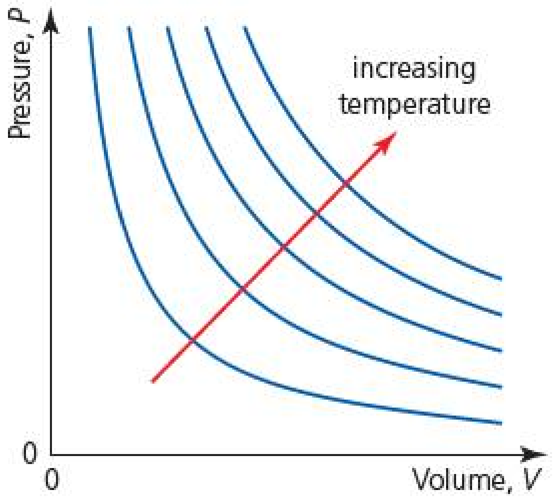


Figure 10.13 Isotherms on a P–V diagram

P–V diagrams

Three quantities – pressure, volume and temperature – are needed to specify the state of a given amount of a gas. We would need a three-dimensional graph to represent all three possibilities, as in Figure 10.12. Instead, we usually use a P–V diagram with the third variable, temperature, shown by one or more curved lines on the diagram, as shown in Figure 10.13.

A line on a P–V diagram that represents possible changes of state of the gas at the same temperature is called an **isothermal line** or **isotherm**. Several isotherms can be shown on the same P–V plot, as in Figure 10.13. Each one of these lines represents the inverse proportionality of a Boyle's law relationship, as previously mentioned in Chapter 3. Note that if a sample of gas is heated within the same volume to a higher temperature, then the pressure will be higher, so that on P–V diagrams 'higher' isothermals correspond to higher temperatures.

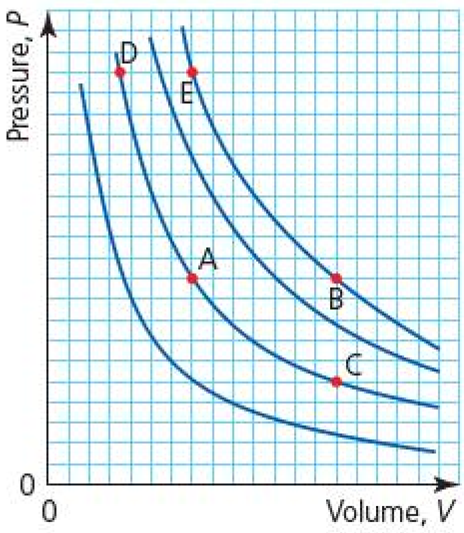


Figure 10.14 Changes of state on a P–V diagram

During a change of state the system will move to a new position on the P–V diagram, but the value of PV/T remains unaltered for a fixed amount of an ideal gas.

Interpreting P–V diagrams is an important skill and Figure 10.14 shows some examples. Consider an ideal gas in a state shown by point A. Doubling the volume at constant pressure is represented by moving to point B, which is at a higher temperature. Point C also represents a doubling of volume, but at constant temperature and lower pressure. Moving to point D would be achieved by doubling the pressure at the same temperature in half the volume. Finally, point E represents doubling the pressure at a higher temperature in the same volume.

Thermodynamic processes for an ideal gas

10.2.4 **Describe** the isochoric (isovolumetric), isobaric, isothermal and adiabatic changes of state of an ideal gas.
10.2.5 **Draw** and **annotate** thermodynamic processes and cycles on P–V diagrams.

Among all the various changes of state that an ideal gas might experience, it is convenient to consider the first law of thermodynamics ($\Delta Q = \Delta U + \Delta W$) under four extremes: $\Delta U = 0$, $\Delta W = 0$, $\Delta Q = 0$, $\Delta P = 0$ (see Figure 10.15).

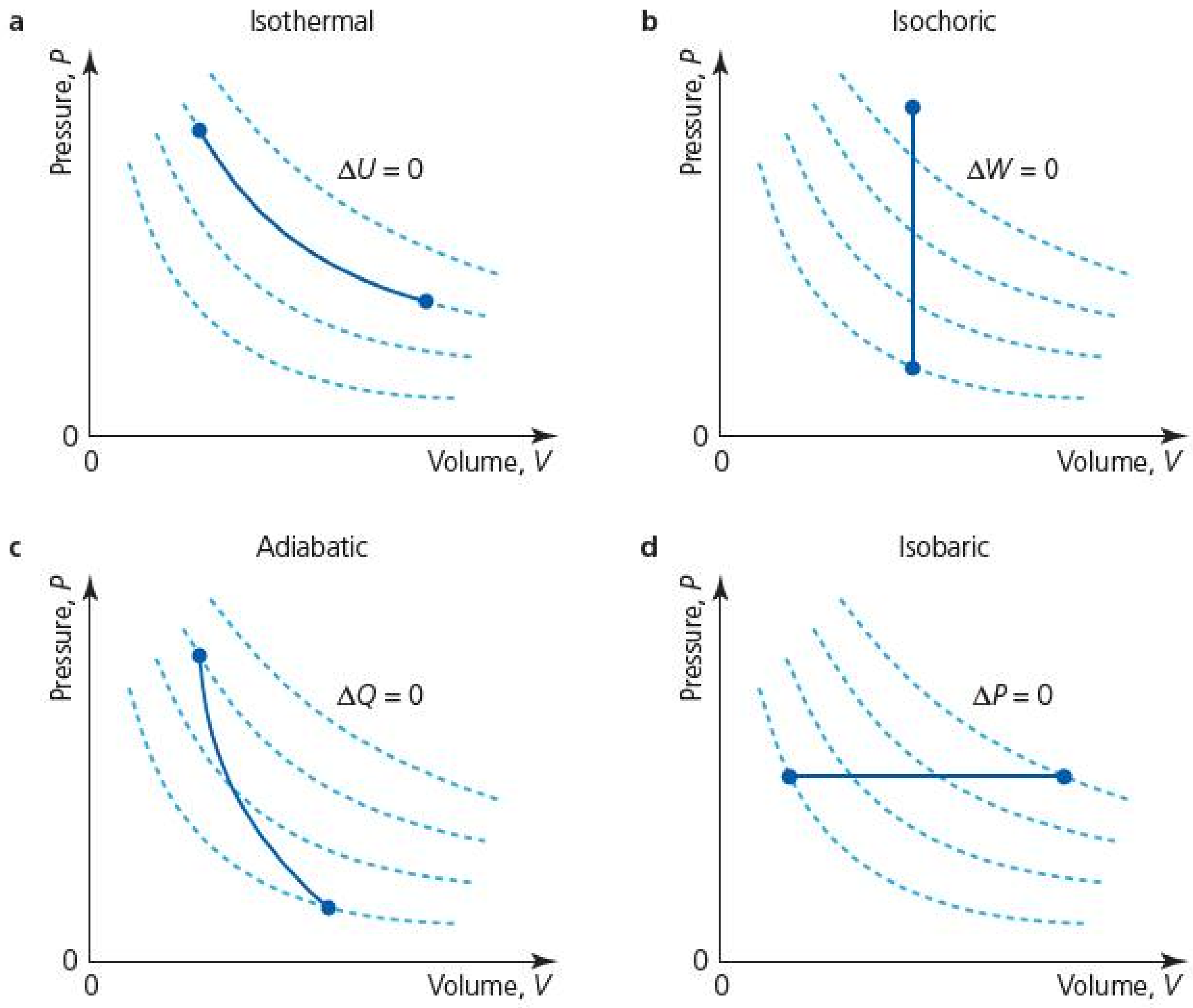


Figure 10.15 Four thermodynamic processes

a **$\Delta U = 0$ (an isothermal process)**. There is no change in the internal energy of the gas. That is, its temperature is constant. Therefore $\Delta Q = 0 + \Delta W$.

In an isothermal expansion all the work done by the gas on the surroundings is supplied by the thermal energy transferred into the gas. In an isothermal compression, the work done on the gas is all transferred away from the gas as thermal energy. For a process to approximate to the ideal of being isothermal, the change must be as slow as possible.

b **$\Delta W = 0$ (an isochoric – or isovolumetric – process)**. There is no work done by or on the gas. This means there is no change in volume. Therefore $\Delta Q = \Delta U + 0$.

In this straightforward process, if thermal energy is transferred into the gas it simply gains internal energy and its temperature rises. If thermal energy is transferred away from the gas its internal energy and temperature decrease.

c **$\Delta Q = 0$ (an adiabatic process)**. No thermal energy is transferred between the gas and its surroundings. Therefore $0 = \Delta U + \Delta W$.

In an adiabatic expansion all the work done by the gas is transferred from the internal energy within the gas, ΔU is negative and the temperature decreases. In an adiabatic compression all the work done on the gas ($-\Delta W$) is transferred to the internal energy of the gas, which gets hotter. When gas molecules hit the inwardly moving piston they gain kinetic energy. Note that adiabatic lines on P–V diagrams must be steeper than isothermal lines simply because, in equal expansions, the temperature falls during an adiabatic change but not (by definition) during an isothermal change.

For a process to approximate to the ideal of being adiabatic, the change must be as rapid as possible in a well-insulated container.

d **$\Delta P = 0$ (an isobaric change)**. Any expansion or compression occurs at constant pressure ($\Delta Q = \Delta U + \Delta W$).

Isobaric changes usually occur when gases are allowed to expand freely keeping their pressure the same as the surrounding pressure.

Worked example

4 An ideal gas of volume $0.08\,\text{m}^3$ and pressure 1.4×10^5 Pa expands to a volume of $0.11\,\text{m}^3$ at constant pressure when 7.4×10^2 J of thermal energy are supplied.

a Name the thermodynamic process.
b Calculate the work done by the gas.
c What is the change in the internal energy of the gas?

a An isobaric process.
b $W = P\Delta V$
$W = (1.4 \times 10^5) \times (0.11 - 0.08) = 4200\,\text{J}$
c $\Delta Q = \Delta U + \Delta W$
$(7.4 \times 10^2) = \Delta U + 4200$
$\Delta U = -3460\,\text{J}$

10.2.7 Solve problems involving state changes of a gas.

12 Explain the differences between the concepts of internal energy, thermal energy and temperature.

13 When a fixed mass of a particular gas expands by a certain amount, compare the amount of work that would be done in isobaric, adiabatic and isothermal changes. Explain your answer.

14 When 100 J of thermal energy were transferred to a gas it expanded and did 30 J of work. What was the change in internal energy?

15 Explain why a gas gets hotter when it is compressed rapidly.

16 When a gas expanded isothermally, 3000 J of work were done.
a Was the work done on the gas or by the gas?
b Was the thermal energy transferred into the gas zero, 3000 J, greater than 3000 J or less than 3000 J?

17 The piston of a bicycle pump is pulled out very slowly.
a What happens to the temperature, internal energy and pressure of the gas?
b What assumption did you make?

18 Explain why an adiabatic expansion shown on a P–V diagram must be steeper than an isothermal expansion starting at the same point.

19 A sample of an ideal gas expands from $2.23\,\text{m}^3$ to $3.47\,\text{m}^3$. The gas pressure was originally 3.63×10^5 Pa.
a If the change was isothermal, what was the final pressure of the gas?
b Draw a suitable P–V diagram and include these two states.
c Draw a line on the diagram to show how the gas went between these two states.
d Estimate the work done during expansion of the gas.

Thermodynamic cycles

An expanding gas (sometimes called a **working substance**) can do useful work, for example by making a piston move along a cylinder. However, it cannot expand forever, so any practical device transferring thermal energy to mechanical work must work in **cycles**, involving repeated expansion followed by compression, followed by expansion, etc. In this section we will discuss some of the physics principles fundamental to cyclical processes in an ideal gas. However, it is important to realize that we are not describing the actual mechanical processes in any particular type of engine.

The essential process in a heat engine is the transfer of thermal energy causing expansion while mechanical work is done by the gas. This is represented by the path AB in Figure 10.16a. The shaded area under the graph represents the work done *by* the gas in this process.

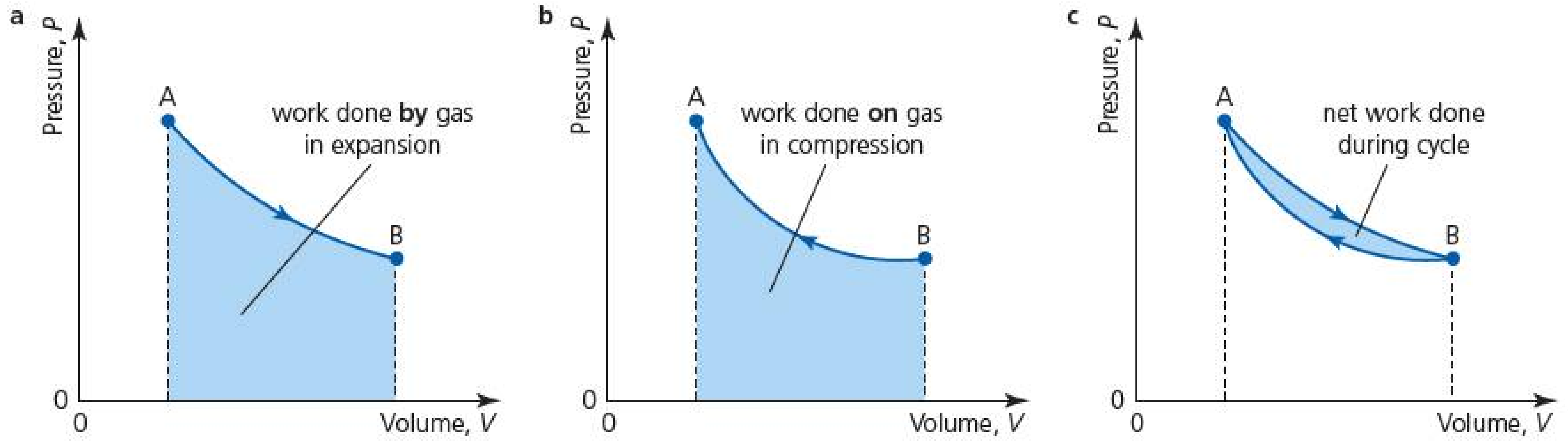


Figure 10.16 Work done in a thermodynamic cycle

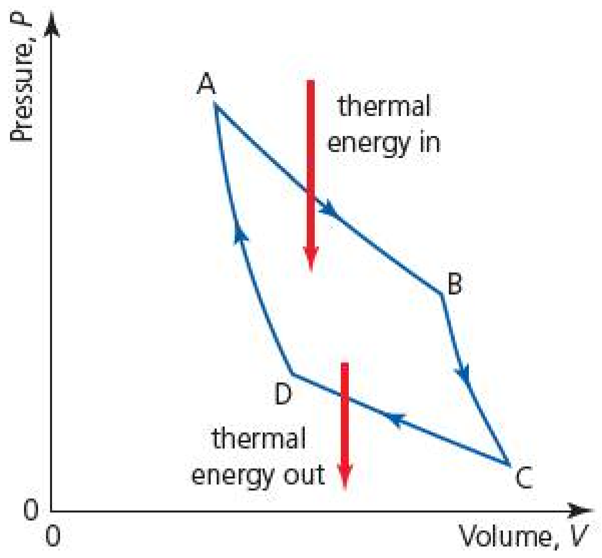


Figure 10.17 The most efficient thermodynamic cycle

In a cyclical process, the gas has to be compressed and returned to its original state. Assume, for the sake of simplicity, that this is represented by the path shown in Figure 10.16b. The area under this graph represents the work done *on* the compressed gas *by* the surroundings. The difference in areas, shown in Figure 10.16c, is the *net useful* work done *by* the gas during one cycle. Of course, if we imagined the impossible situation that, when the gas was compressed, it returned along exactly the same path that it followed during expansion, there would be no useful work done and no energy wasted.

The exact path followed by a heat engine may vary considerably, but the cycle that produces the maximum theoretical efficiency for an ideal gas is shown in Figure 10.17. (It is called the **Carnot cycle**.)

It is a four-stage process: an isothermal expansion (AB) is followed by an adiabatic expansion (BC); the gas then returns to its original state by isothermal (CD) and adiabatic compressions (DA). Thermal energy is transferred during the two isothermal stages.

■ Additional Perspectives

Diesel engines

In car engines (internal combustion) the fuel and air mixture is ignited to cause the rapid, controlled explosions that provide the power to the vehicle. In cars with petrol (gasoline) engines, this is achieved with an electrically produced spark across a 'spark plug' inserted into the gas cylinder. Diesel engines work on a different principle.

Changes to a gas that occur rapidly are approximately adiabatic, so that the work done on the gas when it is compressed causes a rapid increase in temperature. If the compression ratio is high enough (maybe 20:1), the rise in temperature can be large enough to ignite the gas. (The compression ratio is the ratio maximum volume/minimum volume of the working space of the cylinder, as the piston moves up and down.) This is what happens in a diesel engine, a type of engine named after the German, Rudolf Diesel, who produced the first successful engine without spark plugs in 1897.

Because of their higher compression ratios, diesel engines need to be stronger than other engines and they also use a different fuel. They are more efficient than other forms of internal combustion engines and they are widely used where large and powerful engines are needed,

Figure 10.18 Diesel-powered train

especially when the extra mass of a diesel engine is not an important factor, such as in trains (see Figure 10.18) and ships. They have become more and more popular for cars in recent years and about half the cars sold in Europe are now powered by diesel engines.

Question

1 Suppose you were about to buy a new car. Carry out research and a make a list of the advantages and disadvantages of owning a diesel-powered car. Which would you choose?

Calculating the work done in a thermodynamic cycle

10.2.6 Calculate from a P–V diagram the work done in a thermodynamic cycle.

We have seen that the area under a single line on a P–V diagram can be used to calculate the energy transferred when an ideal gas changes state. By considering the energy transfers during each stage of a thermodynamic cycle, we can calculate the *net* energy transfer during a complete cycle. This is shown in the following simplified example.

Worked example

5 Figure 10.19 shows one simplified cycle ABCA of an ideal gas in a particular heat engine, during which time 1.3×10^5 J of thermal energy flowed into the gas.

a Calculate the work done during the process AB.
b Name the processes AB and BC.
c Estimate the net useful work done by the gas during the cycle.
d What is the approximate efficiency of the engine?

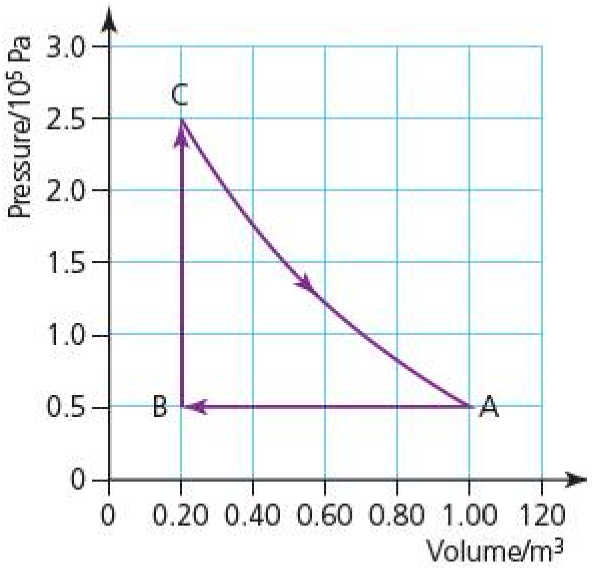


Figure 10.19

a $W = P\Delta V$ = area under AB
$W = (0.50 \times 10^5) \times (1.0 - 0.20)$
$W = 4.0 \times 10^4$ J (done on the gas)

b AB occurs at constant pressure: isobaric compression
BC occurs at constant volume: isochoric temperature increase.

c Net work done by gas = area enclosed in cycle.
$W = \frac{1}{2} \times (0.85 - 0.20) \times (2.5 - 0.5.0) \times 10^5$ (Estimated from a triangle having about the same area, as judged by eye)
$W = 6.5 \times 10^4$ J

d $\text{Efficiency} = \dfrac{\text{useful work output}}{\text{total energy input}}$

$= \dfrac{6.5 \times 10^4}{1.3 \times 10^5} = 0.50$ (or 50%)

10.2.7 Solve problems involving state changes of a gas.

20 Figure 10.17 shows a Carnot cycle. Identify the processes involved and apply the first law of thermodynamics to each of the four processes.

21 Figure 10.20 shows the four-stage cycle of a heat engine.
a Which stage is the compression of the gas?
b The temperature at A is 320 K. Calculate the amount of gas in moles.
c Calculate the temperature at point B.
d Estimate the area ABCD. What does it represent?

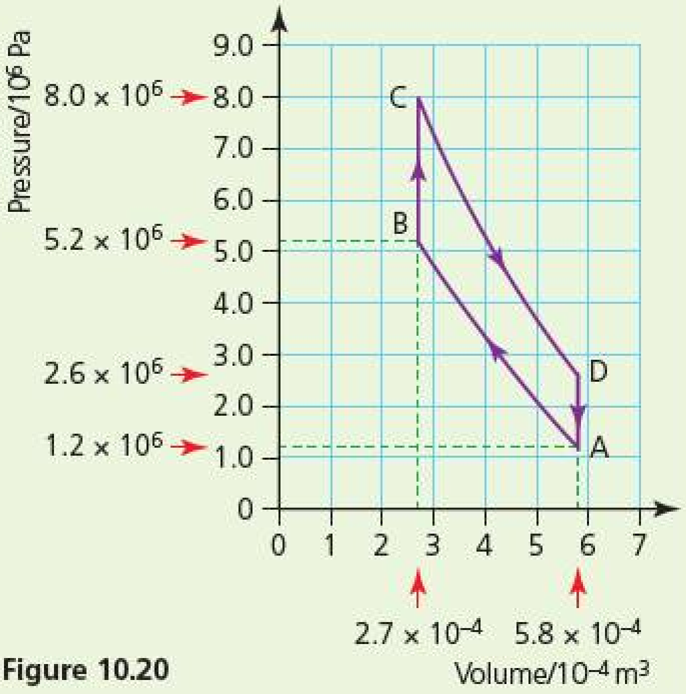


Figure 10.20

22 Using graph paper, make a sketch of the following four consecutive processes in a heat engine. Start your graph at a volume of 20 cm³ and a pressure of 6.0×10^6 Pa.
a An isobaric expansion increasing the volume by a factor of 5.
b An adiabatic expansion doubling the volume to a pressure of 1.5×10^6 Pa.
c An isochoric reduction in pressure to 0.5×10^6 Pa.
d An adiabatic return to its original state.
e Mark on your graph when work is done on the gas.
f Estimate the net work done by the gas during the cycle.

Additional Perspectives

Heat engines and heat pumps

The flow of energy during the continuous conversion of thermal energy to mechanical work in a heat engine is represented in Figure 10.21. A temperature difference, $T_H - T_C$, is needed between hot and cold 'reservoirs' so that there is a resulting flow of thermal energy, which operates the engine. Thermal energy Q_H flows out of the hot reservoir and Q_C flows into the cold reservoir. The difference in thermal energy is transferred to doing useful mechanical work, W.

It is not possible to convert all of the thermal energy into work. (This is a version of the second law of thermodynamics, which is introduced later in this chapter.) For more than 200 years scientists and engineers have spent considerable amounts of time and effort in trying to improve the efficiency of heat engines of various designs. Unfortunately, the laws of physics limit the efficiency that can be achieved. The following simple equation can be used to calculate the maximum efficiency possible (with a Carnot cycle).

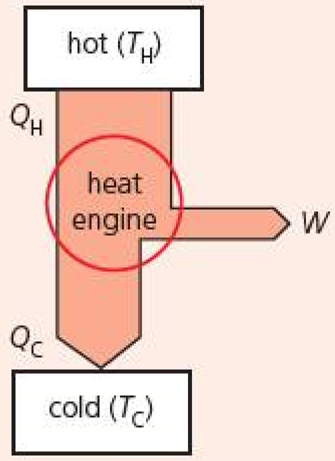


Figure 10.21 Energy flow in a heat engine

$$\text{efficiency} = \frac{(T_H - T_C)}{T_H}$$

Question

1 Calculate the maximum theoretical efficiency of a heat engine operating at an input temperature of 200 °C and releasing heat to the surroundings at 40 °C. How could engineers try to improve that efficiency?

Your answer should be about 34%. In other words, at best, 66% of the energy is not used usefully by the heat engine. Can you think of any way in which this 'wasted energy' could be used?

A heat pump works like a heat engine, but in reverse, using a work input to enable the transfer of thermal energy from colder to hotter, as in a refrigerator or an air conditioner. It is possible to remove thermal energy from the environment and use it to help to keep a house warm in winter, although in many circumstances this is not practicable.

Questions

2 Draw a schematic diagram, similar to Figure 10.21, to represent the principle of operation of a heat pump.

3 Use the Internet to investigate the circumstances under which it may be efficient to use a heat pump to warm a house.

10.3 Second law of thermodynamics and entropy

Order and disorder

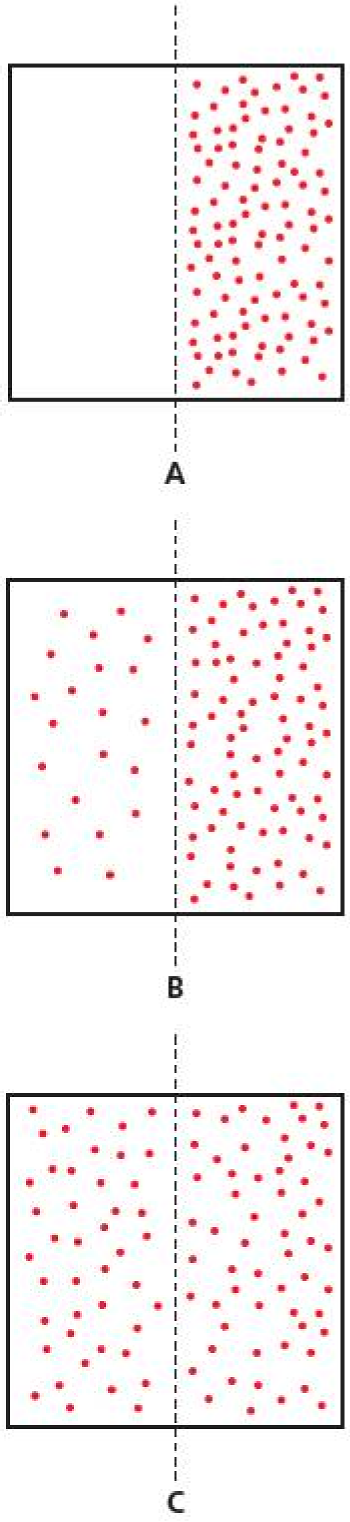


Figure 10.22 Gas molecules spreading out in a container

The molecules of an ideal gas move in a completely random and uncontrolled way. What happens to them is simply the most likely (probable) outcome. It is possible for all the molecules in the room to go out of an open window at the same time. The only reason that this does not happen is simply that it is statistically extremely unlikely.

Consider the three diagrams in Figure 10.22, which show the distribution of the same number of gas molecules in a container. (The dotted line represents an imaginary line dividing the container into two equal halves.) We can be (almost) sure that A occurred before B, and that B occurred before C (maybe the gas was released at first into the right-hand side of the container).

Because it is so unlikely for molecules moving randomly, we simply cannot believe that C occurred before B and A. (In the same way, statistically, we would not believe that if 100 coins were tossed, they could all land 'heads' up.) These diagrams only show about 100 molecules drawn to represent the gas. In even a very small sample of a real gas there will be as many as 10^{19} molecules, turning a highly probable behaviour into a certainty. The simplest way we have of explaining this is that, in the process of going from A to B to C (moving forward in time), the system becomes more **disordered**.

Similarly, the fact that energy is exchanged *randomly* between molecules leads to the conclusion that molecular energies will become more and more disordered and spread out as time goes on. Thermal energy will inevitably spread from places where molecules have higher average kinetic energy (hotter) to places with lower average molecular kinetic energies (colder). This is simply random molecular behaviour producing greater disorder.

Statistical analysis shows that we can be very sure that *every* isolated system of any kind will become more disordered as time progresses. Put simply, this is because everything is made up of particles, and individual atoms and molecules are usually uncontrollable. *Everything* that happens, occurs because of the random behaviour of individual particles. Of course we may wish to control and order molecules, for example by turning water into ice, but this would not be an isolated system: in imposing more order on the water molecules we must remove thermal energy and this will result in an even greater molecular disorder in the surroundings.

Two everyday examples may help our understanding: why is it much more likely that a pack of playing cards will be disordered rather than in any particular arrangement? Why is a desk, or a room, much more likely to be untidy rather than tidy? Because, left to the normal course of events, things get disorganized. To produce order from disorder requires intervention and is difficult, or even impossible. There are a countless number of ways to disorganize a system, but only few ways to organize it.

Entropy

10.3.2 State that entropy is a system property that expresses the degree of disorder in the system.

The disorder of a system can be calculated. It is known as the **entropy** of the system. However, in this course we are not concerned with a precise mathematical definition of entropy, nor how changes of entropy can be measured.

The concept of entropy numerically expresses the degree of disorder in a system.

The concept of entropy is profound and very important. It is relevant everywhere, to every process in every system, to everything that happens anywhere and at any time in the universe.

The principle that molecular disorder is always increasing is neatly summarized by the second law of thermodynamics.

Second law of thermodynamics

10.3.3 State the second law of thermodynamics in terms of entropy changes.

The second law of thermodynamics states that in every process, the total entropy of any isolated system, or the universe as a whole, always increases.

This is sometimes expressed by the statement 'entropy can never decrease'. But it should be stressed that it is certainly possible to reduce the 'local' entropy of *part* of a system, but in the process, another part of the system will gain even more entropy. For example, the growth of a child reduces the entropy of the molecules that come to be inside the growing body, but there will be an even greater increase in the entropy of all the other molecules that were involved in the chemical and biological processes.

The statistical analysis of the behaviour of enormous numbers of uncontrollable particles leads to the inescapable conclusion that differences in the macroscopic properties of any system, like energy, temperature and pressure, must even out over time. This is represented quantitatively by a continuously increasing entropy. Eventually, of course, all energy will be spread out, all differences in temperature will be eliminated and entropy will reach a final steady, maximum value. This is often described as the 'heat death' of the universe.

An alternative way of expressing the second law

10.3.1 State that the second law of thermodynamics implies that thermal energy cannot spontaneously transfer from a region of low temperature to a region of high temperature.

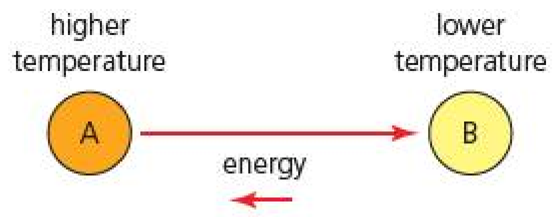


Figure 10.23 Exchanges of energy

Consider any two objects at different temperatures placed in thermal contact in the same system with no external influences, as shown in Figure 10.23.

Energy can flow from A to B and from B to A, but the flow of energy from A to B is more likely because the energy is more 'concentrated' in A. Therefore, the *net* flow of energy will always be from hotter to colder. Obviously, this is a matter of common observation, but it is also the basis of an alternative statement of the second law:

Thermal energy cannot spontaneously transfer from a region of lower temperature to a region of higher temperature.

Thermal energy *always* flows from hotter to colder. Insulation can be used to reduce the rate of energy transfer, but can never stop it completely.

Entropy changes

10.3.4 Discuss examples of natural processes in terms of entropy changes.

Overall energy is always conserved, and overall entropy always increases.

It is not possible to study physics for long without becoming familiar with the fact that in every macroscopic process some useful energy is 'lost' or 'wasted'.

As an example, consider a bouncing ball: the height of each bounce gets lower and lower as shown in Figure 10.24a. If you were to see the opposite happening, with a ball bouncing higher and higher, as in Figure 10.24b, you would probably be amazed, thinking it impossible unless the ball contained an unseen source of energy.

As the ball moves through the air, and as it bounces on the ground, frictional and contact forces result in an increase of internal energy (the ball gets warmer) and then thermal energy is dissipated into the surroundings. We can now interpret this in terms of the second law of thermodynamics and entropy. Figure 10.24b shows an 'impossible' situation because it would involve a decrease in entropy of the system as the energy became more ordered.

Because the molecules in the ball are all moving in the same direction, we can describe the kinetic energy of the ball as *ordered* (disregarding the ball's initial internal energy). As the ball's temperature increases, the increased random kinetic energy of the molecules is *disordered* energy. The overall entropy of the universe increases as the ball loses kinetic energy and gains internal energy.

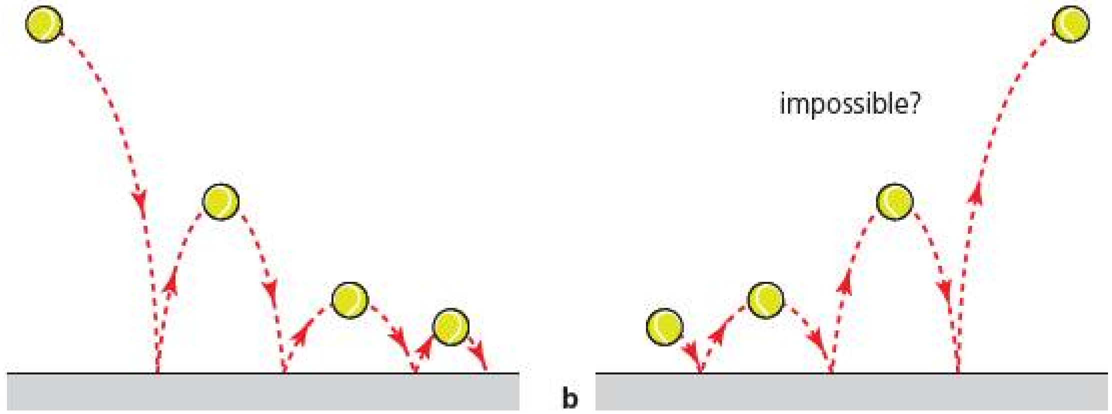


Figure 10.24 Bouncing ball

The same principles apply to all macroscopic processes. We can say that in such processes, some, or all, of the energy is **degraded** because it is been transferred from a useful form to a form in which it can no longer be useful to us, that is, it can no longer perform useful work. The same ideas can be expressed in everyday terms in many different ways, for example, we may talk casually about 'energy being shared out, lost or wasted'. But a higher level of understanding is suggested by reference to thermal energy being **dispersed** or **dissipated** to the surroundings.

The dissipation and degradation of energy occurs because of the chaotic behaviour of molecules. It is simply much more likely that energy will be spread out and become disordered. Given any opportunity, that is exactly what the random behaviour of molecules will *always* produce. This profound and very important idea explains why we are surrounded by 'one-way' processes – events that simply could not happen in reverse. The concepts of **irreversibility**, the 'arrow of time', and why so many things cannot happen even though the principle of conservation of energy is not broken, can all be explained using the powerful ideas of entropy and the second law of thermodynamics.

Figure 10.25 A refrigerator transfers thermal energy from the food and reduces entropy, but where does the energy go?

It may seem that there are some exceptions to the second law. Refrigerators (and air-conditioners) are designed to make something that may already be cooler than the surroundings even cooler by transferring thermal energy from colder to hotter. But these are not 'spontaneous' energy transfers, and in the cooling processes the *overall* entropy of the refrigerator and its surroundings will have increased. The *local* entropy of the contents of the refrigerator will not decrease as much as the entropy of the surroundings will increase (Figure 10.25).

The following quote from Sir Arthur Stanley Eddington (*The Nature of the Physical World*, 1927) may help to convey the importance of this law:

> *The law that entropy always increases holds, I think, the supreme position among the laws of Nature. If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations – then so much the worse for Maxwell's equations. If it is found to be contradicted by observation – well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation.*

23 What happens to the internal energy and the entropy of an ideal gas when it undergoes an isothermal expansion?

24 By discussing what happens to the molecules of the gas, explain the entropy change when a balloon bursts.

25 Coffee, sugar and milk are put in hot water to make a drink. Why it is difficult to reverse the process?

26 Could you make a video recording of a simple event that would not look ridiculous if shown in reverse? Explain.

27 There are four laws of thermodynamics, although only the first and second are included in this book. They are often summarized in the following humorous form.

Zeroth: You must play the game.
First: You can't win.
Second: You can't break even.
Third: You can't quit the game.

What are these comments on the first and second laws suggesting about energy?

TOK Link: Quotes from Richard Feynman

Richard Feynman (1918–1988) received the Nobel Prize for physics in 1965. Although primarily involved in research, he was also a leading personality in the world of physics for many years. He is especially well known for his lectures, books and TV appearances, in which he tried to explain the complexities of physics to the general public. The following quotations give a flavour of his thought-provoking approach to science.

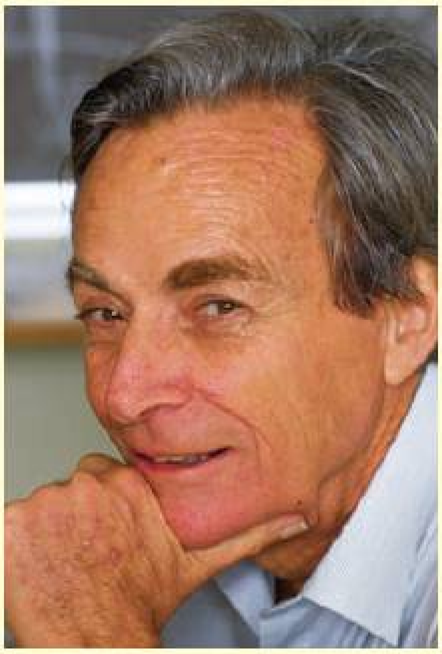


Figure 10.26 Richard Feynman

If, in some cataclysm, all scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic hypothesis (or atomic fact, or whatever you wish to call it) that all things are made of atoms – little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that one sentence you will see an enormous amount of information about the world, if just a little imagination and thinking are applied.

Science is the belief in the ignorance of experts.

You can know the name of a bird in all the languages of the world, but when you're finished, you'll know absolutely nothing whatever about the bird … So let's look at the bird and see what it's doing – that's what counts. I learned very early the difference between knowing the name of something and knowing something.

A great deal more is known than has been proved.

We can't define anything precisely. If we attempt to, we get into that paralysis of thought that comes to philosophers … one saying to the other: 'you don't know what you are talking about!'. The second one says: 'what do you mean by talking? What do you mean by you? What do you mean by know?'

Our imagination is stretched to the utmost, not, as in fiction, to imagine things which are not really there, but just to comprehend those things which are there.

I have approximate answers, and possible beliefs, and different degrees of certainty about different things, but I'm not absolutely sure of anything, … But I don't have to know an answer. I don't feel frightened by not knowing things, by being lost in a mysterious universe without having any purpose, which is the way it really is, as far as I can tell, possibly. It doesn't frighten me.

The Pleasure of Finding Things Out: The Best Short Works of Richard Feynman, edited by Jeffery Robbins

From a long view of the history of mankind – seen from, say, ten thousand years from now, there can be little doubt that the most significant event of the 19th century will be judged as Maxwell's discovery of the laws of electrodynamics. The American Civil War will pale into provincial insignificance in comparison with this important scientific event of the same decade.

'The exception tests the rule.' Or, put another way, 'The exception proves that the rule is wrong.' That is the principle of science. If there is an exception to any rule, and if it can be proved by observation, that rule is wrong.

Question

1 You may detect in some of these quotes a little disrespect for accepted scientific procedures. Do you think that such an attitude is useful for a good scientist? Explain your answer.

SUMMARY OF KNOWLEDGE

10.1 Thermodynamics

- The kinetic model of an ideal gas (introduced in Chapter 3) can be extended to predict properties of gases that we can measure. This theory produces the equation of state for an ideal gas ($PV = nRT$), which defines the macroscopic behaviour of ideal gases. R is known as the universal gas constant.
- Usually real gases behave like ideal gases, except at low temperatures and high pressures, when the assumptions about ideal gases are no longer valid. An ideal gas cannot be turned into a liquid.
- When an ideal gas is cooled, its pressure reduces linearly with decreasing temperature. At −273 °C the pressure of an ideal gas becomes zero because (almost) all molecular motion has stopped. −273 °C is known as absolute zero and it is the basis for the Kelvin scale of temperature.

10.2 Processes

- Heat engines use the expansion of a gas (when supplied with thermal energy) to do useful mechanical work. Heat engines must work in repeating cycles.
- The gas is known as the system, everything else is called the surroundings.
- During a heat engine cycle, each change of state of the gas (system) can usually be approximated to one of four types:
  - Isothermal processes occur at constant temperature, so that there is no change to the internal energy of the gas ($\Delta U = 0$).
  - Isochoric processes occur at constant volume, so that there is no work done ($\Delta W = 0$).
  - Adiabatic processes occur when no thermal energy is transferred to or from the gas ($\Delta Q = 0$).
  - Isobaric changes occur at constant pressure.

  These stages can be respresented by characteristic lines on P–V diagrams.
- Real changes to a gas may be approximately adiabatic if they are done quickly and are well insulated. Isothermal changes need to be done slowly.
- Applying the principle of conservation of energy to thermodynamic changes, we obtain: $\Delta Q = \Delta U + \Delta W$. This is known as the first law of thermodynamics.
- When a fixed mass of gas is compressed or expands, the work done (on or by the gas) can be shown to be determined from $W = P\Delta V$, if the pressure is constant. If the pressure varies, the work done can be determined from the area under a P–V diagram. In a cyclical process, the net work done by the gas is equal to the area enclosed on the P–V diagram.

10.3 Second law of thermodynamics and entropy

- Left to themselves, ordered things naturally become disordered. Because of the random, uncontrollable nature of uncountable molecular motions and energy transfers, everything that ever happens in the universe increases overall molecular and energy disorder.
- Because of this, useful energy always becomes degraded as it gets dissipated into the surroundings. (It cannot be recovered later to do any useful work.)
- Entropy is a numerical measure of the disorder of a system.
- The second law of thermodynamics summarizes these ideas: all processes increase the entropy of an isolated system (and the universe as a whole).
- The second law can also be stated as: thermal energy cannot spontaneously transfer from a region of lower temperature to a region of higher temperature. This is because that would mean there was a decrease in entropy (the molecules would be more ordered).
- It is possible to reduce entropy artificially on a local scale, for example by freezing water, but the energy dissipated from the system will always increase the entropy of the surroundings even more.

Examination questions – a selection

Paper 1 IB questions and IB style questions

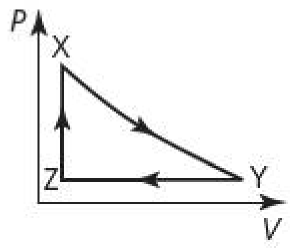
Q1 What entropy changes occur when water is placed in a freezer and turns into ice?

A The entropy of the water decreases and the entropy of the universe decreases.
B The entropy of the water decreases and the entropy of the universe increases.
C The entropy of the water increases and the entropy of the universe increases.
D The entropy of the water increases and the entropy of the universe decreases.

Q2 The graph below shows the variation of the pressure P with volume V of an ideal gas during one cycle of an engine.

Which of the following correctly names the thermodynamic process associated with the parts Y → Z and Z → X of the cycle?

	Y → Z	Z → X
A	isobaric	isochoric
B	isobaric	isothermal
C	isochoric	isobaric
D	isochoric	isothermal



Higher Level Paper 1, Nov 09 TZ2, Q13

Q3 An ideal gas is placed in a container which is divided into two halves of equal volume by a partition. The pressure in both halves of the container is P. If the partition is removed the total pressure will be

A P **B** $\frac{P}{2}$
C $2P$ **D** $\frac{P}{\sqrt{2}}$

Q4 When an ideal gas expanded isothermally 5000 J of work was done. In this process the thermal energy absorbed by the gas was

A 0 J
B less than 5000 J
C 5000 J
D more than 5000 J

Q5 According to the second law of thermodynamics

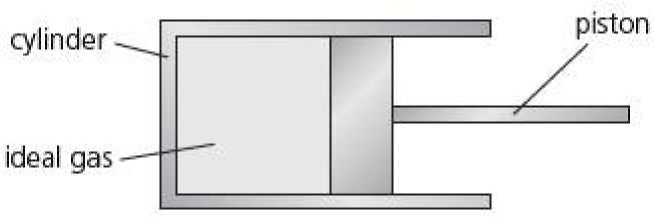
A the entropy of any closed system will always tend to decrease.
B thermal energy flows from hot objects to cold objects.
C it is possible to fully transfer thermal energy to do work in a continuous cyclical process.
D the total entropy of the universe is constant.

Q6 A container of volume V contains n moles of an ideal gas at pressure P and temperature T (K). Another container of volume $2V$ contains $4n$ moles of an ideal gas at temperature $T/2$ (K). What is the pressure in the second container?

A P **B** $\frac{P}{2}$
C $2P$ **D** $4P$

Paper 2 IB question

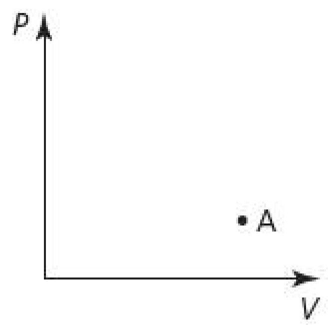
Q1 a An ideal gas is contained in a cylinder by means of a frictionless piston.



At temperature 290 K and pressure 4.8×10^5 Pa, the gas has volume 9.2×10^{-4} m^3.

i Calculate the number of moles of gas. [2]
ii The gas is compressed isothermally to a volume of 2.3×10^{-4} m^3. Determine the pressure P of the gas. [2]
iii The gas is now heated at constant volume to a temperature of 420 K. Show that the pressure of the gas is now 2.8×10^6 Pa. [1]

b The gas in **a iii** is now expanded adiabatically so that its temperature and pressure return to 290 K and 4.8×10^5 Pa respectively. This state is shown as point A.



c i Copy the graph and sketch a pressure–volume (P–V) diagram for the changes in **a ii**, **a iii** and **b**. [3]
ii On your diagram in **c i**, identify with the letter H any change or changes where the gas does external work on its surroundings. [1]
iii Describe how a P–V diagram may be used to estimate a value for the useful work done in one cycle of operation of an engine. [2]

Higher Level Paper 2, May 09 TZ2, QB2 (Part 1)

11 Wave phenomena

STARTING POINTS

- Displacement, amplitude, frequency, wavelength, speed, intensity, phase and phase difference are all important terms used to describe waves.
- The wavelength, λ, frequency, f, and speed, v, of a wave are connected by the equation $v = f\lambda$.
- Travelling waves transfer energy and they can be transverse or longitudinal.
- Electromagnetic waves (like light) are transverse. Sound is a longitudinal wave.
- Two-dimensional waves are represented in diagrams using wavefronts.
- Waves are diffracted when they pass obstacles or go through gaps. Diffraction is greatest when the gap and the wavelength are the same size.
- When waves meet, the result can be determined using the principle of superposition.
- The results of superposition at a point depend on the path difference and phase difference between the waves arriving at that point. When waves arrive in phase constructive interference occurs. If waves arrive out of phase this is described as destructive interference.
- A system can be made to resonate by an external frequency equal to its natural frequency.

11.1 Standing (stationary) waves

Describing standing waves

11.1.1 **Describe** the nature of standing (stationary) waves.

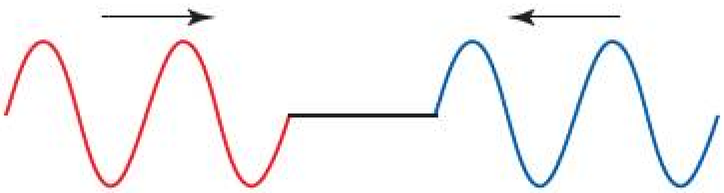


Figure 11.1 Two sinusoidal waves travelling towards each other

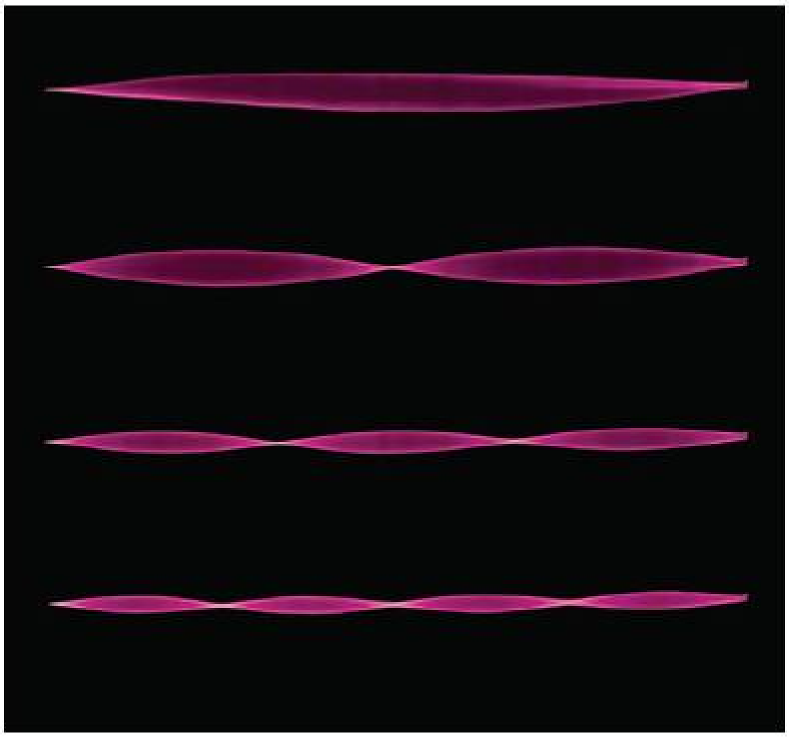


Figure 11.2 Standing waves on a stretched string

Consider two travelling waves of the same shape, frequency, wavelength and amplitude moving in opposite directions, such as shown in Figure 11.1, which could represent transverse waves on a string or rope.

As these waves pass through each other they will combine to produce an oscillating wave pattern that does not change its position. Such patterns are called **standing waves** (sometimes stationary waves) and typical examples are shown in Figure 11.2. Note that a camera produces an image over a short period of time (not an instantaneous image) and that is why the fast-moving string appears blurred. This is equally true when we view such a string with our eyes.

There are points in a standing wave where the displacement is *always* zero. These points are called **nodes**. At positions between the nodes, the oscillations are all **in phase** but the amplitude of the oscillations will vary. Midway between nodes, the amplitude is at its maximum. These positions are called **antinodes**. Figure 11.3 represents the third wave in the photograph in Figure 11.2 diagrammatically. Note that the distance between alternate nodes (or antinodes) is one wavelength.

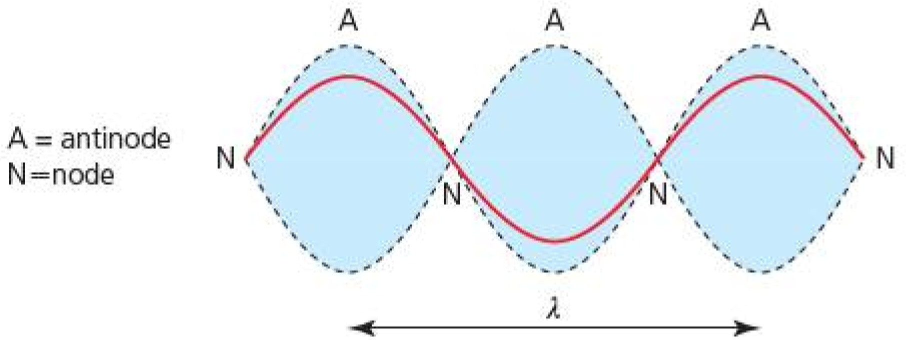


Figure 11.3 Nodes and antinodes in a standing wave on a stretched string. The solid line represents a possible position of the string at one moment

There is energy associated with a standing wave and, without dissipative forces, the oscillations would continue for ever. Energy is *not* transferred by a standing wave.

Explaining standing waves

11.1.2 Explain the formation of one-dimensional standing waves.

When waves similar to those shown in Figure 11.1 meet, **interference** occurs. We can explain the standing wave pattern by determining the resultant at any place and time, and we can do this by using the *principle of superposition* (Chapter 4). The overall displacement is the sum of the two individual displacements at that moment. Nodes occur at places where the two waves are *always* out of phase. At other places, the displacements will oscillate between zero and a maximum value which depends on the phase difference. At the antinodes the two waves are always perfectly in phase. (Students are recommended to use a computer simulation to illustrate this time-changing concept.)

Standing waves are possible with any kind of wave moving in one, two or three dimensions. For simplicity, discussion has been confined to one-dimensional waves, such as transverse waves on a stretched string.

Modes of vibration of transverse waves on strings

Stretched strings

11.1.3 Discuss the modes of vibration of strings and air in open and closed pipes.

Standing waves occur most commonly when waves are repeatedly reflected back from boundaries in a confined space, like waves on a string stretched between two fixed ends. If a stretched string is struck or plucked, it can usually only vibrate if it sets up a standing wave with nodes at both fixed ends. The simplest way in which it can vibrate is shown at the top of Figure 11.4. This is known as the **fundamental mode** of vibration (also called the **first harmonic**). It is usually the most important mode of vibration, but a series of other harmonics is possible and can occur at the same time. Some of these harmonics are also shown in Figure 11.4.

The wavelength, λ_0, of the fundamental mode (first harmonic) is $2l$, where l is the length of the string. The speed of the wave, v, along the string depends on the tension and the mass per unit length. The fundamental frequency, f_0, can be calculated from:

$$f_0 = \frac{v}{\lambda_0}$$

$$= \frac{v}{2l}$$

For a given type of string under constant tension, the fundamental frequency is inversely proportional to the length.

The wavelengths of the harmonics are, starting with the longest, $2l, \frac{2l}{2}, \frac{2l}{3}, \frac{2l}{4}$... and so on.

The corresponding frequencies, starting with the lowest, are f_0, $2f_0$, $3f_0$, $4f_0$ ... and so on.

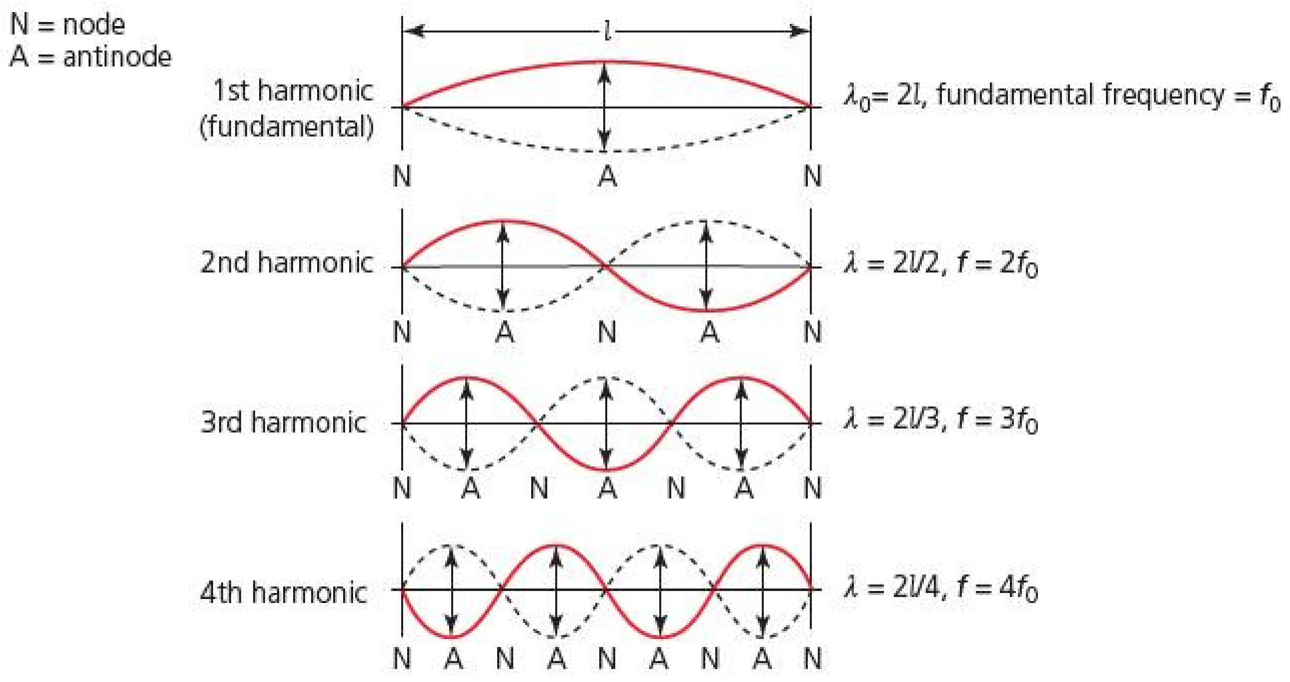


Figure 11.4 Modes of vibration of a stretched string

Additional Perspectives

Stringed musical instruments

When musical notes are played on stringed musical instruments, like guitars, cellos (see Figure 11.5b) and pianos, the strings vibrate mainly in their fundamental modes, but various other harmonics will also be present; this is one reason why each instrument has its own, unique sound. Figure 11.5a shows a range of frequencies that might be obtained from a vibrating guitar string. The factors affecting the fundamental frequency of a note are the length of the string, the tension and the mass per unit length, for example middle C has a frequency of 261.6 Hz. The standing transverse waves of the vibrating strings are used to make the rest of the musical instrument oscillate at the same frequency and, when the vibrating surfaces strike the surrounding air, travelling longitudinal sound waves propagate away from the instrument to our ears.

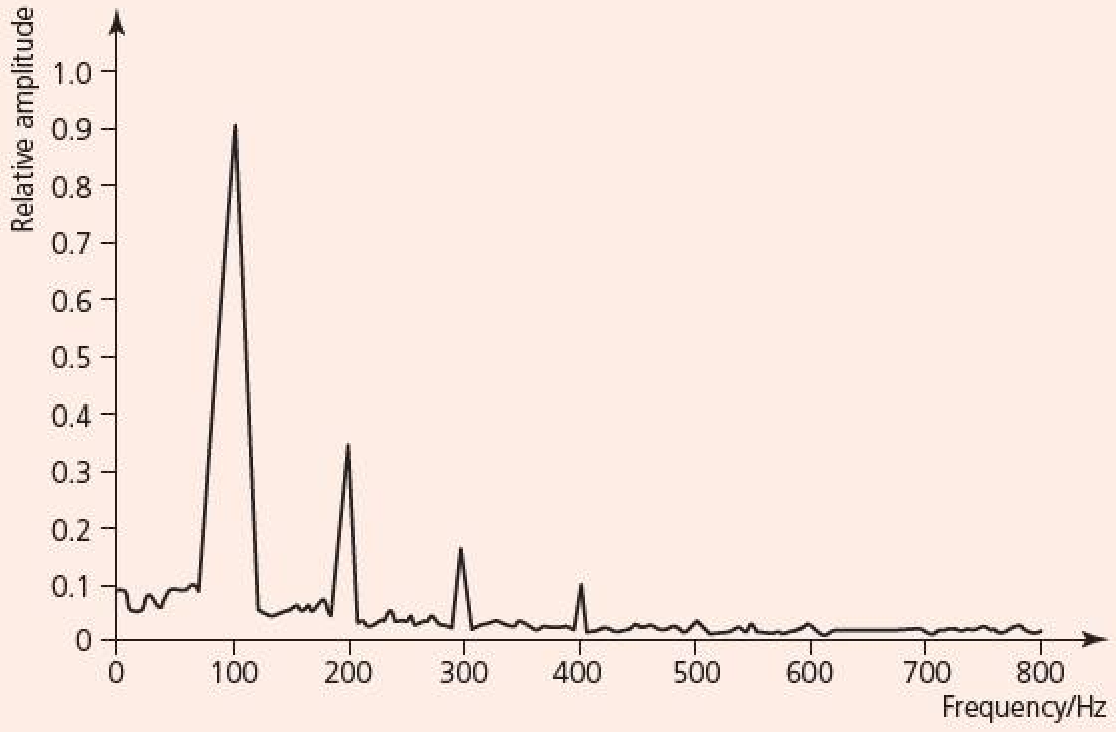


Figure 11.5a Frequency spectrum from a guitar string

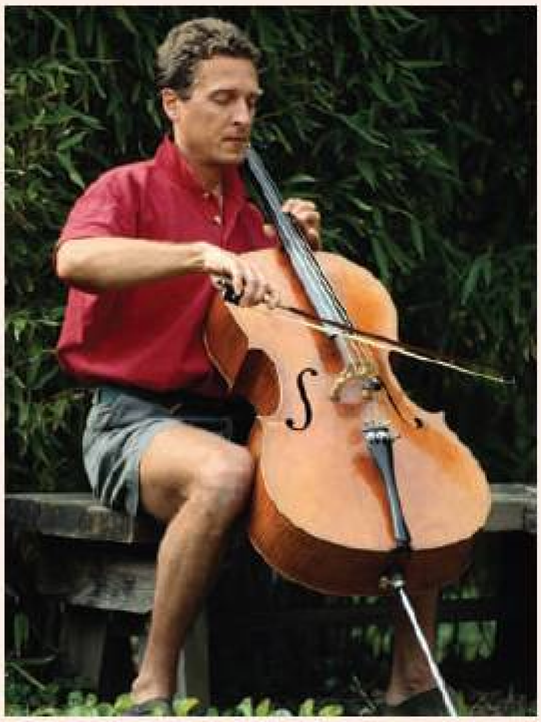


Figure 11.5b Creating standing waves on a cello

Worked example

1 A string has a length of 1.2 m and the speed of transverse waves on it is $8.0\,\mathrm{m\,s^{-1}}$.

a What is the wavelength of the fundamental mode (first harmonic)?
b Draw sketches of the first four harmonics.
c What is the frequency of the third harmonic?

a $\lambda_0 = 2l = 2 \times 1.2 = 2.4\,\mathrm{m}$

b See Figure 11.4.

c $\lambda = \frac{2.4}{3} = 0.8\,\mathrm{m}$

$f = \frac{v}{\lambda} = \frac{8.0}{0.8} = 10\,\mathrm{Hz}$

Longitudinal sound waves in pipes

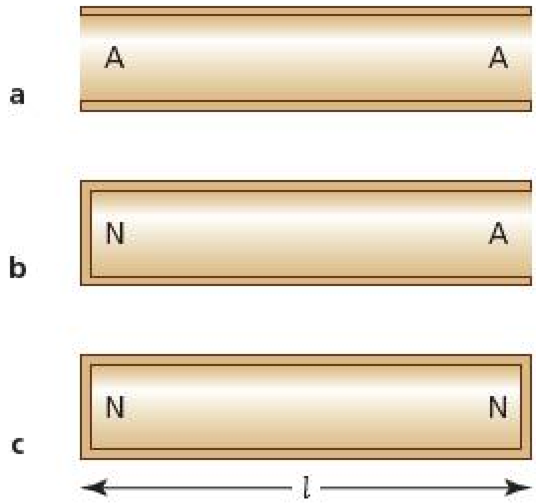


Figure 11.6 Nodes and antinodes at the ends of open and closed pipes

Air can be made to vibrate and produce standing longitudinal sound waves in various containers and tubes. The sound produced by blowing across the top of an empty bottle is an everyday example of this. Many musical instruments, such as a flute or a clarinet, use the same idea. For simplicity, we will only consider standing waves in **pipes** of uniform shape (sometimes called air **columns**).

As with strings, in order to understand what wavelengths and frequencies can be produced, we need to consider the length of the pipe and what happens at the end points (boundaries) of the wave, sometimes called the **boundary conditions**.

This is illustrated in Figure 11.6. In **a** the pipe is open at both ends, so it must have antinodes, A, at the ends, and at least one node in between. In **b** the pipe is open at one end (antinode) and closed at the other end (node, N). In **c** the pipe is closed at both ends, so it must have nodes at the ends, and at least one antinode in between.

Figure 11.7 shows the first three harmonics for a pipe open at both ends. The fundamental wavelength (twice the distance between adjacent nodes or antinodes) is $2l$. The fundamental frequency is therefore inversely proportional to the length. (Note that drawings of standing longitudinal waves can be confusing and the curved lines in the diagram are an indication of the amplitude of vibration. They should not be mistaken for transverse waves.)

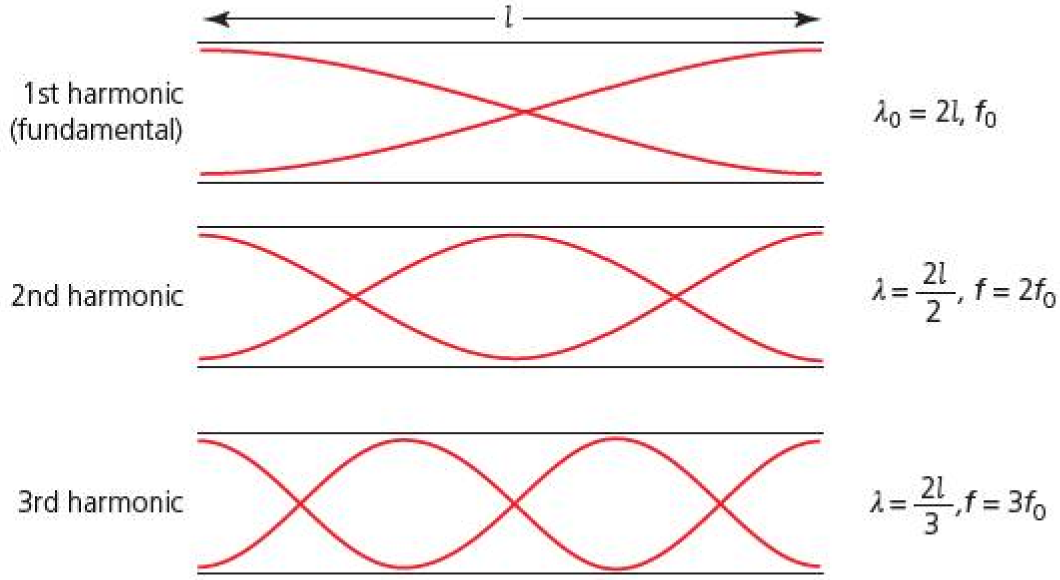


Figure 11.7 The first three harmonics in a pipe open at both ends

A pipe closed at both ends also has a fundamental wavelength of $2l$.

Figure 11.8 shows the first three possible harmonics for a pipe open at one end and closed at the other. Only odd harmonics are possible under these circumstances. The fundamental wavelength is $4l$.

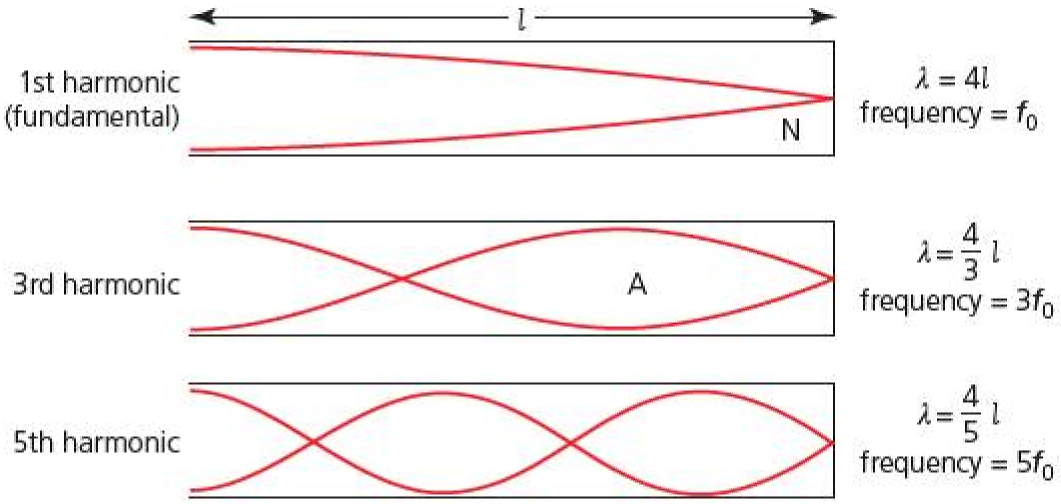


Figure 11.8 Harmonics in a pipe open at one end

Figure 11.9 shows a way of demonstrating standing waves with sound. A speaker is placed close to the open end of a long transparent pipe which is closed at the other end. Some powder is scattered all along the pipe and when the loudspeaker is turned on and the frequency carefully adjusted, the powder is seen to move into separate piles. This is because the powder tends to move from places where the vibrations are large (antinodes) to the nodes, where there are no vibrations.

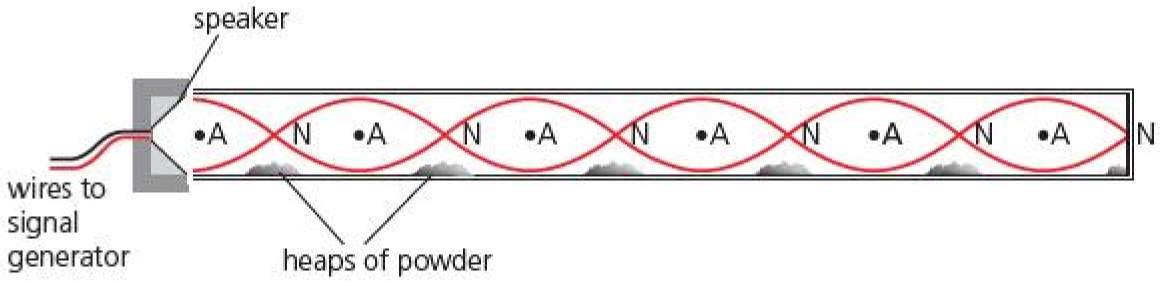


Figure 11.9 Demonstrating a standing wave with sound

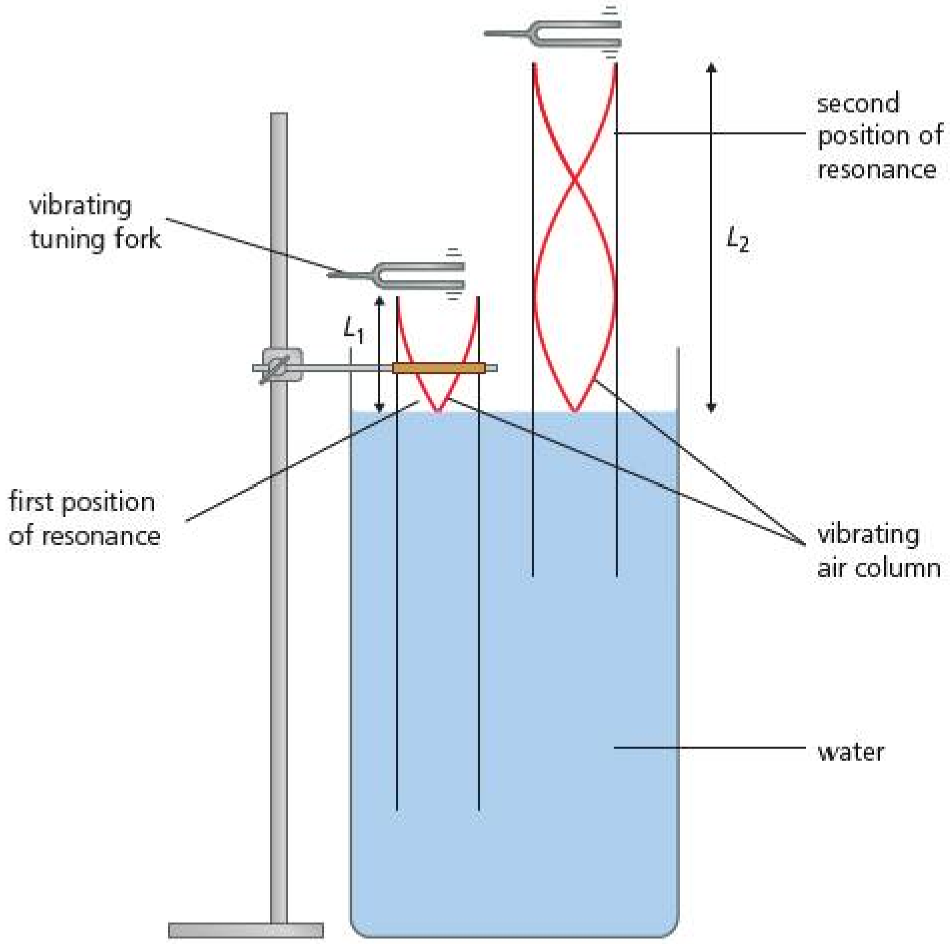


Figure 11.10 Demonstrating resonance with a tuning fork

The experiment shown in Figure 11.9 is a further demonstration of *resonance* (discussed in Chapter 4) because the applied frequency must be the same as one of the pipe's harmonic frequencies for energy to be transferred to move the powder.

Another way of demonstrating resonance in a pipe is by using a **tuning fork** of a known frequency, as shown in Figure 11.10. A vibrating tuning fork is held above the open end of a pipe. The pipe is open at the top and closed at the bottom by the level of water. The length of the pipe above the water is slowly increased until a louder sound is heard. This will be the first position of resonance. If the length of the pipe above water is increased again, then further positions of resonance may be found. Resonance will only occur when the length of the pipe above water is such that one of its harmonic frequencies is the same as the frequency of the tuning fork. Measurements can be made during this demonstration that will enable a value for the speed of sound to be determined.

Worked example

2 a If the tuning fork in Figure 11.10 had a frequency of 659 Hz, calculate the length L_1. (Assume the speed of sound in air is 340 m s^{-1}.)

b How far will the pipe need to be raised to obtain the next position of resonance?

a $\lambda = \frac{v}{f} = \frac{340}{659} = 0.516\,\text{m}$

This wavelength will be four times the length of the tube (see Figure 11.8).

So, $L_1 = \frac{0.516}{4} = 0.129\,\text{m}$

In reality this is only an approximate answer because antinodes do not occur exactly at the open ends of tubes. (If a more accurate answer is needed in a calculation, it is possible to use an 'end correction' which is related to the diameter of the tube.)

b Refer to Figure 11.10. In the first position of resonance the pipe contains one-quarter of a wavelength. In the second position the pipe contains three-quarters of a wavelength. Therefore, it must be raised half a wavelength, or 0.258 m.

Summary of differences between standing waves and travelling waves

11.1.4 Compare standing waves and travelling waves.

Table 11.1 Comparison of standing waves and travelling waves

	Standing waves	**Travelling waves**
Wave pattern	Stationary/standing	Progressive/travelling
Energy transfer	No energy is transferred	Energy is transferred through the medium
Amplitude (assuming no energy dissipation)	Amplitude at any one place is constant but it varies with position between nodes. Maximum amplitude at antinodes; zero amplitude at nodes	All oscillations have the same amplitude
Phase	All oscillations between adjacent nodes are in phase	Oscillations one wavelength apart are in phase. Oscillations between are not in phase
Frequency	All oscillations have the same frequency	All oscillations have the same frequency
Wavelength	Twice the distance between adjacent nodes	Shortest distance between points in phase

11.1.5 Solve problems involving standing waves.

1 A first harmonic is seen on a string of length 123 cm at a frequency of 23.8 Hz.
  a What is the wave speed?
  b What is the frequency of the third harmonic?
  c What is the wavelength of the fifth harmonic?

2 a What is the phase difference between two points on a standing wave which are:
    i one wavelength apart
    ii half a wavelength apart?
  b The distance between adjacent nodes of the third harmonic on a stretched string is 18.0 cm, with a frequency of 76.4 Hz. Sketch the waveform of this harmonic.
  c On the same drawing add the waveforms of the fundamental mode and the fourth harmonic.
  d Calculate the wavelength and frequency of the fifth harmonic.
  e What was the wave speed?

3 Explain why observing standing waves on a stretched string using a mechanical vibrator can be considered as a demonstration of resonance.

4 A certain guitar string has a fundamental frequency of 262 Hz.
  a If the tension in the string is increased, suggest what happens to the speed of waves along it.
  b If the string is adjusted so that the speed increases, explain what will happen to the fundamental frequency.
  c Suggest why a wave will travel more slowly along a thicker string of the same material, under the same tension.
  d Explain why thicker strings of the same material, same length and same tension produce notes of lower frequency.

5 Draw the first three harmonics for a pipe which is closed at both ends.

6 A and B are two similar pipes of the same length. A is closed at one end, but B is open at both ends. If the fundamental frequency of A is 180 Hz, what is the fundamental frequency of B?

7 If the frequency used in the demonstration shown in Figure 11.9 was 6.75 kHz and the piles of powder were 2.5 cm apart, what was the speed of the sound waves?

8 What length must an organ pipe (open at one end, see Figure 11.11) have if it is to produce a note of fundamental frequency 90 Hz?
(Use speed of sound = 340 m s^{-1})

9 An organ pipe open at both ends has a second harmonic of frequency 228 Hz.
  a What is its length?
  b What is the frequency of the third harmonic?
  c What is the wavelength of the fourth harmonic?
  d Suggest one advantage of using organ pipes that are closed at one end, rather than open at both ends.

Figure 11.11 On this church organ the wooden pipes are open at one end and the metal pipes behind are open at both ends

Additional Perspectives

Microwave ovens

The frequency of the microwaves which are used to cook food is chosen so that it is absorbed by water and other polarized molecules in the food. Such molecules are positively charged at one end and negative at the other end. They respond to the oscillating electromagnetic field of the microwave by gaining kinetic energy because of their increased vibrations, which means that the food gets hotter. Most microwave ovens operate at a frequency of 2.45 GHz. This frequency of radiation is penetrating, which means that the food its not just heated on the outside.

To ensure that the microwaves do not pass out of the oven into the surroundings, the walls, floor and ceiling are metallic, although the door may have a metallic mesh (with holes) that will reflect microwaves, but allow light, with its much smaller wavelength, to pass through, so that the contents of the oven can be seen from outside.

The microwaves reflect off the oven walls, which means that the walls do not absorb energy, as they do in other types of oven. The 'trapping' of the microwaves is the main reason why microwave ovens cook food quickly and efficiently. However, reflected microwaves can combine to produce various kinds of standing wave in the oven and this can result in the food being cooked unevenly. To reduce this effect the microwaves can be 'stirred' by a rotating deflector as they enter the cooking space, or the food can be rotated on a turntable.

Questions

1 Calculate the wavelength of microwaves used for cooking.
2 Design an experiment which would investigate if there were significant nodes and antinodes in a microwave oven. How far apart would you expect any nodes to be?

11.2 The Doppler effect

11.2.1 **Describe** what is meant by the Doppler effect.
11.2.2 **Explain** the Doppler effect by reference to wavefront diagrams for moving-detector and moving-source situations.

When we hear a sound we can usually assume that the frequency (pitch) that is heard by our ears is the same as the frequency that was emitted by the source. But if the source of the sound is moving towards us (or away from us) we will hear a sound with a different frequency. This is usually only noticeable if the movement is fast; the most common example is the sound heard from a car or train that moves quickly past us.

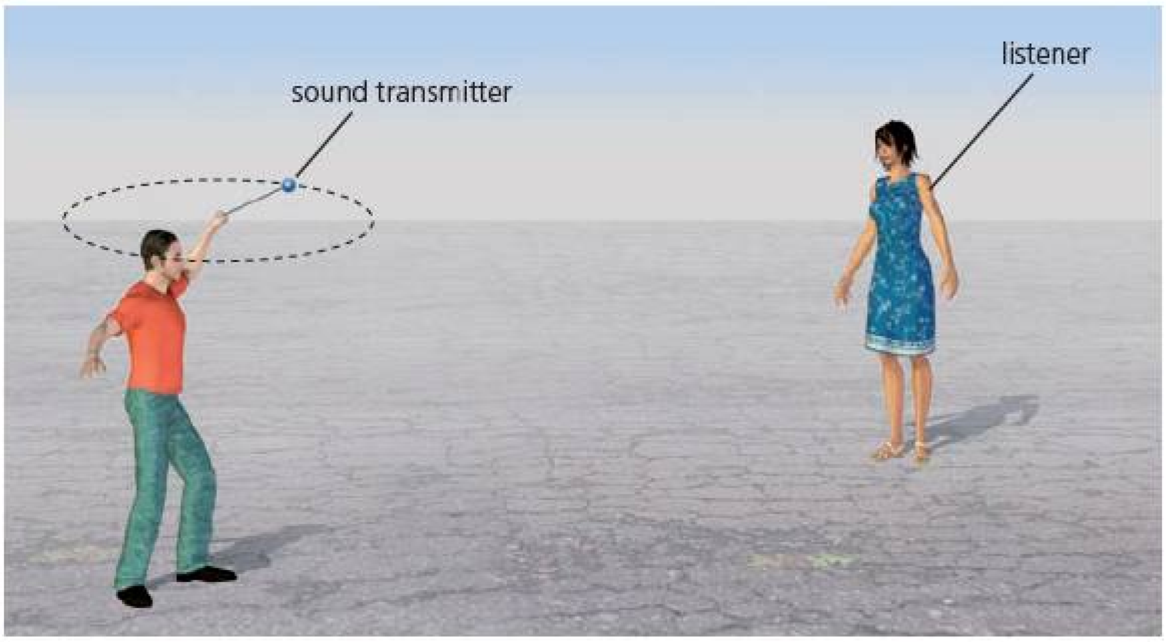


Figure 11.12 Demonstrating the Doppler effect with sound waves

This change of frequency that is detected when there is relative motion between a source and a receiver of waves is called the **Doppler effect**. (The Doppler effect is named after the Austrian physicist, Christian Doppler, who first proposed it in 1842.) The Doppler effect may occur with any kind of wave.

Figure 11.12 shows a way in which the Doppler effect with sound can be demonstrated. A small source of sound (of a single frequency) is spun around in a circle. When the source is moving towards the listener a higher frequency is heard; when it is moving away, a lower frequency is heard.

The easiest way to explain the Doppler effect is by considering wavefronts (Figure 11.13). Figure 11.13a shows the common situation in which a stationary source, S, emits waves which travel towards a stationary detector, D. Figure 11.13b shows a detector moving towards a stationary source and Figure 11.13c shows a source moving directly towards a stationary detector. Similar diagrams can be drawn to represent the situations where the source and detector are moving apart.

The detector in **b** will meet more wavefronts in a given time than if it remained in the same place, so that the received frequency, f', is greater than the emitted frequency, f. In **c** the distance between the wavefronts (the wavelength λ) between the source and the observer is reduced, which again means that the received frequency will be greater than the emitted frequency. (Frequency = v/λ and the wave speed, v, is constant. The speed of sound through air does not vary with the motion of the source or observer.)

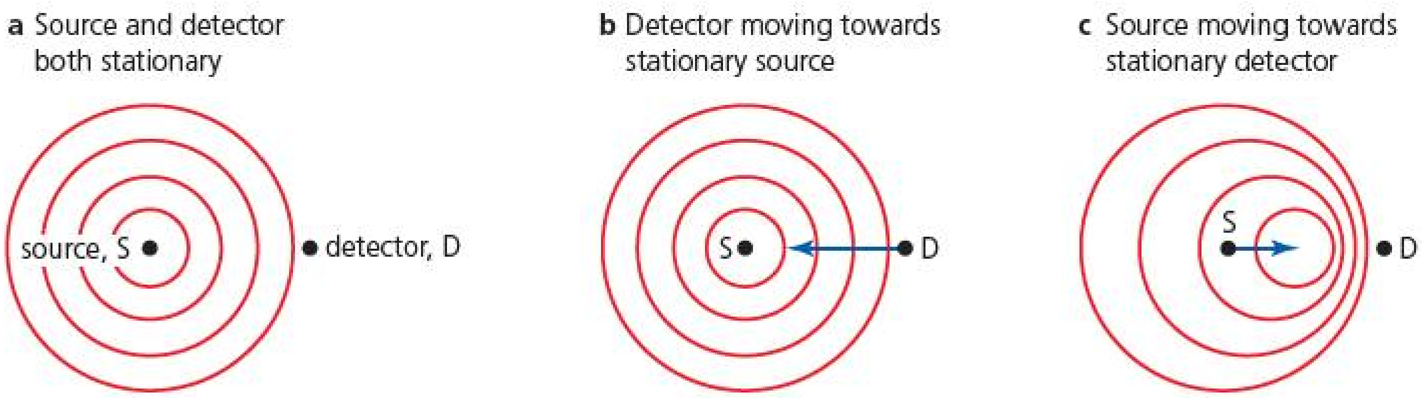


Figure 11.13 Wavefront diagrams to demonstrate the Doppler effect

Equations for the Doppler effect

11.2.3 Apply the Doppler effect equations for sound.

Figure 11.14a shows waves of frequency, f, and wavelength, λ, travelling at a speed, v, between a stationary source, S, and a stationary observer, O. (The term *observer* can be used with any kind of waves, not just light.) In the time, t, that it takes the first wavefront emitted from the source to reach the observer, the wave has travelled a distance vt. The number of waves between the source and observer is ft. The wavelength, λ, equals the total distance divided by the number of waves $= vt/ft = v/f$, as we would expect.

Figure 11.14b represents exactly the same waves emitted in the same time from a source moving towards a stationary observer with a speed u_s. In time, t, the source has moved from S_1 to S_2. The number of waves is the same as in **a**, but because the source has moved forwards a distance, u_st, the waves between the source and the observer are now compressed within a length, $(vt - u_st)$.

This means that the observed (received) wavelength, λ', equals the total distance divided by the number of waves:

$$\lambda' = \frac{vt - u_st}{ft} = \frac{v - u_s}{f}$$

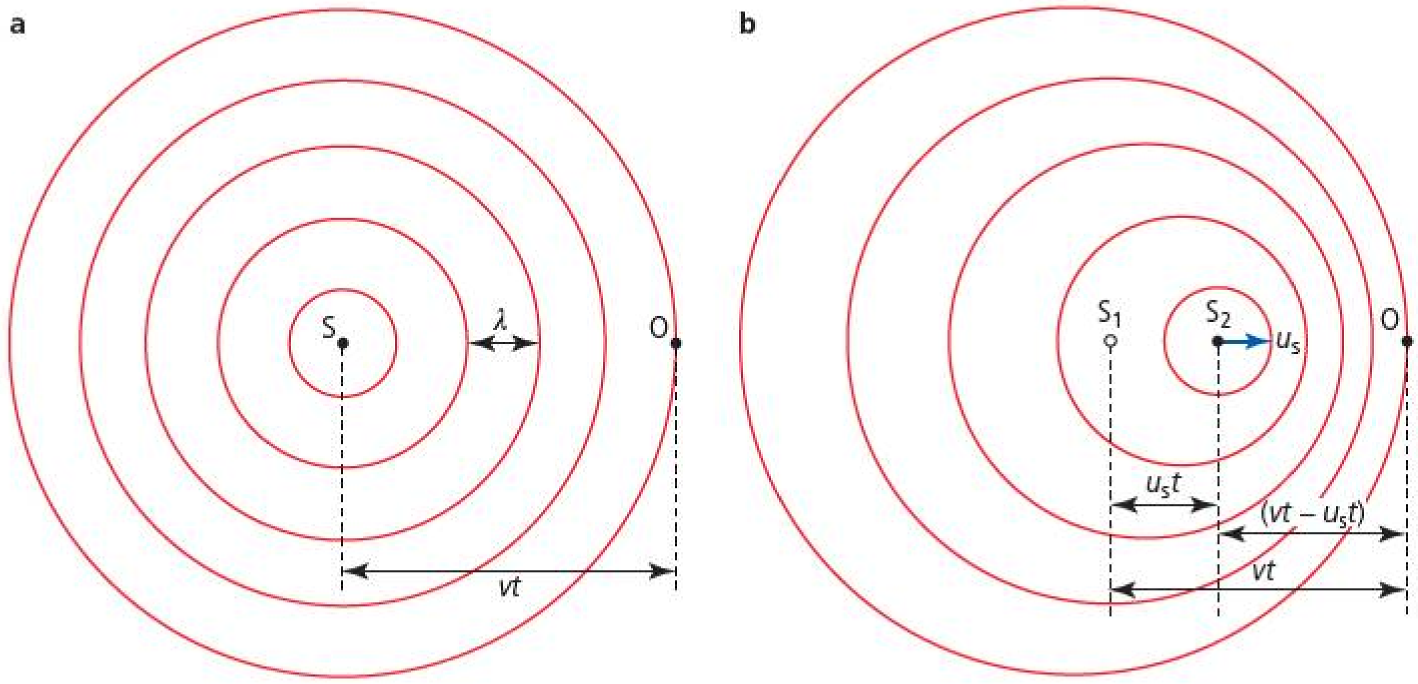


Figure 11.14a Waves between a stationary source and a stationary observer **b** Waves between a moving source and a stationary observer

The observed (received) frequency, f', is given by:

$$f' = \frac{v}{\lambda} = \frac{vf}{v - u_s}$$

If the source is moving away from the observer, the equation becomes

$$f' = \frac{vf}{v + u_s}$$

In general, we can write:

$$f' = \left(\frac{v}{v \pm u_s}\right)$$

This is the equation for the Doppler effect from a *moving source* detected by a stationary observer and it is given in the IB *Physics data booklet*.

In a similar situation, the equation for the frequency detected by a *moving observer* from a stationary source is:

$$f' = f\left(\frac{v \pm u_o}{v}\right)$$

This equation is also given in the IB *Physics data booklet*.

If the source of the sound and the observer are getting closer together, along a line directly between them, a higher frequency sound (than that emitted) will be detected by the observer and that frequency will be constant, although it will increase in intensity. Similarly, if the source and the observer are moving apart, the observed sound will have a lower, constant frequency (than that emitted) and it will decrease in intensity. The frequency must change quickly at the moment the source and detector move past each other. (If the motion is not directly between the source and the observer, or both the source and the observer are moving, the principles are the same, but the mathematics is more complicated and it is not included in this course.)

Worked examples

3 a A source of sound emitting a frequency of 480 Hz is moving directly towards a stationary observer at $50\,\text{m s}^{-1}$. If it is a hot day and the speed of sound is $350\,\text{m s}^{-1}$, what frequency is received?

b What frequency would be heard on a cold day when the speed of sound was $330\,\text{m s}^{-1}$?

c Explain why the speed of sound is less on a colder day.

a $f' = \dfrac{vf}{(v - u_s)}$

$f' = \dfrac{(350 \times 480)}{(350 - 50)} = 560\,\text{Hz}$

b $f' = \dfrac{(330 \times 480)}{(330 - 50)} = 566\,\text{Hz}$

c Sound is transferred though air by moving air molecules. On a colder day the molecules will have a lower average speed.

4 What frequency will be received by an observer moving at $24\,\text{m s}^{-1}$ directly away from a stationary source of sound waves of frequency 980 Hz? (Take the speed of sound to be $342\,\text{m s}^{-1}$.)

$f' = \dfrac{f(v \pm u_o)}{v}$

$f' = \dfrac{980 \times (342 - 24)}{342} = 910\,\text{Hz}$

11.2.4 Solve problems on the Doppler effect for sound.

10 An observer receives sound of frequency 436 Hz from a train moving at $18\,\text{m s}^{-1}$ directly towards him. What was the emitted frequency? (Take the speed of sound to be $342\,\text{m s}^{-1}$.)

11 An observer is moving directly towards a source of sound emitted at 256.0 Hz, with a speed of $26.0\,\text{m s}^{-1}$. If the received sound has a frequency of 275.7 Hz, what was the speed of sound?

12 A car emitting a sound of 190 Hz is moving directly away from an observer who detects a sound of frequency 174 Hz. What was the speed of the car? (Take the speed of sound to be $342\,\text{m s}^{-1}$.)

13 A train is moving at constant speed along a track as shown in Figure 11.15 and is emitting a sound of constant frequency.

a Suggest how the sound heard by an observer at point P will change as the train moves from A to B.

b Describe the sound heard by someone sitting on the train during the same time.

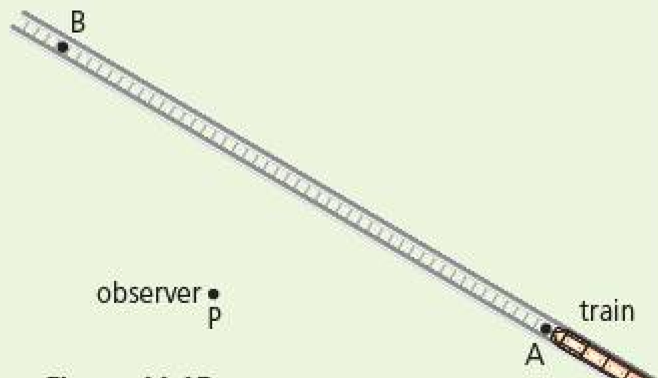


Figure 11.15

14 The source of sound shown in Figure 11.12 is rotating at 4.2 revolutions per second in a circle of radius 1.3 m. If the emitted sound has a frequency of 287 Hz, what is the difference between the highest and lowest frequencies which are heard? (Take the speed of sound to be $340\,\text{m s}^{-1}$.)

The Doppler effect with electromagnetic waves

11.2.5 Solve problems on the Doppler effect for electromagnetic waves using the approximation $\Delta f = \frac{v}{c}f$.

The Doppler effect also occurs with electromagnetic waves, but the situation is made more complicated because the received speed of electromagnetic waves is unaffected by the speed of the observer. The equations given in the previous section cannot be used with electromagnetic waves.

However, the following equation for the change (shift) in frequency, Δf, can be used if the relative speed between source and observer, v, is very much less than the speed of the electromagnetic waves, c, ($v << c$).

$$\Delta f = \frac{v}{c}f$$

This equation is given in the IB *Physics data booklet*.

Since the speed of electromagnetic waves is so high ($c = 3.00 \times 10^8\,\text{m s}^{-1}$ in vacuum or air) this equation can nearly always be used with accuracy.

Worked example

5 A plane travelling at a speed of $250\,\text{m s}^{-1}$ transmits a radio signal at a frequency of 130 MHz. What change of frequency will be detected by the airport it is travelling towards?

$$\Delta f = \frac{v}{c}f = \frac{250}{(3.00 \times 10^8)} \times (1.3 \times 10^8) = 108\,\text{Hz}$$

The airport will receive a frequency 108 Hz higher than 1.3×10^8 Hz. This is a very small increase, requiring good quality electronic circuits to be detected.

Clearly the shift in frequencies of electromagnetic waves is not something we will observe in everyday life. It becomes more significant for very fast moving objects like stars.

Using the Doppler effect to measure speed

11.2.6 Outline an example in which the Doppler effect is used to measure speed.

In order to measure the (average) velocity of a moving object we usually observe the position of the object on two occasions between known time intervals. One way of doing this, especially if the object is inaccessible, like a plane for example, is to send waves towards it and then detect the waves as they are reflected back to the transmitter. From this data, the direction from the transmitter to the plane can be calculated. The distance between the transmitter and the plane can be calculated from the time delay between the sent and the received signals. If this process is repeated, the velocity of the plane can be determined. In this example of a simple **radar** system, microwaves are used.

The Doppler effect, however, provides a better way of determining the speed of a moving object, like a plane, using the same kind of waves (see Figure 11.16). Some animals, like bats (see Figure 11.17) and dolphins, use the same principles but with sound or ultrasonic waves.

Figure 11.16 Air traffic control uses the Doppler effect

If waves of a known speed, v, and frequency, f, are directed towards a moving object and reflected back, the object will effectively be acting like a moving source of waves; the Doppler equations can be used to determine the speed of the object if the received frequency, f', can be measured.

The measurement of the speeds of cars, as well as planes and other vehicles, is a common example of the use of the Doppler effect. In many countries the police reflect electromagnetic radiation off cars in order to determine their speeds. Microwaves and infrared radiation are commonly used for this purpose. The speed of athletes or balls in sports can also be determined using the Doppler effect. The measurement of the rate of blood flow in an artery is another interesting example, which is shown in Figure 11.18.

Figure 11.17 These bats in Malaysia use the Doppler effect to navigate

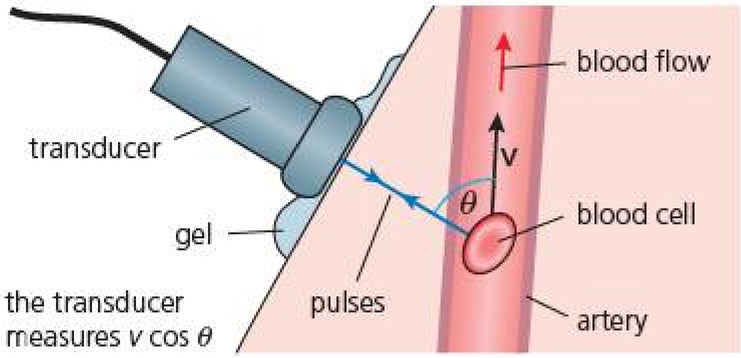


Figure 11.18 Measuring blood flow rate using the Doppler effect

Pulses of ultrasonic waves are sent into the body from the **transducer** and are reflected back from blood cells flowing in an artery. The received waves have a different frequency because of the Doppler effect and the change of frequency can be used to calculate the speed of the blood. This information can be used by doctors to diagnose many medical problems. Because the waves usually cannot be directed along the line of blood flow, the calculated speed will be the component ($v \cos\theta$).

There are also some very important applications of the Doppler effect used in astronomy. The decrease in frequency of radiation received from distant galaxies is known as the 'red shift', and it provides very important information about the nature and age of the universe.

15 **a** What change of frequency will be received back from a car moving directly away at 135 km h^{-1} if the radiation used in the 'speed gun' has a frequency of 24 GHz?
b Suggest reasons why ultrasonic waves are not usually used in speed guns.

16 **a** An airport radar system using microwaves of frequency 98.2 MHz sends out a pulse of waves that is reflected off a plane which is flying directly away. If the reflected signal was received back at the airport 8.43×10^{-5} s later, at a frequency 85.2 Hz lower than was emitted, what was the speed of the plane and how far away was it?
b Suggest how it might be possible for a plane to avoid being detected by radar.

17 In the television broadcast of some sports, for example, baseball, tennis and cricket, viewers can see a replay of the trajectory (path) of a moving ball. Use the Internet to find out how this is done. (Is the Doppler effect used?)

18 A star emits radiation of frequency 1.42×10^9 Hz. When received on Earth the frequency is 1.38×10^9 Hz. What is the speed of the star? Is it moving towards the Earth, or away from the Earth?

■ **Additional Perspectives**

Breaking the sound barrier

As an object, like a plane, flies faster and faster, the sound waves that it makes get closer and closer together in front of it. When a plane reaches the speed of sound, at about 1200 km h^{-1}, the waves superpose to create a 'shock wave'. This is shown in Figure 11.20.

When a plane reaches the speed of sound it is said to be travelling at 'Mach 1' (named after the Austrian physicist, Ernst Mach). Faster speeds are described as 'supersonic' and twice the speed of sound is called Mach 2, etc. As Figure 11.19 shows, the shock wave travels away from the side of the plane and may be heard on the ground as a 'sonic boom'. Similarly, 'bow waves' can often be seen spreading from the front (bow) of a boat because boats usually travel faster than the water waves they create. The energy transferred by such waves can cause a lot of damage to the land at the edges of rivers (river banks).

For many years some engineers doubted if the sound barrier could ever be broken. The first confirmed supersonic flight (with a pilot) was in 1947. Now it is common for military aircraft to travel faster than Mach 1, but Concorde and Tupolev 144 were the only supersonic passenger aircraft in regular service.

It is possible to use a whip to break the sound barrier. If the whip gets thinner towards its end then the speed of a wave along it can increase until the tip is travelling faster than sound (in air). The sound it produces is often described as a whip 'cracking'.

Figure 11.19 Plane breaking the sound barrier

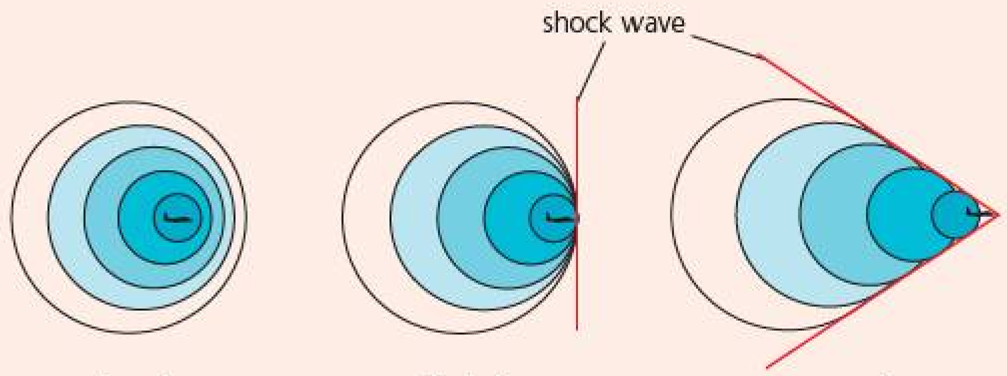


Figure 11.20 Creating a shock wave in air

Question

1 In a world in which time is so important for so many people, suggest reasons why there are no longer supersonic aircraft in commercial operation.

11.3 Diffraction at a single slit

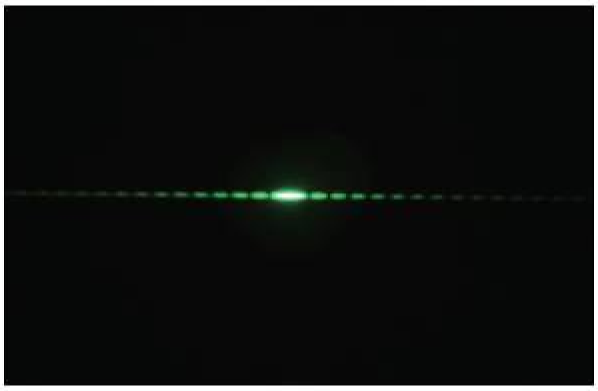


Figure 11.21 Diffraction pattern of monochromatic light passing through a narrow slit

Figure 11.21 shows a photograph of a diffraction pattern for light passing through a narrow vertical slit, which is the type of **aperture** (opening) usually used to produce simple diffraction patterns. The diffraction pattern seen on the screen is a series of bands (fringes) of light and dark. The central band is brighter and about twice the width of the others.

Figure 11.22 shows the diffraction pattern produced when light passes through a small circular aperture.

By comparing Figures 11.21 and 11.22, we can see that the shape of the pattern depends on the shape of the aperture. The light used to produce these patterns was **monochromatic**, which means that it has only one wavelength (and frequency). Light from most sources will consist of many different wavelengths and will not produce such a simple pattern.

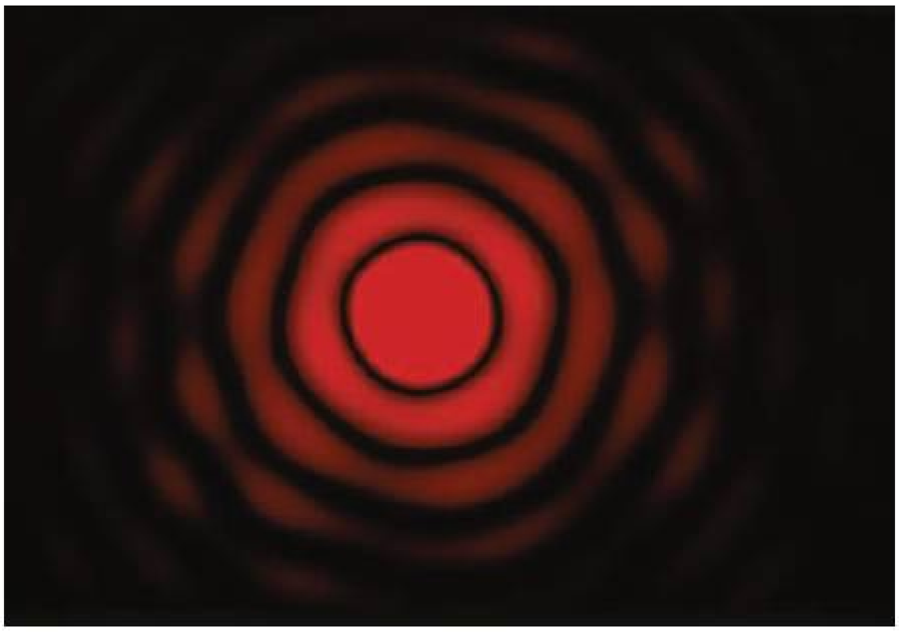


Figure 11.22 Monochromatic light diffracted by a very small circular aperture

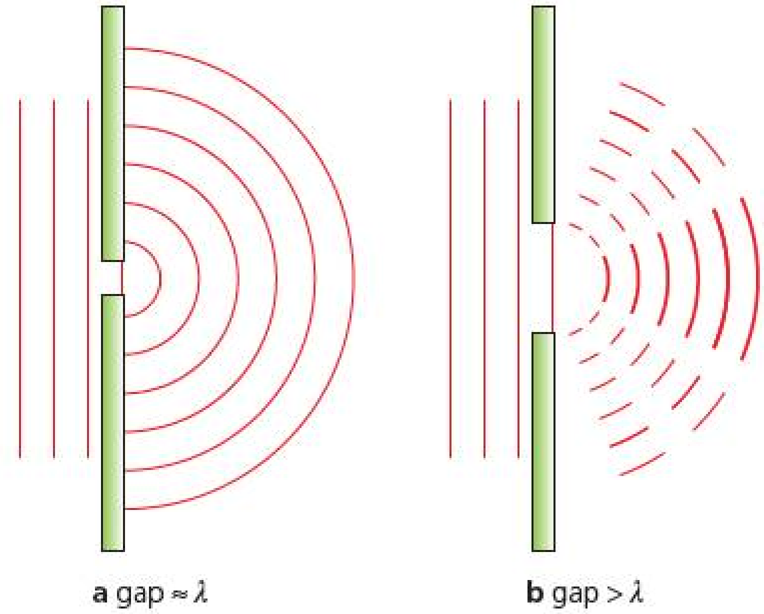


Figure 11.23 Diffraction of waves by different sized apertures

In order to produce these diffraction patterns for light, the waves must travel away from the aperture in some directions, but not in others. The simple diffraction theory covered in Chapter 4 does not explain this. Figure 11.23a shows the diffracted waves coming away from a gap of about the same width as the wavelength (as discussed in Chapter 4). The gap acts effectively like a point source, spreading waves forward equally in *all* directions. This *cannot* be used to help to explain the diffraction of light because the wavelengths of light are so much smaller than the width of even a very small hole. For example, a hole of diameter 0.5 mm is 1000 times greater than the wavelength of green light (approximately 5×10^{-7} m).

To produce a diffraction pattern for light, such as shown in Figure 11.21, we must develop a theory which explains why waves spread away from a gap that is larger than the wavelength, as shown in Figure 11.23b.

Explaining the diffraction pattern produced by a single slit

11.3.2 Derive the formula $\theta = \lambda/b$ for the position of the first minimum of the diffraction pattern produced at a single slit.

Diffraction patterns can be explained by considering that each point on a wavefront passing through the slit acts as a point source of **secondary waves**. (This idea was first suggested in 1678 by the Dutch physicist Christian Huygens, when he proposed that light waves propagate forwards because *all* the points on any wavefront act as sources of secondary waves.) What happens after the wave has passed through the slit depends on how these secondary waves **interfere** with each other.

Figure 11.24 shows a direction, θ, in which secondary waves travel away from a slit of width, b. If θ is zero, all the secondary waves will interfere constructively in this direction (straight through the slit) because there is no **path difference** between them. (Of course, in theory, waves travelling parallel to each other in the same direction cannot meet and interfere, so we will assume that the waves' directions are *very nearly* parallel.)

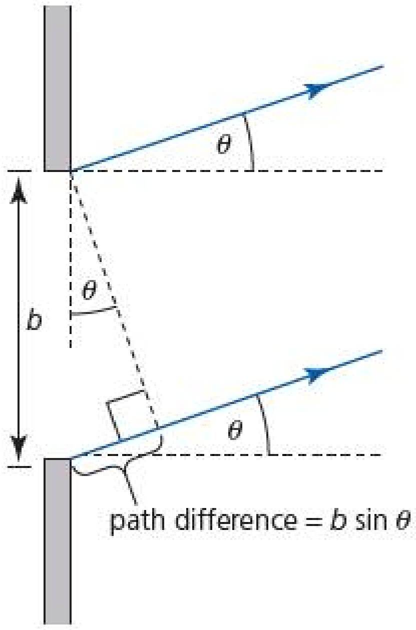


Figure 11.24 Path differences and interference

Consider what happens for angles increasingly greater than zero. The path difference, as shown in Figure 11.24, equals $b\sin\theta$ and this increases as the angle θ increases. There will be an angle at which the waves from the two edges of the slit interfere constructively because the path difference has increased to become one wavelength, λ.

But if secondary waves from the edges of the slit interfere constructively, what about interference between all the other secondary waves? Consider Figure 11.25 on page 386 in which the slit has been divided into a number of point sources of secondary waves. (Ten points have been chosen, but it could be many more.)

If the angle, θ, is such that secondary waves from points 1 and 10 would interfere constructively because the path difference is one wavelength, then secondary waves from 1 and 6 must have a path difference of half a wavelength and interfere destructively. Similarly, waves from points 2 and 7, points 3 and 8, points 4 and 9 and points 5 and 10 must all interfere destructively. In this way waves from *all* points can be 'paired off' with

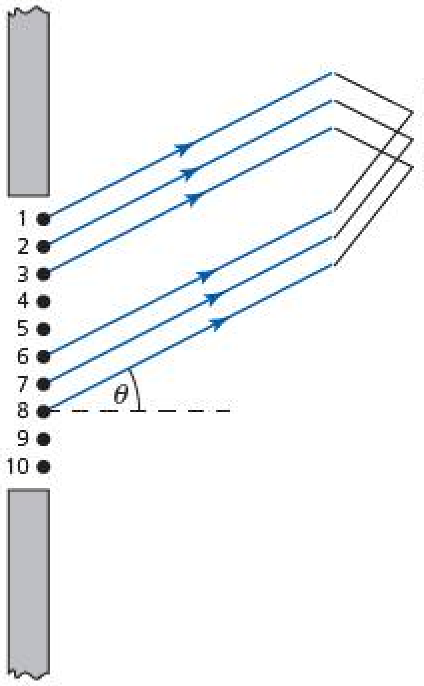


Figure 11.25 Secondary waves which will interfere destructively can be 'paired off'

others, so that the first *minimum* of the diffraction pattern occurs at such an angle that waves from the edges of the slit would otherwise interfere constructively.

The first minimum of the diffraction pattern occurs when the path difference between secondary waves from the edge of the slit is equal to one wavelength. That is, if $b\sin\theta = \lambda$.

For the diffraction of light, the angle θ is usually small and approximately equal to $\sin\theta$ if the angle is expressed in radians. (This was first explained in Chapter 4.)

The angle for the first minimum of a single slit diffraction pattern is $\theta = \dfrac{\lambda}{b}$.

This equation is given in the IB *Physics data booklet*.

Relative intensity–angle graph for single slit diffraction

11.3.1 Sketch the variation with angle of diffraction of the relative intensity of light diffracted at a single slit.

Similar reasoning to that discussed in the previous section can be used to show that further diffraction minima occur at angles $2\lambda/b$, $3\lambda/b$, $4\lambda/b$, etc. Figure 11.26 shows a graphical interpretation of these figures and approximately how it corresponds to a drawing of a single slit diffraction pattern.

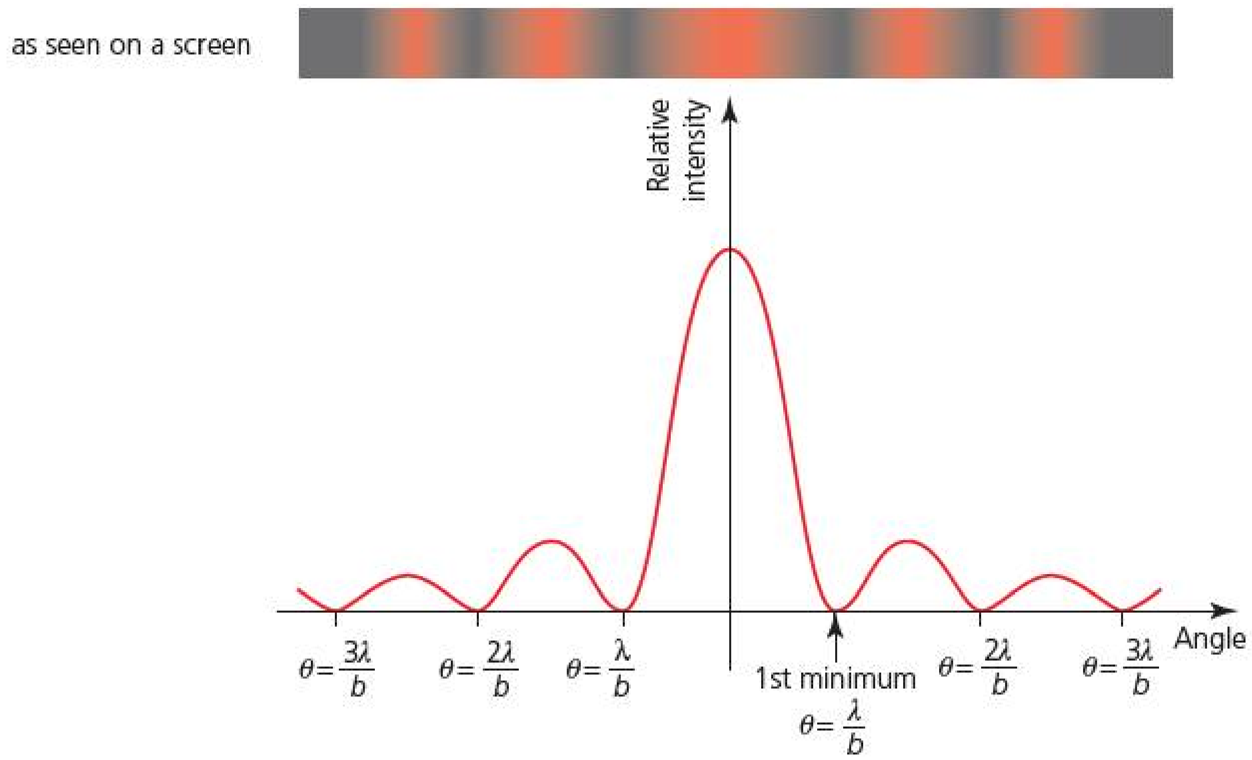


Figure 11.26 Variation of intensity with angle for single slit diffraction

Worked example

6 Monochromatic light of wavelength 663 nm is shone through a gap of width 0.0730 mm.

a At what angle is the first minimum of the diffraction pattern formed?

b If the pattern is observed on a screen which is 2.83 m from the slit, what is the width of the central maximum?

a $\theta = \lambda/b$

$$\theta = \frac{663 \times 10^{-9}}{7.30 \times 10^{-5}}$$

$$= 9.08 \times 10^{-3} \text{ radians}$$

b See Figure 11.27.

$$\theta \approx \sin\theta$$

$$= \frac{\text{half width of central maximum}}{\text{slit to screen distance}}$$

half width of central maximum = $(9.08 \times 10^{-3}) \times 2.83 = 0.0257\,\text{m}$
width of central maximum = $0.0257 \times 2 = 0.0514\,\text{m}$

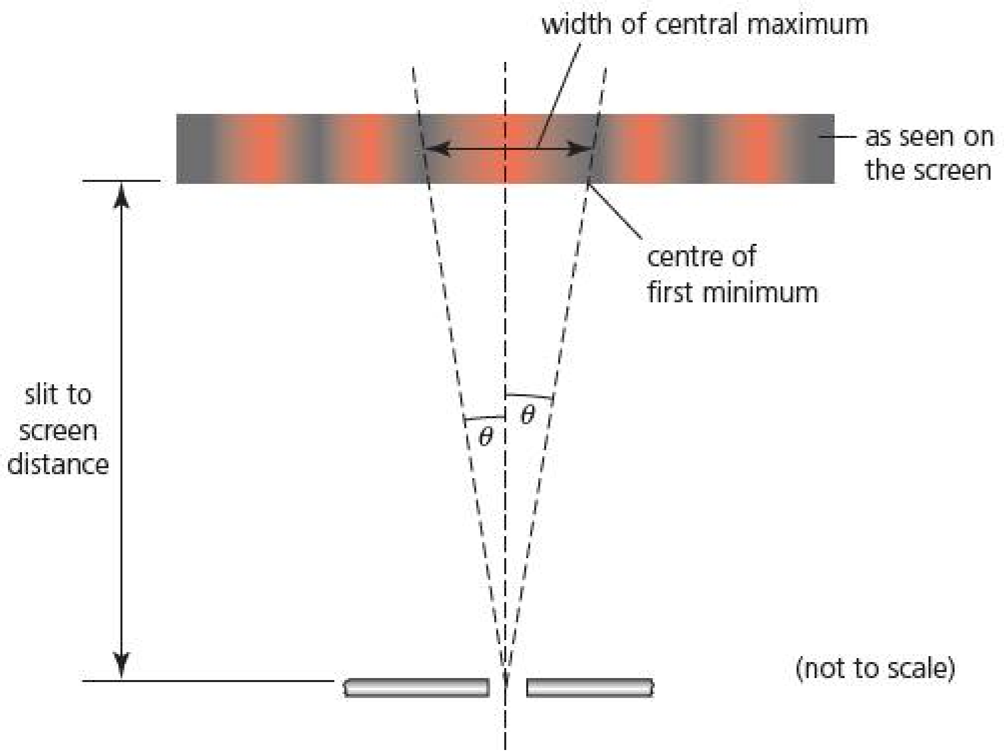


Figure 11.27

11.3.1 Solve problems involving single slit diffraction.

19 Electromagnetic radiation of wavelength $2.37 \times 10^{-7}\,\text{m}$ passes through a narrow slit of width $4.70 \times 10^{-5}\,\text{m}$.
  a In what part of the electromagnetic spectrum is this radiation?
  b Suggest how it could be detected.
  c Calculate the angle of the first minimum of the diffraction pattern.

20 What is the wavelength of light that has a first diffraction minimum at an angle of 0.0038 radians when it passes through a slit of width 0.15 mm?

21 When light of wavelength $6.2 \times 10^{-7}\,\text{m}$ was diffracted through a narrow slit, the central maximum had a width of 2.8 cm on a screen which was 1.92 m from the slit. What was the width of the slit?

22 **a** Sketch and label a relative intensity against angle graph for the diffraction of red light of wavelength $6.4 \times 10^{-7}\,\text{m}$ through a slit of width 0.082 mm. Include at least five peaks of intensity.
  b Add to your graph a sketch to show how monochromatic blue light would be affected by the same slit.

The importance of diffraction

The diffraction of light is further evidence confirming the wave nature of light, and when it was discovered that electrons could be diffracted, their wave nature was also understood for the first time (see page 392 and Chapter 13). Other electromagnetic radiations will be diffracted significantly by objects and gaps that have a size comparable to their wavelengths. For example, the diffraction of X-rays by atoms, ions and molecules can be used to provide detailed information about solids with regular structure. The diffraction of radio waves used in communication is also of considerable importance when choosing which wavelength to use in order to get information efficiently from the transmitter to the receiver.

Waves travelling from place to place can have their direction changed by diffraction, but diffraction can also be important when waves are emitted or received. They can then be considered to be passing through an aperture, and the ratio λ/b will be important when considering how the waves spread away from a source, or what direction they appear to come from when they are received.

■ **Additional Perspectives**

Diffraction by multiple apertures

We have discussed how light waves are affected by passing through a single slit, but diffraction effects can be greatly increased by the use of many identical slits arranged close together in a regular pattern. Such pieces of apparatus are called diffraction gratings. Figure 11.28 shows a diffraction grating, which will typically have 500 or more slits in every millimetre.

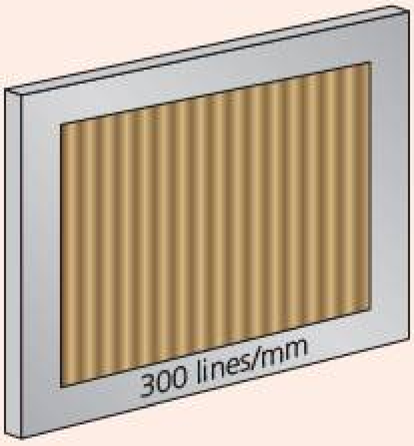


Figure 11.28 A diffraction grating

The use of multiple apertures greatly increases the intensity and the sharpness of diffraction patterns. Diffraction gratings are often used in a similar way to prisms for the analysis of light (discussed in Chapter 7).

Question

1 What is the approximate ratio λ/b for each slit in a diffraction grating that has 500 slits per millimetre?

11.4 Resolution

If you stand on a beach and look down at the sand, you probably will not be able to see the separate grains of sand. Similarly, if you look at a tree which is a long way away, you will not be able to see the separate leaves. In scientific terms, we say that you cannot **resolve** the detail. If you assume that a person has good eyesight, the ability of their eyes to resolve detail depends on the diffraction of light as it enters the eye through the **pupil** (aperture) and the separation of the light receptors on the **retina** of the eye. The larger the pupil, the less the diffraction and, so, the better the resolution. You can check this by looking at the world through a *very* small hole made in a piece of paper held in front of your eye (increasing the diffraction of light as it enters your eye).

When discussing resolution, in order to improve understanding, we usually simplify the situation by only considering waves of a single frequency coming from two point sources of light.

How wavelength and aperture width affect resolution

11.4.1 **Sketch** the variation with angle of diffraction of the relative intensity of light emitted by two point sources that has been diffracted at a single slit.

Figure 11.29 shows an eye looking at two distant identical point objects, O_1 and O_2; θ is called the **angular separation** of the objects. (This is sometimes called the angle **subtended** at the eye.)

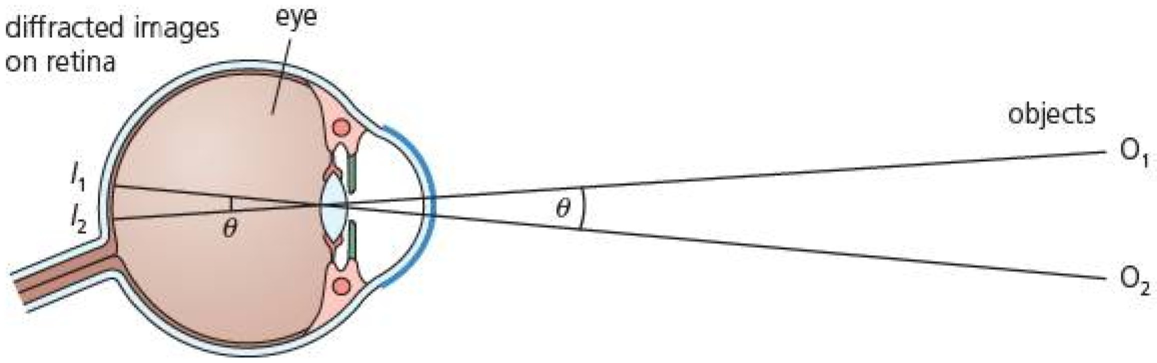


Figure 11.29 An eye receiving light from two separate objects

When the light from each of the objects enters the eye it will be diffracted and two single slit diffraction patterns (similar to that shown in Figure 11.26) will be received by the light receptors on the retina at the back of the eye. So, point objects do not form point images. (Note that the light rays are not shown being refracted by the eye in this example because they are passing through the optical centre of the lens.)

The ability of this eye to resolve two separate images depends on how much the two diffraction patterns overlap (assuming that the receptors in the retina are close enough together). Consider the diagrams in Figure 11.30, which show intensities which might be received by the eye. In **a** the two sources can easily be resolved because their diffraction patterns do not overlap. In **c** the sources are so close together that their diffraction patterns merge together and the eye cannot detect any fall in the resultant intensity between them; the images are not resolved.

Figure 11.30b represents the situation in which the sources can *just* be resolved because, although the images are close, there is a detectable fall in resultant intensity between them.

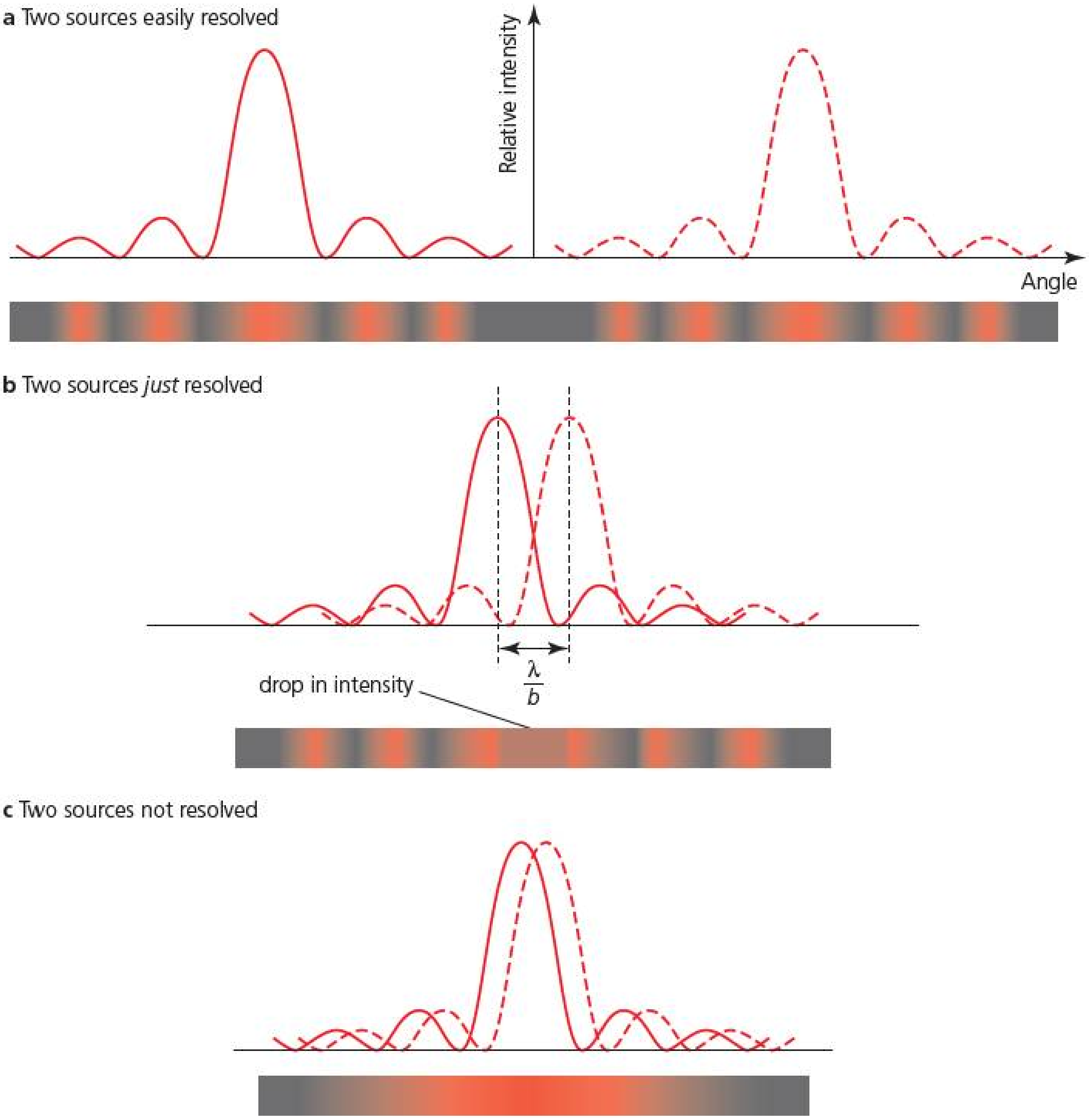


Figure 11.30 Intensities received by the eye as a result of two diffraction patterns

Rayleigh criterion

11.4.2 State the Rayleigh criterion for images of two sources to be just resolved.

Figure 11.31a shows two point sources viewed through circular apertures.

Rayleigh's criterion states that two point sources are *just* resolved if the first minimum of the diffraction pattern of one occurs at the same angle as the central maximum of the other (Figure 11.31b).

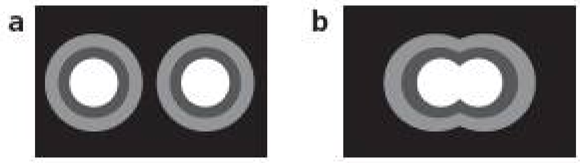


Figure 11.31 Images of two point sources observed through circular apertures that are **a** easily resolvable and **b** just resolvable

Rayleigh's criterion is a useful guide, not a law of physics, and there may be factors other than diffraction which can affect resolution. This can be expressed mathematically, as follows.

The images of two sources can *just* be resolved through a *narrow slit*, of width b, if they have an angular separation of $\theta = \frac{\lambda}{b}$.

If they subtend a larger angle they will be resolvable but if they subtend a smaller angle then they cannot be resolved.

When we consider eyes, telescopes and microscopes it is clear that light waves are usually received and detected through circular apertures, rather than slits. The resolution of a circular aperture will not be as good as with a slit of the same dimensions, so the criterion is adjusted, as follows.

The images of two sources can *just* be resolved through a *circular aperture* if they have an angular separation of $\theta = 1.22\frac{\lambda}{b}$.

This equation is given in the IB *Physics data booklet*.

Worked example

7 Two small point sources of light separated by 1.1 cm are placed 3.6 m away from an observer who has a pupil diameter of 1.9 mm. Can they be seen as separate when the average wavelength of the light used was 5.0×10^{-7} m?

They will be resolvable if their angular separation is greater than or equal to $1.22\lambda/b$.

$$1.22\frac{\lambda}{b} = \frac{1.22 \times 5.0 \times 10^{-7}}{1.9 \times 10^{-3}} = 3.2 \times 10^{-4} \text{ radians}$$

$$\text{Angular separation of sources} = \frac{1.1 \times 10^{-2}}{3.6} = 3.1 \times 10^{-3} \text{ radians}$$

The angular separation is much greater than $1.22\lambda/b$, so they are easily resolvable.

Significance of resolution

Radio telescopes

11.4.3 Describe the significance of resolution in the development of devices such as CDs and DVDs, the electron microscope and radio telescopes.

Stars emit many other kinds of electromagnetic radiation apart from light and infrared. Radio telescopes, like the one shown in Figure 11.32, detect wavelengths much longer than visible light and, therefore, would have very poor resolution if they were not made with big diameters.

Figure 11.32 The Jodrell Bank radio telescope in England, UK

Worked example

8 a The Jodrell Bank radio telescope in the UK has a diameter of 76 m. When used with radio waves of wavelength 21 cm, is it capable of resolving two stars which are 2.7×10^{11} m apart if they are both 1.23×10^{16} m from Earth?

b Compare the ability of this radio telescope to resolve with that of a typical human eye.

a $1.22\dfrac{\lambda}{b} = \dfrac{1.22 \times 21 \times 10^{-2}}{76} = 3.4 \times 10^{-3}$ radians

Angular separation of sources $= \dfrac{2.7 \times 10^{11}}{1.23 \times 10^{16}} = 2.2 \times 10^{-5}$ radians

The angular separation is much less than $1.22\lambda/b$, so they cannot be resolved – they will appear like a single star.

b Using the data from Worked example 7 as an example, a human eye can resolve two objects if their angular separation is about 3×10^{-4} radians or greater, but the radio telescope in this question requires the objects to be at least 3.4×10^{-3} radians apart. This angle is about ten times bigger. The eye has a much better resolution than a radio telescope mainly because it uses waves of a much smaller wavelength. The resolution of radio telescopes can be improved by making them even larger, but there are constructional limits to how big they can be made.

Additional Perspectives

More about radio telescopes

Despite their poor resolution (compared to the human eye), radio telescopes collect a lot of information about the universe that is not available from visible light radiation. Light waves from distant stars are affected when they pass through the Earth's atmosphere and this will decrease the possible resolution. This is why optical telescopes are often placed on mountain tops or on orbiting satellites. Radio telescopes do not have this problem.

Hydrogen is the most common element in the universe and its atoms emit electromagnetic waves with wavelength 21 cm, in the part of the electromagnetic spectrum known as radio waves. Many other similar wavelengths (from a few centimetres to many metres) are received on Earth from space and radio telescopes are designed to detect these waves.

In a dish telescope, such as that shown in Figure 11.32, the radio waves reflect off a parabolic reflector and are focused on a central receiver (like a reflecting optical telescope.) As we have seen, the resolution of a radio telescope will be disappointing unless its dish has a large aperture. Of course, having a large aperture also means that the telescope is able to detect fainter and more distant objects because more energy is received from them when using a larger dish.

Figure 11.33 An array of linked radio telescopes

There are many different designs of radio telescope, but the largest with a single dish is that built at Arecibo in Puerto Rico (diameter of 305 m), which uses a natural hollow in the ground to help provide support. However, a radio telescope is being built in China with a diameter of 500 m, which is due to be operational in 2013.

In radio *interferometry* signals received from individual telescopes, grouped together in a regular pattern (an **array**, as in Figure 11.33), are combined to produce a superposition (interference) pattern which has a much narrower spacing than the diffraction pattern of each individual telescope. This greatly improves the resolution of the system.

Question

1 Find out what you can about China's *FAST* telescope.

Electron microscopes

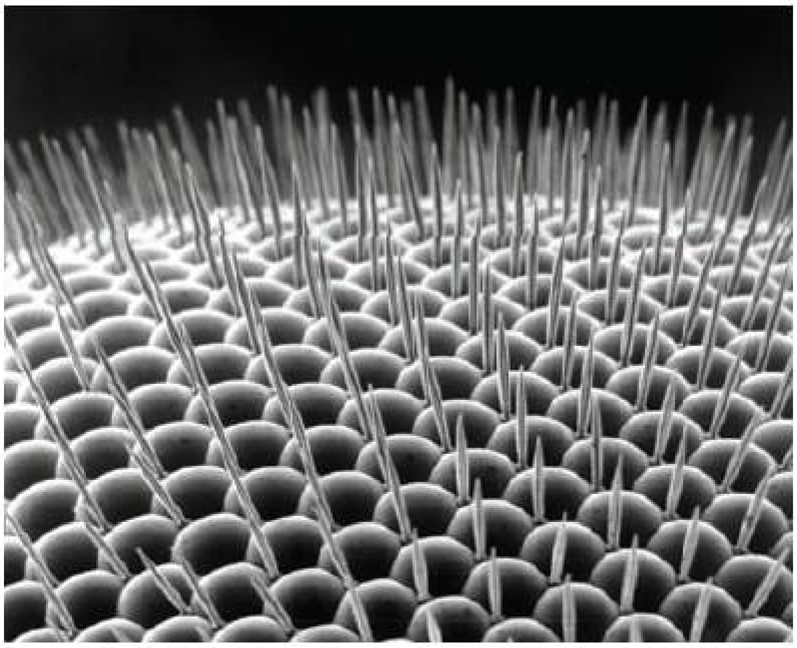


Figure 11.34 Image of an insect eye taken with an electron microscope

If a large enough voltage is used, electrons can be produced which have wavelengths as small as 5×10^{-12} m. This is about 10^5 smaller than visible light waves and makes electron waves ideal for examining very small objects with high magnification and resolution.

The electron waves are detected by a fluorescent screen or sensors for conversion to visible images. The size of atoms is much smaller than the wavelength of visible light, so it will never be possible to see an atom with an optical microscope, but blurry images of individual atoms can be obtained with electron microscopes.

Electronic displays

The number of picture elements **(pixels)** on an electronic display is usually given in a form such as 1366 × 768, meaning that there are 1366 pixels in each horizontal row and 768 in each vertical row, making a total of 1 049 088 pixels, or approximately one megapixel (1 MPx). If the screen size is 41 cm by 23 cm, then the centres of pixels are an average distance of 0.03 cm apart in their rows and columns. The spaces between the pixels are much smaller. (Confusingly, the number of pixels is often called the 'resolution' of the display.) Pixels, digital cameras and the storage of digital data are discussed in more detail in Chapter 14.

When we look at a screen, we do not want to see individual pixels, so the angular separation as seen by a viewer must be smaller than $1.22\lambda/b$. Using 4.0×10^{-7} m as a value for the lowest wavelength produced by the screen, we can calculate a rough guide to the minimum distance at which a viewer (with a pupil of size of 2 mm) would need to be positioned so that they could not resolve individual pixels:

$$\frac{(1.22 \times 4.0 \times 10^{-7})}{(2 \times 10^{-3})} = \frac{\text{pixel separation}}{\text{minimum distance between viewer and screen}}$$

Using a pixel separation of 0.03 mm gives a minimum distance of about 10 cm.

Digital cameras

The quality of the lens and the sizes of the lens and aperture are the most important factors in determining the quality of the images produced by a camera. But, the distance between the pixels on the image sensor must be small enough to resolve the detail provided by the lens. The 'resolutions' of digital cameras are also described in terms of the numbers of (light-receiving) pixels. The settings used on the camera, the way in which the data is processed and the 'resolution' of the display used will all affect the amount of detail resolvable in the final image.

Optical storage of digital information

Information is stored using tiny bumps (called 'lands') on a plastic disc (CD or DVD). The information is 'read' by light from a laser that reflects off the reflective coating on the lands and/or the 'pits' between them.

The closer the lands (and pits) are located to each other, the greater the amount of information that can be stored on the disc. But, if the lands are *too* close together, the diffracting laser beam will not be able to resolve the difference between them. A laser with a shorter wavelength will diffract less and enable more data to be stored on the same sized disc.

On a typical CD (which stores 700MB of information), the pits are about 8×10^{-6} m in length and they are read by light from a laser of wavelength 7.8×10^{-7} m. DVDs store information in the same way, but they have an improved data capacity; the pits and lands can be closer together because a laser light of shorter wavelength (6.5×10^{-7} m) is used. A single-layer DVD can store about seven times as much information as a CD.

The development of blue lasers, with their smaller wavelengths (4.05×10^{-7} m), has led to Blu-ray technology and greatly increased data storage capacity on discs of the same size. This enables the storage and playback of high definition (HD) video.

The storage of digital information is discussed at length in Chapter 14.

11.4.4 Solve problems involving resolution.

23 **a** Why might you expect a camera with a lens with a larger diameter to produce better pictures?
b Suggest a reason why bigger lenses might produce poorer images.

24 Why do astronomers sometimes prefer to take photographs with blue filters?

25 The pupils in our eyes dilate (get bigger) when the light intensity decreases. Discuss whether this means that people can see better at night.

26 A car is driving towards an observer from a long way away at night. If the headlights are 1.8 m apart, estimate the maximum distance away at which the observer will see two distinct headlights. (Assume that the average wavelength of light is 5.2×10^{-7} m and the pupil diameter is 4.2 mm.)

27 A camera on a satellite orbiting at a height of 230 km above the Earth is required to take photographs which resolve objects that are 1.0 m apart. Assuming a wavelength of 5.5×10^{-7} m, what minimum diameter lens would be needed?

28 **a** A radio telescope with a dish of diameter 64 m is used to detect radiation of wavelength 1.4 m. How far away from the Earth are two stars separated by a distance of 3.8×10^{12} m which can just be resolved?
b What assumption(s) did you make?

29 Could an optical telescope with a (objective) lens of diameter 12 cm be used to read the writing on an advertising sign which is 5.4 km away if the letters are an average of 8.5 cm apart? Explain your answer.

30 The Moon is 3.8×10^{8} m from the Earth. Estimate the smallest distance between two features on the Moon's surface that can just be observed from the Earth by the human eye.

31 Use Figure 11.35 to measure the resolution of your own eyes.

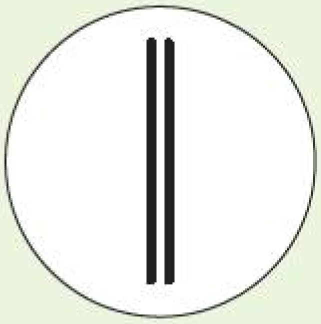


Figure 11.35

TOK Link: Differences between reality and our observations

Figure 11.36 shows a large number of individual pixels, and it may be difficult to decide if this is picture of anything or just dots. But, once we know what the picture is, it becomes instantly recognizable. This is a simple illustration of the fact that, when our brains interpret signals from our eyes (or ears, or other senses), how we interpret what we sense has a lot to do with our past experiences and our expectations.

Optical illusions like that shown in Figure 11.37 trick us because they are unusual experiences and our brain tries hard to make sense of them by interpreting them in terms of more familiar (probably three-dimensional) images. You tend to see what your brain thinks it should be seeing, rather that what is really there.

The expression 'seeing is believing' is far from true, because we cannot trust any of our senses to give us an objective interpretation of the world around us. A person sitting for a long time in a room at 16°C would probably describe the room as cold, but someone coming from outside, where the temperature was 10°C, would certainly think that the room was warm. If we listen to someone speaking a foreign language we may hear no sound that we recognize, but we would immediately recognize our name if it was mentioned.

Of course, this inherent uncertainty in our observations is a very good reason why modern science is based on quantitative measurements, which are much less open to misinterpretation. The earliest scientists (natural philosophers) were restricted by the limitations of qualitative observations because they did not do practical work.

Figure 11.36 A large number of individual dots can be used to create a picture

Figure 11.37 An optical illusion

Question

1 'If a tree falls in a forest and no one is around to hear it, does it make a sound?' is an old philosophical question which raises many issues, including the distinction between what something is and what it appears to be. In what ways do a tree and the sound that it makes exist, if nobody is there to observe them? Should we believe in things of which we have no direct experience? If we believe that a tree has actually fallen and disturbed the surrounding air, but sound is only the name we give to a sensation caused at the human ear, is there any sound in the forest?

11.5 Polarization

Polarized waves

Basic wave properties were covered in Chapter 4, but one interesting property of (transverse) waves was not included – **polarization**. As a simple example, consider sending transverse waves along a rope. If your hand only oscillates vertically, the rope will only oscillate vertically and the wave can be described as **plane polarized** because it is only oscillating in one plane (vertical), as shown in Figure 11.38. If your hand only oscillates horizontally it will produce a wave polarized in the horizontal plane.

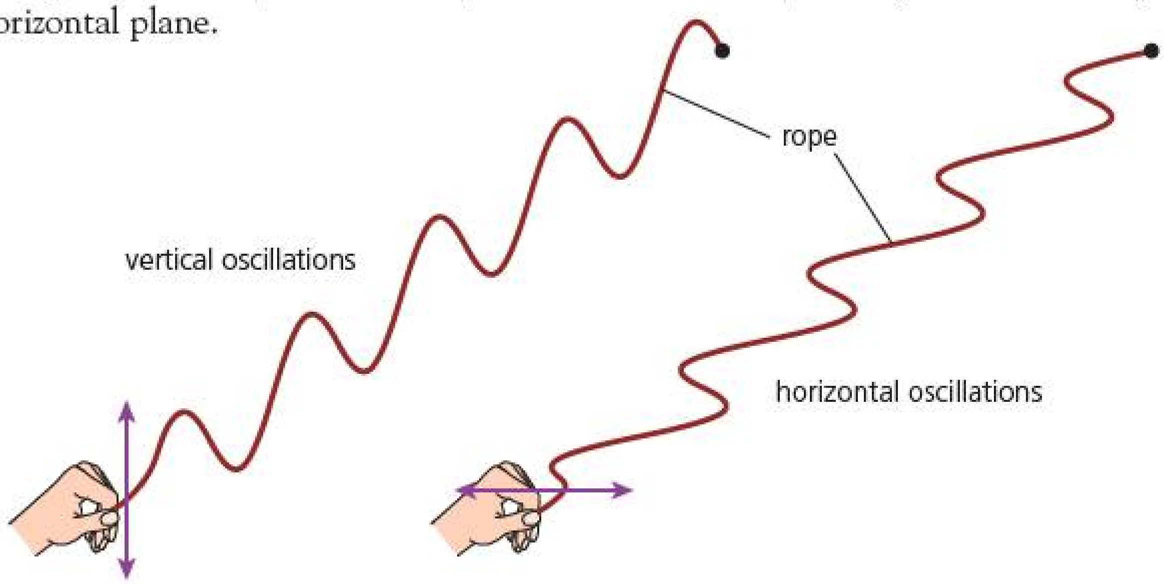


Figure 11.38 Transverse waves on a rope

A transverse wave is (plane) polarized if all the oscillations transferring the wave's energy are in the same plane (called the **plane of polarization**). The plane of polarization must be perpendicular to the direction of wave travel, so it is impossible for longitudinal waves, like sound, to be polarized.

Polarized light and other electromagnetic waves

11.5.1 Describe what is meant by polarized light.

In order to explain the polarization of light, we first need to understand the basic nature of electromagnetic waves. All electromagnetic radiations, including light, are transverse waves of oscillating electric and magnetic fields at right angles to each other. This is shown in Figure 11.39.

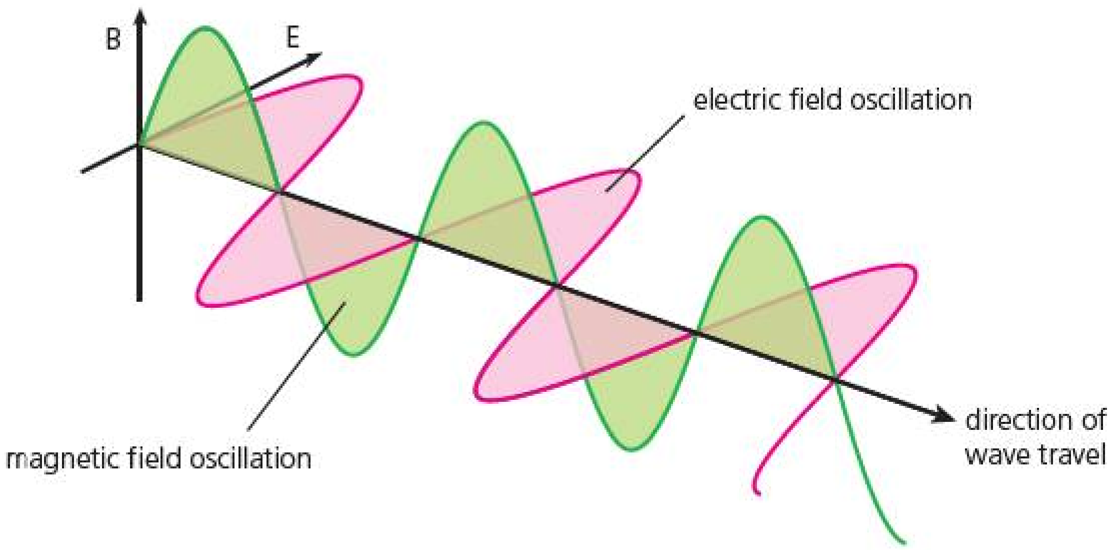


Figure 11.39 Nature of electromagnetic waves

In Figure 11.39 the waves are moving to the right, the electric field oscillations are in the horizontal plane and the magnetic field oscillations are in the vertical plane, but the oscillations could be in any plane which is perpendicular to the direction of wave travel (as long as the electric fields and magnetic fields are perpendicular to each other).

Electromagnetic waves, including light, are usually released during random, unpredictable processes, so we would expect them usually to oscillate in random directions and not be polarized, as shown in Figure 11.40a.

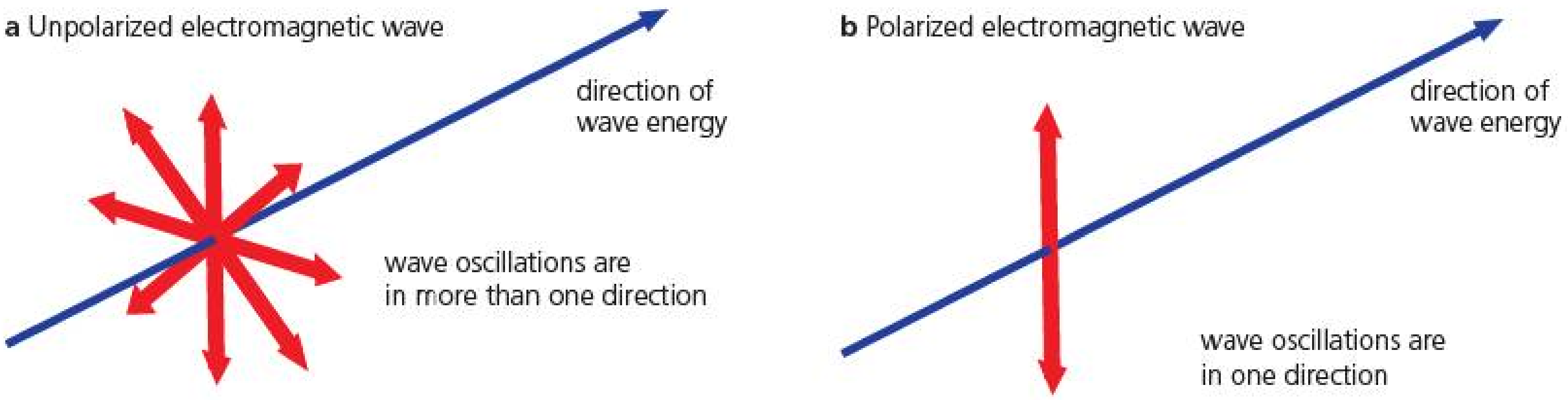


Figure 11.40 **a** unpolarized electromagnetic radiation **b** polarized electromagnetic radiation

Electromagnetic waves, such as light, are said to be (plane) polarized if all the electric field oscillations (or the magnetic field oscillations) are only in one plane, as shown by Figure 11.40b.

Electromagnetic waves which are produced and transmitted by currents oscillating in aerials (for example radio waves and microwaves) will be polarized, with their electric field oscillations parallel to the transmitting aerial.

For example, Figure 11.41 shows the transmission and reception of microwaves. In Figure 11.41a a strong signal is received because the transmitting and receiving aerials are aligned, but in Figure 11.41b no signal is received because the receiving aerial has been rotated through 90°.

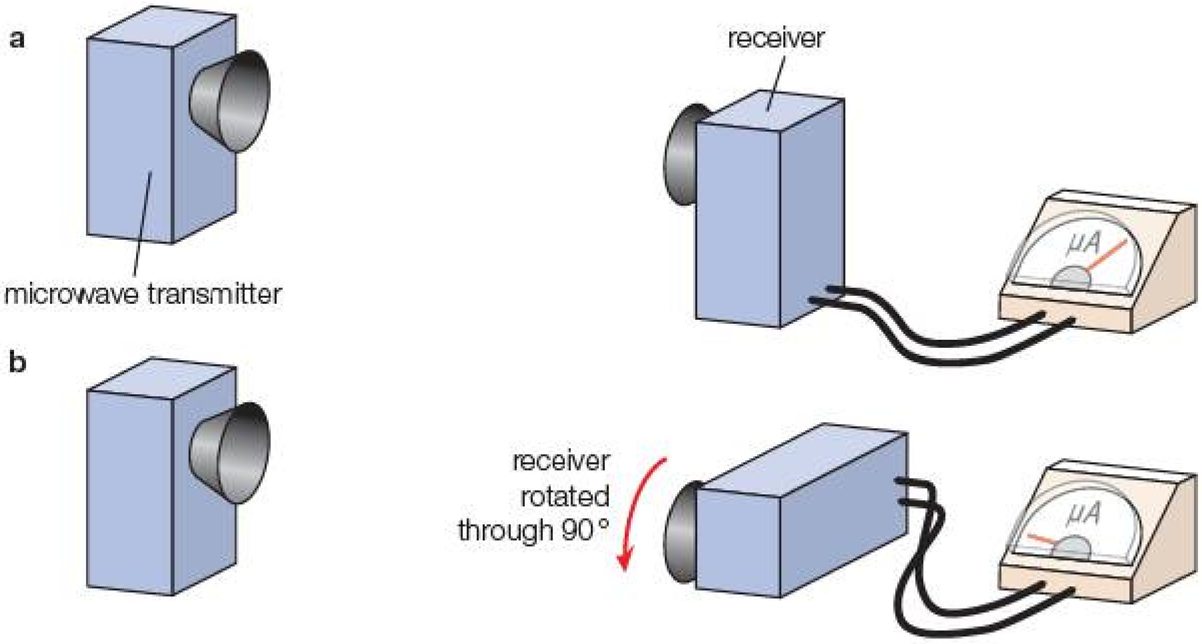


Figure 11.41 The receiver must be aligned in the same plane as the transmitter to detect microwaves

Polarization by absorption

11.5.4 Explain the terms polarizer and analyser.

Normal, unpolarized light can be converted into polarized light by passing it through a special filter, called a **polarizer**, which absorbs oscillations in all planes except one.

Polarizing filters are made of long chain molecules mostly aligned in one direction. Components of the electric field parallel to the long molecules are absorbed; components of the electric field perpendicular to the molecules are transmitted. Because of this, we would expect the transmitted intensity to be about half of the incident intensity, as shown in Figure 11.42.

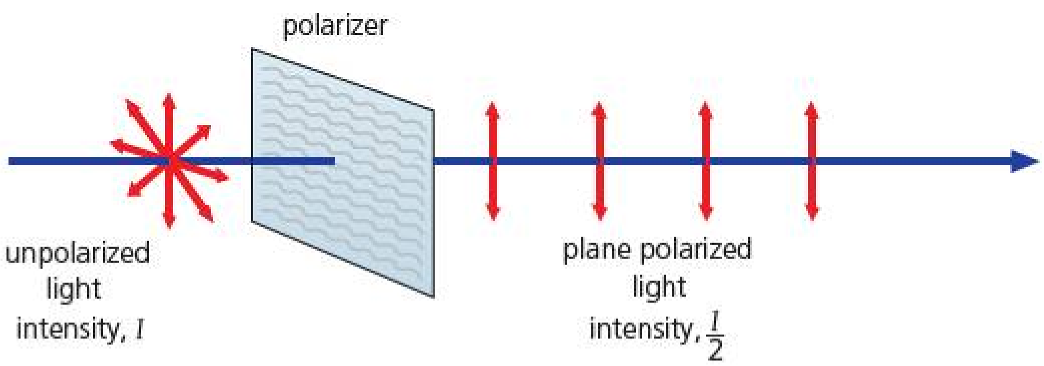


Figure 11.42 Polarizing light using a polarizer (polarizing filter)

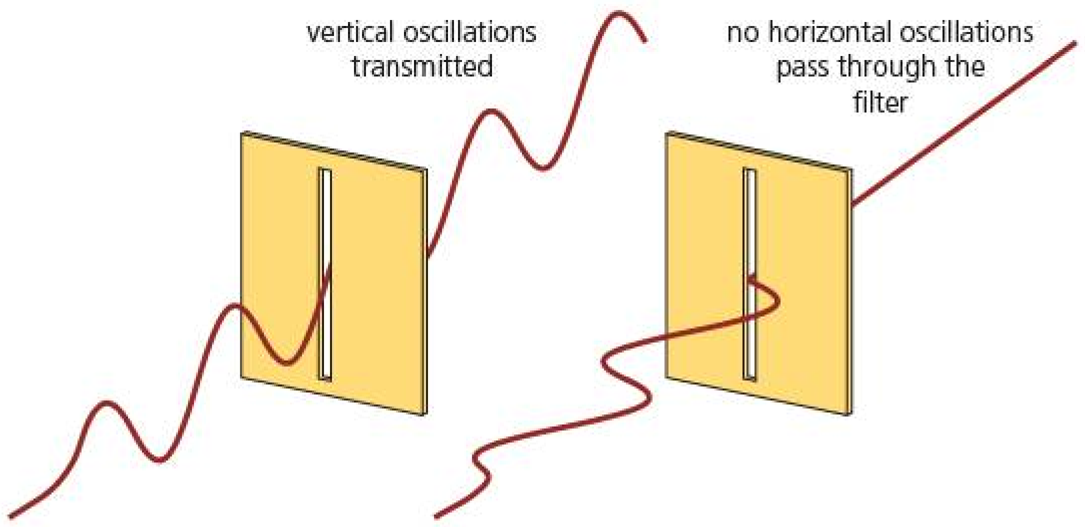


Figure 11.43 A vertical slit acts as a polarizing filter

In order to understand this, it may be helpful to consider how transverse waves made by oscillating a rope (as shown in Figure 11.38) would behave if they had to pass through a vertical slit; see Figure 11.43. Waves oscillating parallel to the slit would pass through, but others would be blocked. The slit acts as a polarizer.

What happens if polarized light is then incident on a second polarizing filter? This depends on how the filters are aligned. If the second filter (called the **analyser**) transmits waves in the same plane as the first (the polarizer), the waves will pass through unaffected, apart from a possible small decrease in intensity. The filters are said to be **in parallel**, as shown in Figure 11.44b. Figure 11.44a shows the situation in which the analyser only allows waves through in a plane which is at right angles to the plane transmitted by the polarizer. The filters are said to be '**crossed**' and no light will be transmitted by the analyser.

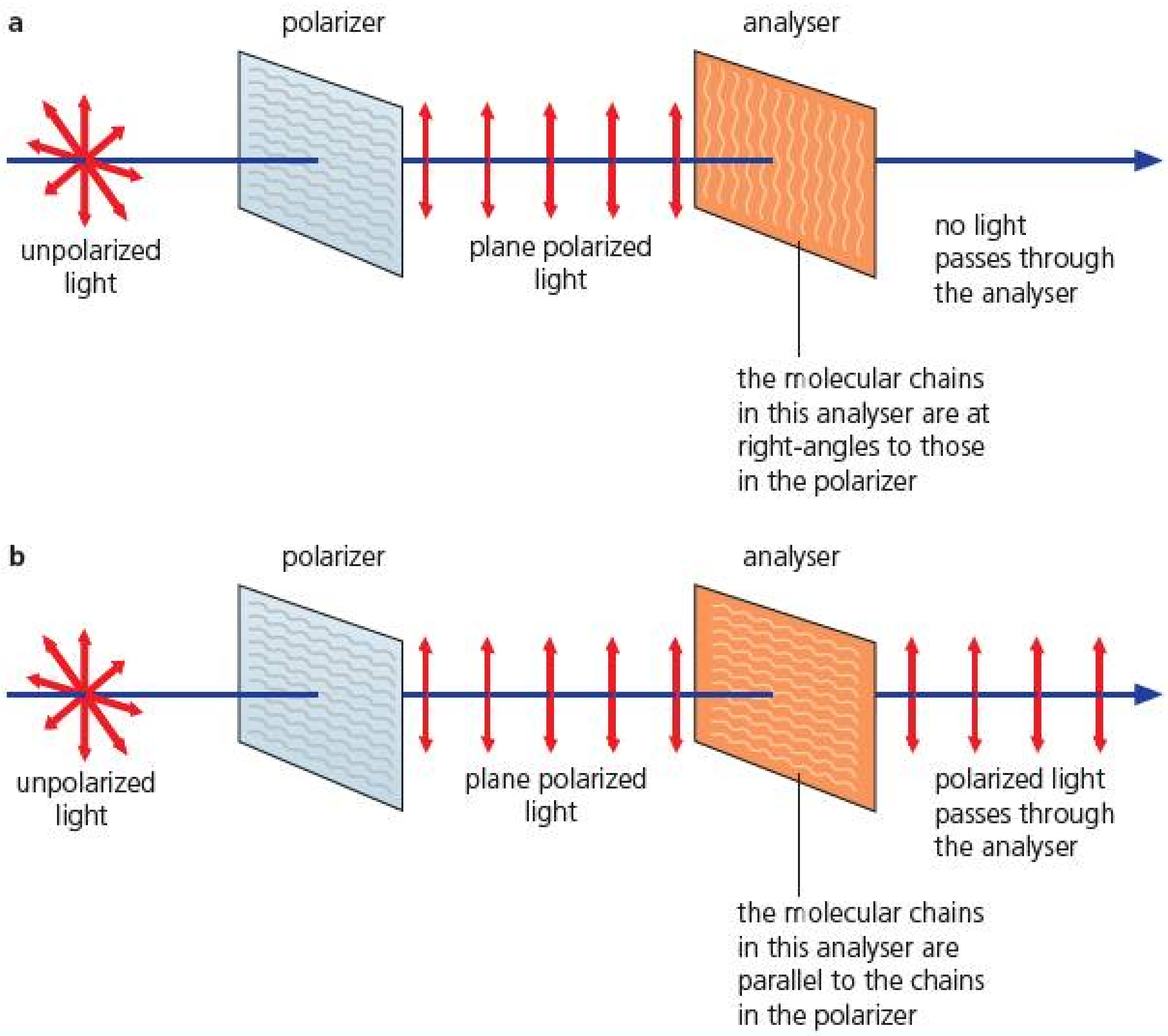


Figure 11.44 Polarizer and analyser: **a** crossed; **b** in parallel

The second filter is called an *analyser* because it can be rotated to analyse light to determine if it is polarized and, if so, in which direction.

If you rotate a polarizing filter in front of your eye, for example, as you look around at the light reflected from various objects, if the intensity changes then the light must be at least partly polarized. The most common type of transparent plastic used for polarizers and analysers is called **Polaroid®**.

Malus' law

11.5.5 Calculate the intensity of a transmitted beam of polarized light using Malus' law.

Figure 11.45 represents a direction (in the plane of the paper) for the electric field oscillations of polarized light which is travelling perpendicularly out of the page. There is an angle θ between the oscillations and the plane in which the polarizer (through which the light is passing) will transmit all of the waves.

If the amplitude of the oscillations incident on the analyser is A_0, then the component in the direction in which waves can be transmitted equals $A_0 \cos\theta$.

We know, from Chapter 4, that the intensity of waves is proportional to amplitude squared, $I \propto A_0^2$, so that the transmitted intensity is represented by the following equation:

$$I = I_0 \cos^2\theta$$

This equation is given in the IB *Physics data booklet*.

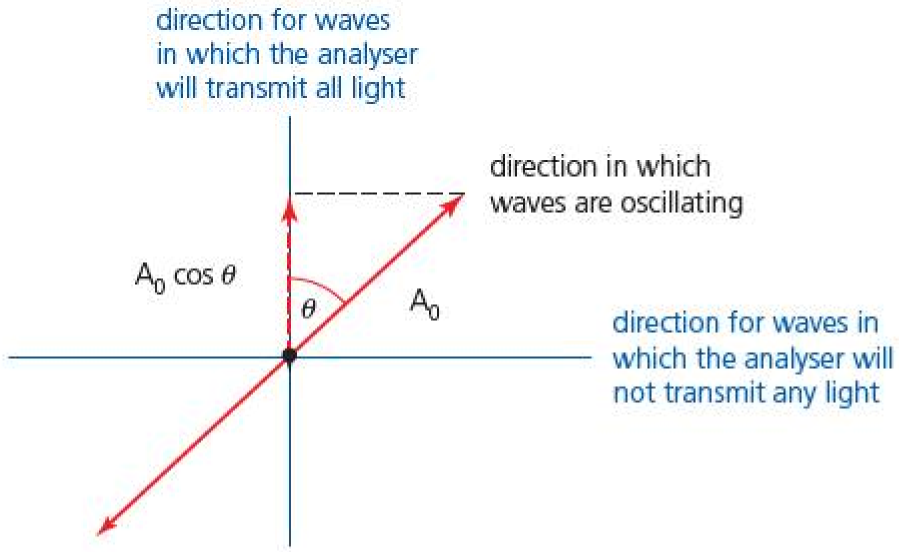


Figure 11.45 Angle between oscillations and the polarizer

Worked example

9 a If vertically polarized light falls on a polarizing filter (analyser) which is positioned so that its transmission direction (axis) is at 30° to the vertical, what percentage of light passes through the analyser?

b Repeat the calculation for an angle of 60°.

c What angle would allow 50% of the light to pass?

d Sketch a graph to show how the transmitted intensity would vary if the analyser was rotated through 360°.

a $I = I_0 \cos^2\theta$

$$\frac{I}{I_0} = \cos^2 30°$$

$$\frac{I}{I_0} = 0.75 \text{ or } 75\%$$

b $\dfrac{I}{I_0} = \cos^2 60° = 0.25 \text{ or } 25\%$

c $\frac{I}{I_0} = 0.50 = \cos^2\theta$

$\cos\theta = \sqrt{0.50} = 0.71$

$\theta = 45°$

d See Figure 11.46. This graph shows the variation with angle of intensity transmitted through a polarizing filter.

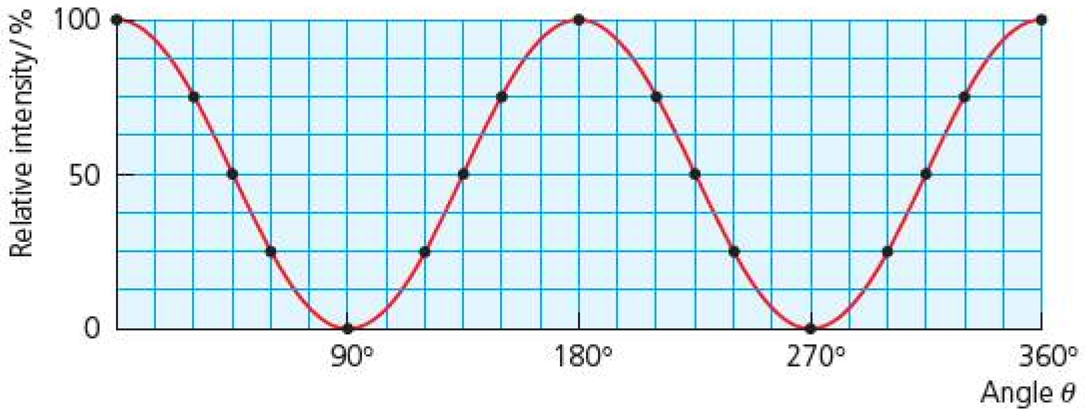


Figure 11.46

Polarization by reflection

11.5.2 Describe polarization by reflection.
11.5.3 State and **apply** Brewster's law.

When (unpolarized) light reflects off an insulator the waves may become polarized and the plane of polarization will be parallel to the reflecting surface.

Figure 11.47 The same scene with and without Polaroid®

The most common examples of polarization by reflection are the reflection of light from water and from glass. Such reflections are usually unwanted and the reflected light (sometimes called **glare**) entering the eyes can be reduced by wearing **polarizing (Polaroid®) sunglasses**, which also reduce the intensity of unpolarized light by half. Figure 11.47 shows an example. The fish under the water can be seen clearly when Polaroid sunglasses are used. The sunglasses greatly reduce the amount of light reflected off the water's surface entering the eye. But the reduction is not the same for all viewing angles because the amount of polarization depends on the angle of incidence (see below). Photographers may place a rotatable polarizing filter over the lens of their camera to reduce the intensity of reflected light.

For transparent materials (for example glass or water), at a particular angle, called the **Brewster angle**, all of the reflected light is polarized. This occurs when the refracted rays and the reflected rays are exactly at 90° to each other, as shown in Figure 11.48. (Remember that rays are lines showing the direction in which waves are travelling.) At the Brewster angle, incident light which is polarized in the right direction will be totally transmitted, without any reflection.

From Chapter 4, we know that Snell's law links refractive indices to angles of incidence and refraction as follows:

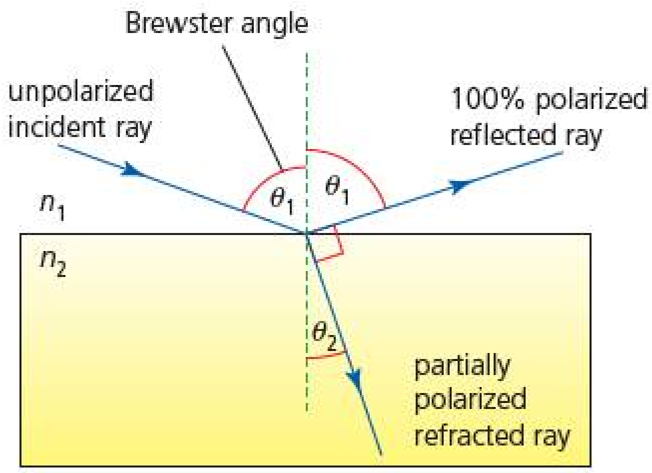


Figure 11.48 Brewster angle

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1}$$

In Figure 11.48 the angle of incidence is θ_1 in a material of refractive index n_1. The wave is then refracted at an angle θ_2 in a material of refractive index n_2.

Looking at the angles in Figure 11.48 it is clear that:

$$\theta_1 + \theta_2 + 90° = 180°$$

So that,

$$\theta_2 = 90° - \theta_1$$

and

$$\sin\theta_2 = \cos\theta_1$$

Substituting this back into the Snell's law equation, we get:

$$\frac{n_2}{n_1} = \frac{\sin\theta_1}{\cos\theta_1} = \tan\theta_1$$

Usually material 1 will be air, so that $n_1 = 1$. Put into words, this tells us that the tangent of the Brewster angle, ϕ $(= \theta_1)$, is equal to the refractive index, n, of the material that the light is entering.

This is known as **Brewster's law**. It is given in the IB *Physics data booklet* in the form:

$$n = \tan\phi$$

Worked example

10 Calculate the Brewster angle for light passing from air into water (refractive index = 1.33).

$$\tan\phi = n = 1.33$$
$$\phi = 53°$$

11.5.10 Solve problems involving the polarization of light.

32 How could you quickly check if sunglasses were made with Polaroid®?

33 Plane polarized light passes through an analyser which has its transmission axis at 75° to the plane of polarization. What percentage of the incident light emerges?

34 When unpolarized light was passed through two polarizing filters only 20% of the incident light emerged. What was the angle between the transmission axes of the two filters?

35 Suggest why the blue light from the sky on a clear, sunny day is partly polarized.

36 **a** What is the Brewster angle for light passing into glass of refractive index 1.53?
b At what angle will light which is polarized parallel to the surface of water of refractive index 1.33 be totally refracted?
c What is the angle of refraction?

37 Use the Internet to find out about polarizing microscopes and their uses.

Applications of polarization

As we have seen, if two polarizing filters are 'crossed', the polarized light which passes through the polarizer cannot pass through the analyser, so no light is transmitted. However, if a transparent material is placed between the two filters it may *rotate the plane of polarization*, allowing some light to be transmitted through the analyser. Figure 11.49 shows a solution of sugar being used between the polarizing filters. This effect has some interesting applications.

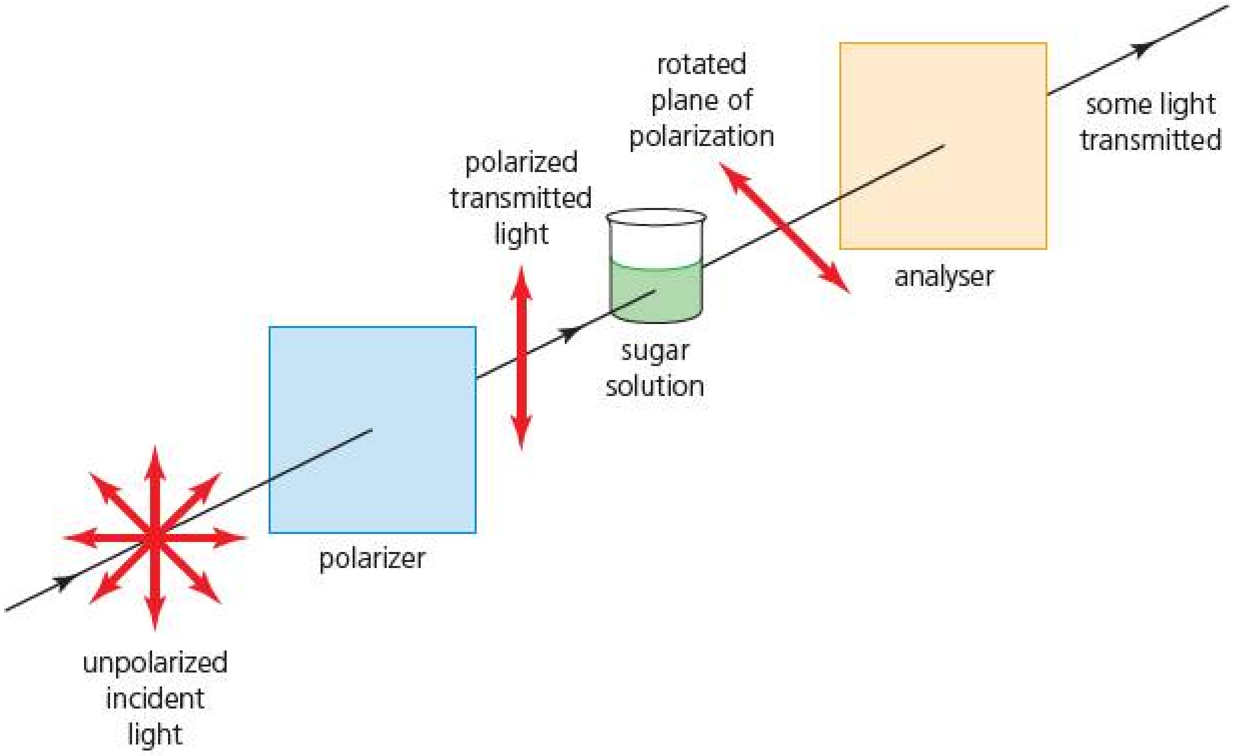


Figure 11.49 Rotating the plane of polarization by placing a transparent material, in this case sugar solution, between the polarizer and the crossed analyser, allowing some light to be transmitted through

Determining the concentration of solutions

11.5.6 Describe what is meant by an optically active substance.
11.5.7 Describe the use of polarization in the determination of the concentration of certain solutions.

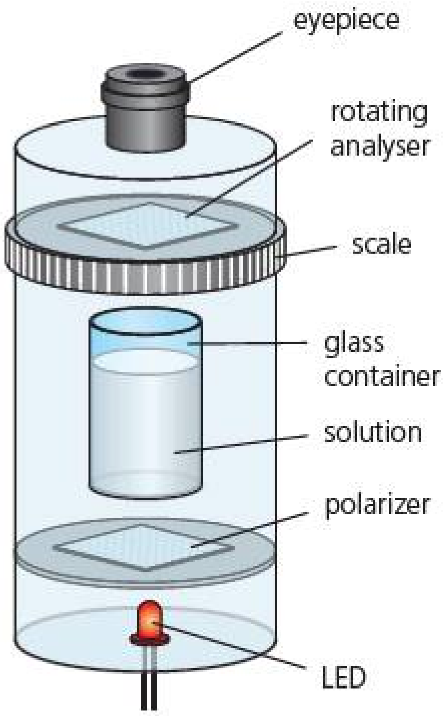


Figure 11.50 Measuring the concentration of a solution with a polarimeter

A substance which rotates the plane of polarization of light waves passing through it is said to be **optically active**.

The angle through which the plane is rotated depends on the substance, the length of the path through the substance, and the wavelength used (and the temperature of the solution). Solutions of some pure compounds are optically active and sugar solutions are a common example (as shown in Figure 11.49). The greater the concentration of the solution, the more the plane of polarization is rotated, so this can be used as a way of experimentally determining the concentration of certain solutions, as in the simple **polarimeter** shown in Figure 11.50.

Using the eyepiece in a polarimeter, the intensity of the light received from the LED at the bottom of the apparatus is observed as the analyser is rotated, and the angle at which no light is received is measured and compared to the result without the solution.

Stress analysis

11.5.8 Outline qualitatively how polarization may be used in stress analysis.

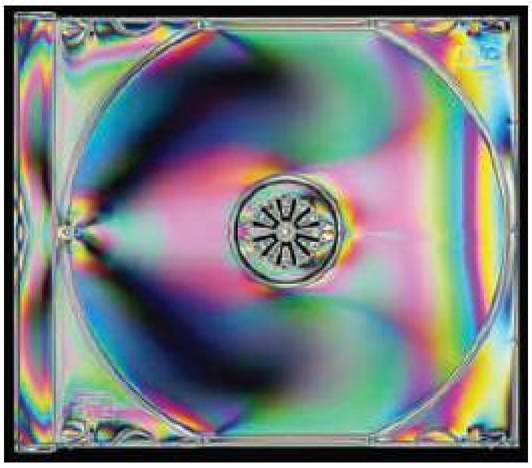


Figure 11.51 Stress concentration in a DVD case seen with polarized light

Some plastics and glasses become optically active and rotate the plane of polarization of light when they are stressed. This can be useful for engineers who can analyse possible concentrations of stress in a structure before it is made.

By building a (scale) model in a transparent material and placing it between polarizing filters, engineers can observe its behaviour under the action of various forces. An example of stress concentration is shown in Figure 11.51. The greater the stress, the greater is the rotation of the plane of polarization. Different colours are seen if white light is used because the effect depends on the wavelength of the light used.

Liquid-crystal displays (LCDs)

11.5.9 Outline qualitatively the action of liquid-crystal displays (LCDs).

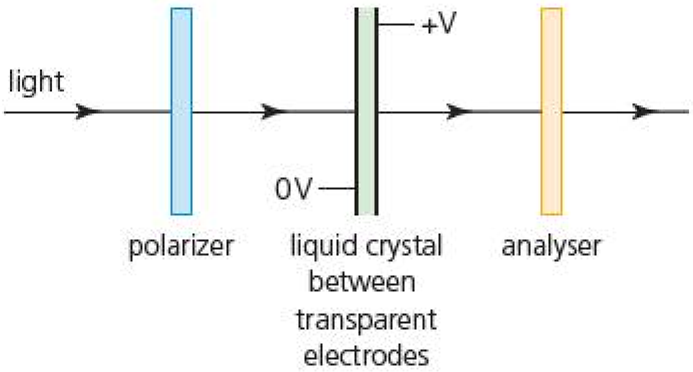


Figure 11.52 Arrangement of parts in a liquid-crystal display

A liquid crystal is a state of matter that has physical properties between those of a liquid and a solid (crystal). Most interestingly, the ability of certain kinds of liquid crystal to rotate the plane of polarization of light can be changed by applying a small potential difference across them, so that their molecules twist in the electric field. Figure 11.52 represents a simplified arrangement. If there is no potential difference (p.d.) across the liquid crystal no light is transmitted out of the analyser.

When a p.d. is applied to the liquid crystal, its molecules change orientation to align with the electric field and the plane of polarization rotates so that some, or all, of the light is now transmitted. The amount of rotation of the plane of polarization and the amount of light transmitted depend on the size of the p.d.

In simple displays (for example those used on many calculators, digital clocks and watches) light entering through the front of the display passes through the liquid crystals and is then reflected back to the viewer. Each segment of the display will then appear dark or light, depending on whether a p.d. has been applied to the liquid crystal (see Figure 11.53).

Using the same principles, large numbers of liquid crystals are used to make the tiny picture elements (pixels) in many computer and mobile phone displays, and televisions. Colours are created by using filters and the light is provided by a fluorescent lamp behind the display.

Figure 11.53 A seven segment liquid-crystal display

11.5.10 Solve problems involving the polarization of light.

38 Light of intensity $0.82\,\text{W m}^{-2}$ is incident on two crossed polarizing filters.
- **a** What is the intensity entering the second filter?
- **b** What intensity emerges from the second filter?
- **c** If a sugar solution placed between the filters rotates the plane of polarization by 28°, what intensity then emerges from the second filter?

39 Use the Internet to research and compare the main advantages and disadvantages of using liquid crystal and LED displays.

■ Additional Perspectives

3-D cinema

Our eyes and brain see objects in three dimensions (3-D) because, when our two eyes look at the same object, they each receive a slightly different image (because our eyes are not located at exactly the same place). This is known as stereoscopic vision. Our brain merges the two images to give the impression of three dimensions or 'depth'. But when we look at a two-dimensional image in a book or on a screen, both eyes receive essentially the same image.

If we want to create a 3-D image from a flat screen, we need to provide a different image for each eye and the use of polarized light makes this possible. (Some earlier, and less effective systems used different coloured filters.) In the simplest modern systems one camera is used to take images which are projected in the cinema as vertically polarized waves. At the same time a second camera, located close by, takes images that are later projected as horizontally polarized waves. Sometimes a second image can be generated by a computer program (rather than by a second camera) to give a 3-D effect.

Figure 11.54 Using polarizing glasses to watch a 3-D movie

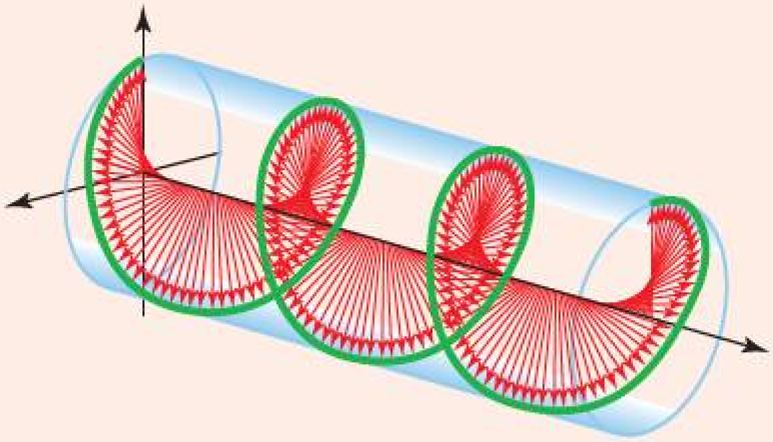


Figure 11.55 Circularly polarized waves

To make sure that each eye receives a different set of images, the viewer wears polarizing glasses, allowing the vertically polarized light into one eye and the horizontally polarized light into the other eye. One problem with using plane polarized waves is that viewers need to keep their heads level, but this can be overcome by using circularly polarized waves in which the direction of the electric field oscillations continually rotates in circles, as in Figure 11.55. A single projector sends images alternating between clockwise and anti-clockwise polarization to the screen.

Question

1 Use the Internet to investigate the latest developments in 3-D television techniques.

SUMMARY OF KNOWLEDGE

11.1 Standing (stationary) waves

- Standing waves may be set up when waves of the same frequency and amplitude, travelling in opposite directions, combine together. The resultant at any place can be found (at any time) by using the principle of superposition.
- Travelling (progressive) waves move away from their source and transfer energy, but a standing wave pattern stays in the same position and energy is not transferred. There are places (called nodes) along a standing wave where the amplitude is always zero. Between nodes the amplitude of the oscillations varies, but all the oscillations are in phase. Places of maximum amplitude are called antinodes. The wavelength is twice the distance between adjacent nodes (or antinodes).
- It is usually possible for a particular system to vibrate with standing waves of different frequencies: these are called different modes of vibration. The lowest frequency mode is called the fundamental mode, or the first harmonic. The frequency of the fundamental (and other modes) will be dependent on the speed of the wave, the length of the system and whether the ends are fixed ($v = f\lambda$).
- There are various modes of vibration for standing transverse waves on strings and longitudinal sound waves in pipes (open and closed).
- Standing waves are often created by an external source of continuous vibration. If the external frequency is the same as a natural frequency of the system, resonance will occur and the amplitude of the standing wave may increase.

11.2 The Doppler effect

- When there is relative motion between the source and the receiver of a wave, the frequency received will not be the same as the frequency emitted. This is known as the Doppler effect. The most common example is the sound heard from a moving vehicle.
- The Doppler effect can be explained by drawing the wavefronts from a moving source.
- The IB *Physics data booklet* provides the equations for the frequency received from a moving source and by a moving observer.
- For electromagnetic waves the change in frequency caused by the Doppler effect can be calculated from $\Delta f = (v/c)f$. This is an approximation which can be used only if the relative speed between source and detector is much less than the speed of electromagnetic radiation (which is true for most calculations).
- The Doppler effect can be used to determine speeds: if waves of a known speed and frequency are directed at a moving object, the change of frequency of the reflected waves can be used to calculate the speed of the object. This is used with ultrasonic waves to measure blood flow rates and with microwaves to measure plane and car speeds.

11.3 Diffraction

- Because light has a very small wavelength, the diffraction of light only becomes important for small apertures. It is most easily understood by considering the diffraction of monochromatic light passing through a narrow slit. It is possible, by considering the path differences, to derive the formula $\theta = \lambda/b$ for the position of the first minimum of the diffraction pattern produced at a single slit.
- Graphs can be drawn to show the variation of relative intensity with angle for light diffracted by a single slit.

11.4 Resolution

- The angular separation of two objects is equal to the angle they subtend at the observer.
- If we can see two objects that have a small angular separation as being separate, we say that they can be resolved. The ability of our eyes to resolve detail depends on our eyesight (obviously) and the diffraction of light entering the eyes.
- Consider the simplest example: if we look at two close and identical point sources of monochromatic light, then our eyes will receive two identical, overlapping diffraction patterns. Rayleigh's criterion states that the two images are just resolvable if the first diffraction minimum of one occurs at the same angle as the central maximum of the other.
- For light passing through a narrow slit, we can say that the images of two sources can be resolved if they have an angular separation of $\theta = \lambda/b$ or greater. For a circular aperture the resolution is poorer and the criterion becomes $\theta = 1.22(\lambda/b)$.
- Diffraction patterns of images can be sketched and interpreted to show images that are well resolved, just resolved and not resolved.
- To see objects in detail with high resolution, the wavelength used should be as small as possible and the aperture as large as possible. The resolution of optical instruments like telescopes, microscopes and cameras is limited by the wavelength of light, but they can be improved by using larger apertures. Radio telescopes need to be large because of the size of the radio waves that they detect. Electron microscopes achieve high resolution because of the small wavelength of electrons.
- Data is stored on CDs and DVDs using lands and pits on the discs. The closer together the lands and pits, the more data that can be stored on a given disc, but their separation is limited by the diffraction of the laser light used to read the data.

11.5 Polarization

- A transverse wave is (plane) polarized if the all the oscillations transferring the wave's energy are in the same plane. Longitudinal waves cannot be polarized.
- Electromagnetic radiations are transverse waves of oscillating electric and magnetic fields at right angles to each other.
- Artificial radio waves and microwaves are polarized because of the controlled way in which they are produced. Other electromagnetic radiations, including light, are usually unpolarized.
- Light can be polarized by passing it through a polarizing filter (called a polarizer) and its intensity will then be halved. If the polarized light is then passed through a second polarizing filter (called an analyser), the intensity of the transmitted light will be zero if the filters are 'crossed', and remains unaltered if the filters are aligned. Malus' law ($I = I_0 \cos^2\theta$) can be used to calculate the transmitted intensity. A sketch graph provides a simple way to show how intensity varies with angle.
- An optically active substance rotates the plane of polarization of light passing through it. This effect can be used in stress analysis and in the determination of the concentration of certain solutions.
- Light which is reflected from the surface of an insulator (like glass or water) may become polarized. This can be checked by observing the reflection through a rotating polarizing filter. Polarized sunglasses are used to reduce the amount of reflected light entering the eyes.

- For light entering transparent materials at a particular angle, called the Brewster angle, ϕ, all of the reflected light is polarized. This occurs when the refracted rays and the reflected rays are at 90° to each other. Brewster's law relates this angle to the refractive index, n, of the material: $n = \tan \phi$.
- The plane of polarization of a liquid crystal can be rotated by applying a p.d. across it. A liquid-crystal display (LCD) contains individual pixels, each of which is connected to a separate p.d., so that the amount of light passing through it can be controlled.

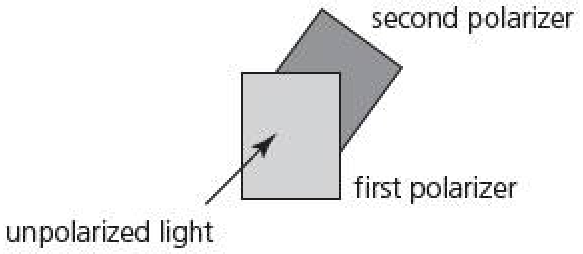
Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 A moving source of sound passes a stationary observer with a speed v. If the speed of sound in air is c, what will be the speed of the waves detected by the observer?

A c
B $c - v$
C $c + v$
D $v - c$

Q2 Unpolarized light of intensity I_0 is incident on a polarizer. The transmitted light is then incident on a second polarizer. The axis of the second polarizer makes an angle of 60° to the axis of the first polarizer.



The cosine of 60° is $\frac{1}{2}$. The intensity of the light transmitted through the second polarizer is:

A I_0 **B** $\frac{I_0}{2}$
C $\frac{I_0}{4}$ **D** $\frac{I_0}{8}$

Higher Level Paper 1, Specimen Paper 09, Q18

Q3 The equation $\theta = \lambda/b$ is used to describe diffraction at a single slit. In this equation

A b is the width of the fringes seen on the screen.
B b is the path difference between waves arriving at adjacent fringes.
C θ is the angular width of the central maximum.
D θ is half the angular width of the central maximum.

Q4 Which of the following describes standing waves on stretched strings?

A All points on the waves move with the same amplitude.
B All points on the waves move with the same phase.
C All points on the waves move with the same frequency.
D The wavelength is the distance between adjacent antinodes.

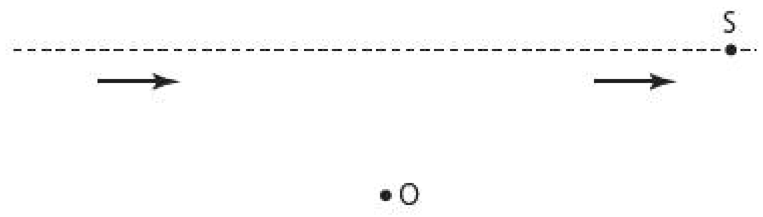
Q5 In order to store more data on an optical storage disc, it is possible to:

A use laser light of a longer wavelength.
B use laser light of a higher frequency.
C use infrared radiation instead of visible light.
D make the pits and lands on the disc larger.

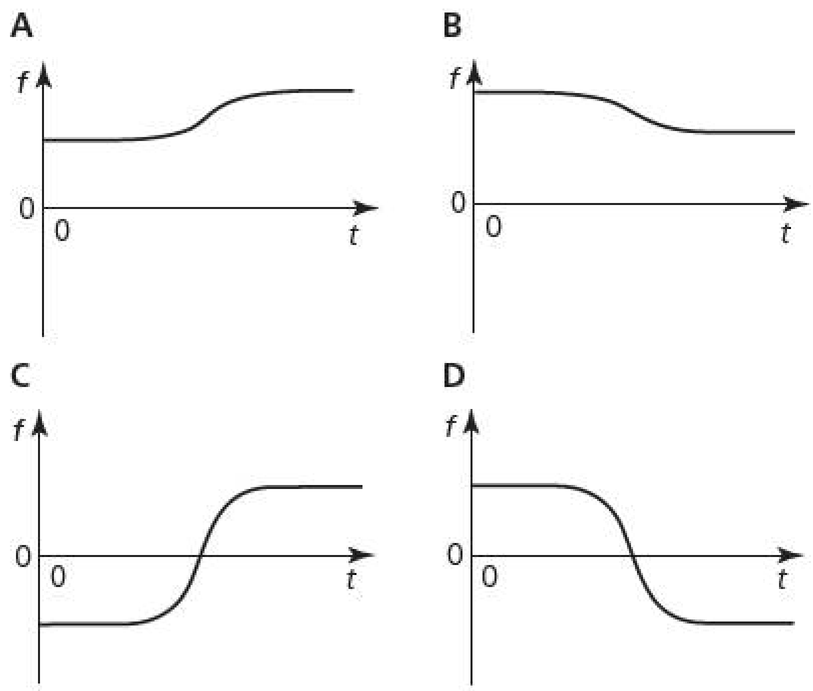
Q6 A stretched string is made to oscillate in its second harmonic. Including the ends, how many nodes, will be seen along the string?

A 1
B 2
C 3
D 4

Q7 A source S, moving at constant speed, emits a sound of constant frequency. The source passes by a stationary observer O, as shown below.



Which of the following shows the variation with time t of the frequency f observed at O as the source S approaches and passes by the observer?



Higher Level Paper 1, Specimen Paper 09, Q17

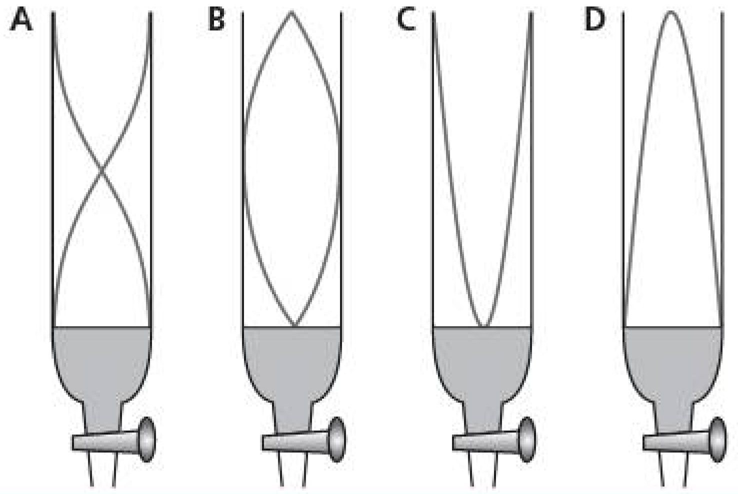
Q8 An optically active substance:

- **A** rotates the direction of transmission of energy of light.
- **B** can be used to investigate stress concentrations.
- **C** can be used to determine the wavelength of light.
- **D** can be used to determine the density of certain solutions.

Q9 Which of the following is not (on its own) a possible way to increase the resolution of an optical telescope?

- **A** increase the diameter of the lenses
- **B** use a blue filter
- **C** use more powerful lenses to produce a bigger magnification
- **D** use better quality lenses

Q10 A source of sound of a single frequency is positioned above the top of a tube which is full of water. When the water is allowed to flow out of the tube a loud sound is first heard when the water is near the bottom, as shown in all the diagrams. Which diagram best represents the standing wave pattern that produced that loud sound?



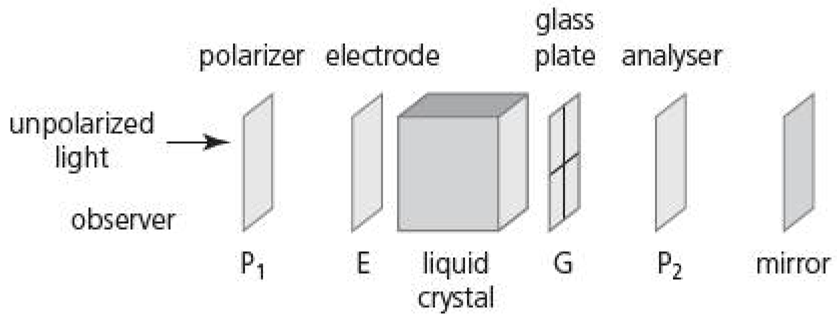
Paper 2 IB questions and IB style questions

Q1 The diagram shows a source of sound, S, which is moving in the direction shown with a constant velocity. The sound emitted by S is received at stationary points P and Q.

- **a** Make a copy of the diagram and show a series of wavefronts that have been emitted by S. [2]
- **b** Use your drawing to explain the differences in the sounds that are heard at points P and Q. [3]
- **c** **i** What is the name of this effect?
  ii Give one example of this effect in everyday life. [2]

Q2 This question is about polarization and liquid crystals.

- **a** A liquid crystal has the property of being able to rotate the plane of polarization of light. Explain what is meant by the expression 'able to rotate the plane of polarization of light'. [2]
- **b** The diagram below is a representation of a liquid-crystal display.

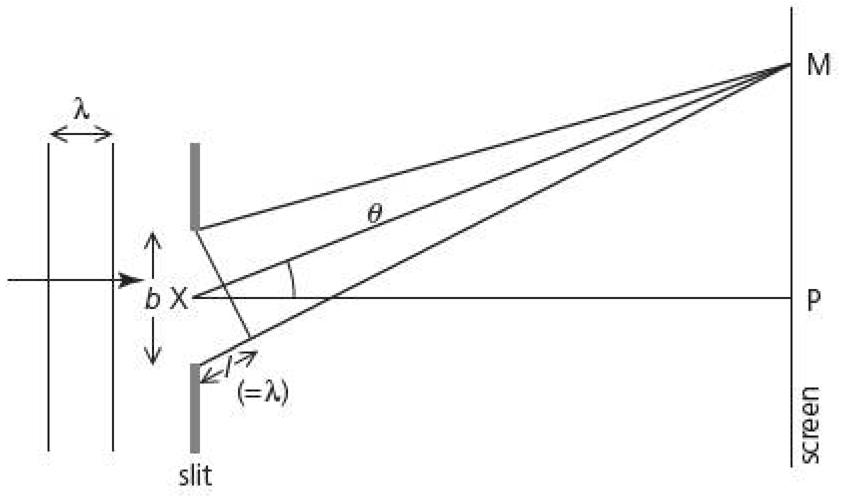


P_1 is a polarizer and P_2 is an analyser. The transmission axis of P_2 is at right angles to that of P_1. E is an electrode. G is a glass plate upon which a shaped electrode is etched. Unpolarized light is incident on P_1.

- **i** State, and explain, what the observer would see if the liquid crystal were not present. [2]
- **ii** Outline how the application of a potential difference between E and the electrode etched on G enables the observer to see the shape of the electrode. [3]

Higher Level Paper 2, Specimen Paper 09, QA3

Q3 a Plane wavefronts of monochromatic light of wavelength λ are incident on a narrow slit. After passing through the slit they are incident on a screen placed a large distance from the slit.

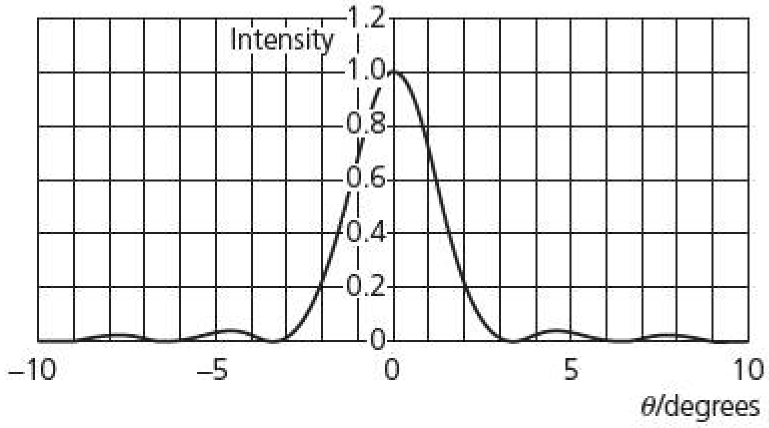


The width of the slit is b and the point X is at the centre of the slit. The point M on the screen is the position of the first minimum of the diffraction pattern formed on the screen. The path difference between light from the top edge of the slit and light from the bottom edge of the slit is l.

Use the diagram to explain why the distance l is equal to λ. [3]

The wavefronts in **a** are from monochromatic point source S_1. Diagram 1 is a sketch of how the intensity of the diffraction pattern formed by the single slit varies with the angle θ. The units on the vertical axis are *arbitrary*.

Diagram 1



Another identical point source S_2 is placed close to S_1 as shown in diagram 2.

Diagram 2

b The diffraction patterns formed by each source are just resolved. Make a copy of diagram 1 and sketch the intensity distribution of the light from source S_2. [2]

c Outline how the Rayleigh criterion affects the design of radio telescopes. [2]

d The dish of the Arecibo radio telescope has a diameter of 300 m. Two distant radio sources are 2.0×10^{12} m apart. The sources are 3.0×10^{16} m from Earth and they emit radio waves of wavelength 21 cm. Determine whether the radio telescope can resolve these sources. [3]

Higher Level Paper 2, May 09 TZ1, QB4 (Part 2)

12 Electromagnetic induction

STARTING POINTS

- Vectors quantities (like field strengths) can be resolved into two perpendicular components.
- Energy cannot be created or destroyed.
- Frequency = 1/period.
- Conductors contain free electrons which can be made to flow in an electric current. The direction of an electric current is always shown to be from positive to negative.
- Potential difference (p.d.), current, resistance and power in electrical components are interconnected by the following equations: $V = IR$, $P = IV$, $P = I^2R$, $P = V^2/R$.
- The emf of a source of electrical energy is the total energy transferred by it divided by the total charge flowing through it.
- The resistance of a wire of length L and area A can be determined from $R = \rho L/A$, where ρ is the resistivity of the conducting material. Good conductors have low resistivity.
- Charges experience forces in electric fields. *Moving* charges experience forces in magnetic fields.
- Fields are represented in drawings by field lines. The direction of an electric field is from positive to negative. The direction of a magnetic field is from magnetic north to magnetic south.
- The force acting on a charge moving perpendicularly across a magnetic field can be determined from $F = qvB$. The force acting on a charge in an electric field can be determined from $F = E/q$.
- The strength of a uniform electric field can be determined from the p.d. divided by the separation of the terminals.
- The magnetic permeability of a material is a measure of its ability to transfer a magnetic field.
- Electromagnetic radiation consists of photons and the energy of each photon can be determined from $E = hf$.
- The movement of conductors across magnetic fields can produce electric currents and this is how the world generates its electrical power. In general, an emf is generated whenever a changing magnetic field passes through any conductor.

12.1 Induced electromotive force (emf)

Inducing an emf by movement

12.1.1 **Describe** the inducing of an emf by relative motion between a conductor and a magnetic field.

Whenever a conductor moves across a magnetic field, or a magnetic field moves across a conductor, an emf may be **induced**. This effect is called **electromagnetic induction**. (Induction in this respect means the act of making something happen.) Although these two effects are essentially the same, we will begin by describing them separately.

Electromagnetic induction by moving a conductor across a magnetic field

Figure 12.1 shows a simple experiment in which an emf is induced when a conductor is moved across a magnetic field. The induced emf is detected because it makes a small current flow through a circuit containing a sensitive ammeter, called a **galvanometer**.

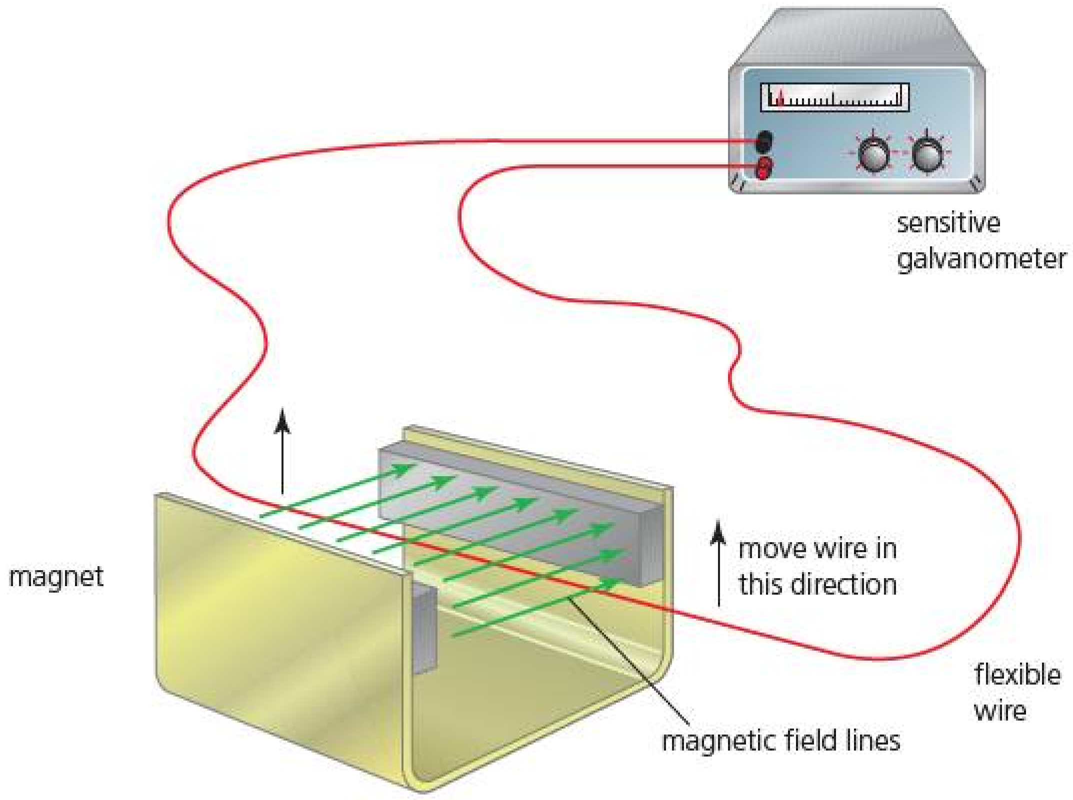


Figure 12.1 Inducing an emf by moving a wire across a magnetic field

The charged particles in the conductor experience forces (as discussed in Chapter 6) because they are moving with the wire as it crosses the magnetic field. Because it is a conductor, it contains **free electrons** that can move along the wire under the action of these forces. Other charges (protons and most of the electrons) also experience forces but are not able to move along the conductor. Moving the wire, containing free electrons, is equivalent to a current of *positive* charge in the opposite direction. This means we can use the left-hand rule to predict the forces on the electrons, as shown in Figure 12.2. In this case the magnetic force pushes the electrons to the left, so the left-hand end of the conductor becomes negatively charged, while the other end becomes positively charged (because some electrons have flowed away). This charge separation produces a potential difference (emf) across the ends of the conductor.

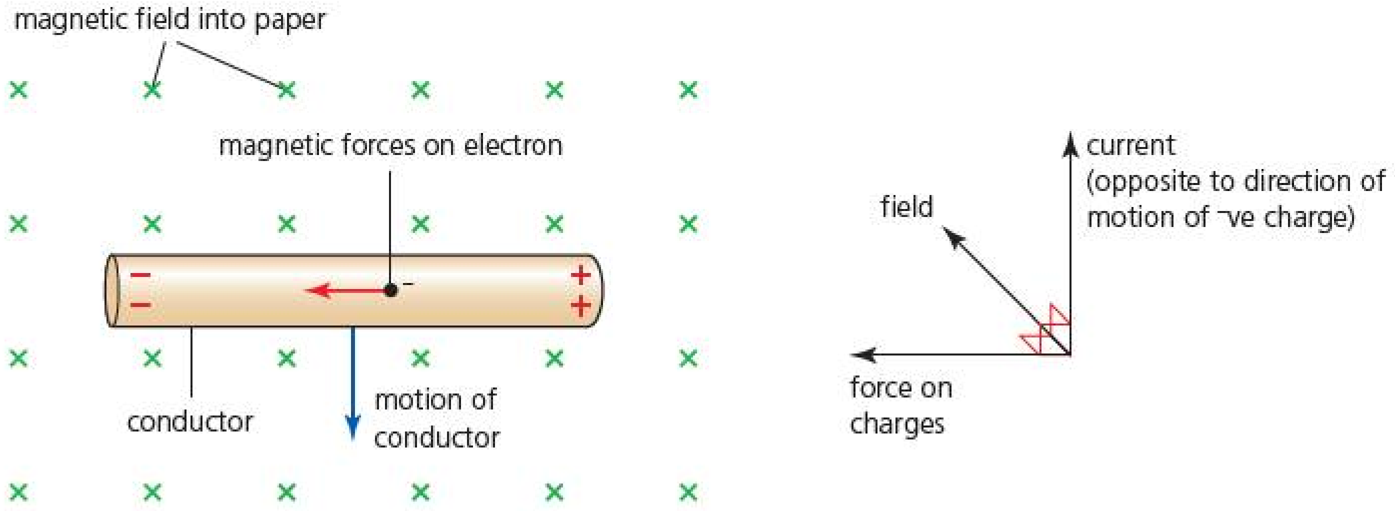


Figure 12.2 Magnetic forces on electrons produces charge separation

If the motion *or* the magnetic field is reversed in direction, then the emf and current are also reversed. (If both the motion *and* the magnetic field are reversed, then the emf and current are unchanged.) If the conductor and the magnetic field are both moving with the same velocity, no emf is induced. For electromagnetic induction to occur there must be *relative motion* between the conductor and the magnetic field. This effect is important and it needs it be well-understood because it has many applications. Most importantly, it is the basic principle behind the generation of the world's electrical energy.

In order to induce an emf, a conductor needs to move *across* a magnetic field. Magnetic fields are represented by *field lines* and the conductor needs to be moving so that it '*cuts*' across

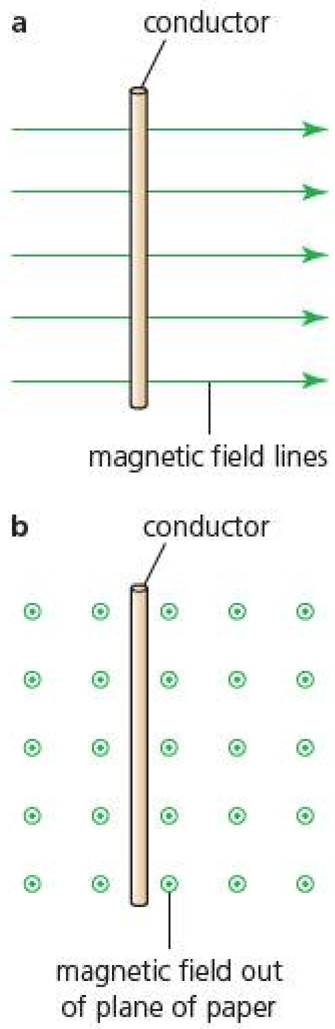


Figure 12.3 Direction of motion that causes electromagnetic induction

(through) the field lines. There will be no induced emf if the conductor is moving in a direction that is parallel to the magnetic field lines. Consider the examples shown in Figure 12.3. To induce an emf in **a** the conductor has to be moved so that its motion has a component into, or out of, the plane of the paper. In **b** the magnetic field is out of the plane of the paper and the conductor would have to be moved so that its motion has a component parallel to the paper. For similar conductors moving at the same speed, the induced emf is greatest if the motion is perpendicular to the magnetic field, as shown in Figure 12.4.

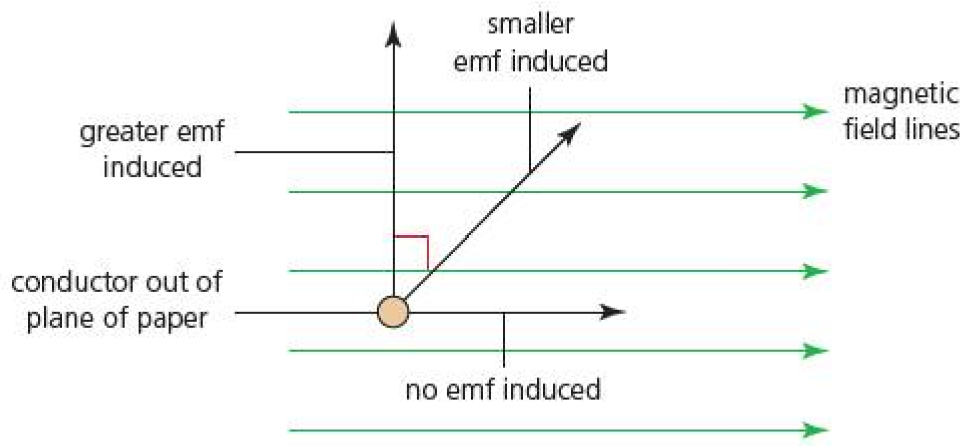


Figure 12.4 The size of an induced emf depends on the direction of motion

The induced emf in Figure 12.1 is very small, but there are simple ways to greatly increase it:

- faster motion
- use a magnet with a greater magnetic field strength
- wrap a long wire into a coil with many turns (and move *one* side of the coil through the field)
- use magnetic fields and coils with larger areas.

1 Explain why no emf is induced across string made of plastic when it is moved through a magnetic field.

2 Figure 12.5 shows a copper wire between the poles of a permanent magnet. In which direction(s) should the wire be moved to induce:
  a the greatest emf
  b zero emf?
  c Explain why no current can be induced in this wire as shown.

3 Draw a diagram to show how to demonstrate the induction of a current using a stationary magnet and a moving coil.

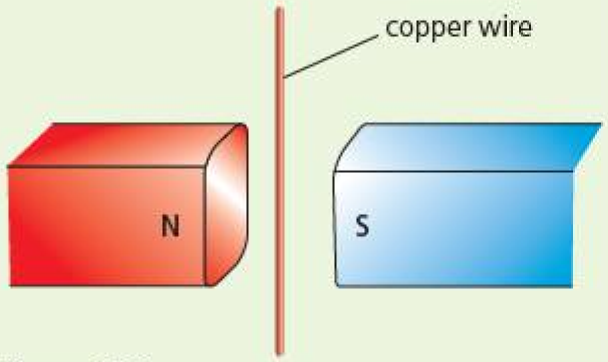


Figure 12.5

Electromagnetic induction by moving a magnetic field across a conductor

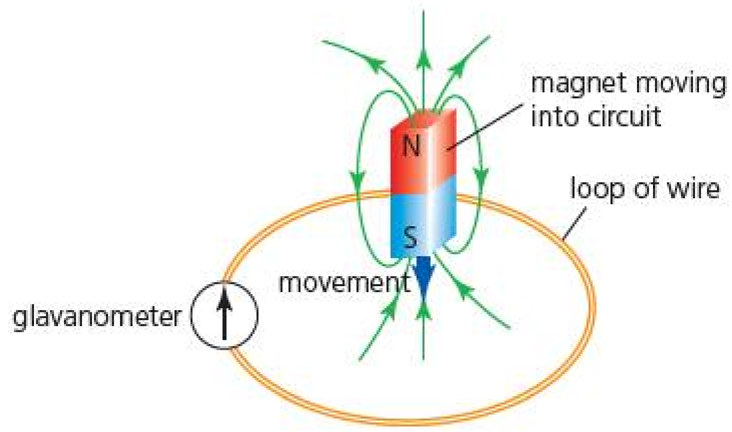


Figure 12.6 Moving a magnet to induce an emf and a current

Figure 12.6 shows electromagnetic induction by moving a magnetic field (around a permanent magnet) through a conductor. Again, the induced emf and current will be very small in this basic example, but Figure 12.7 shows how the effects can be greatly increased by winding the conductor into a coil or *solenoid* with many turns. The direction of the induced current around the coil will be reversed if the motion of the magnet is reversed.

Figure 12.8 shows an electromagnetic induction experiment recorded on a data logger and computer. The data logger records the emf induced at regular time intervals when a magnet is dropped through a coil and the data is used to draw a graph.

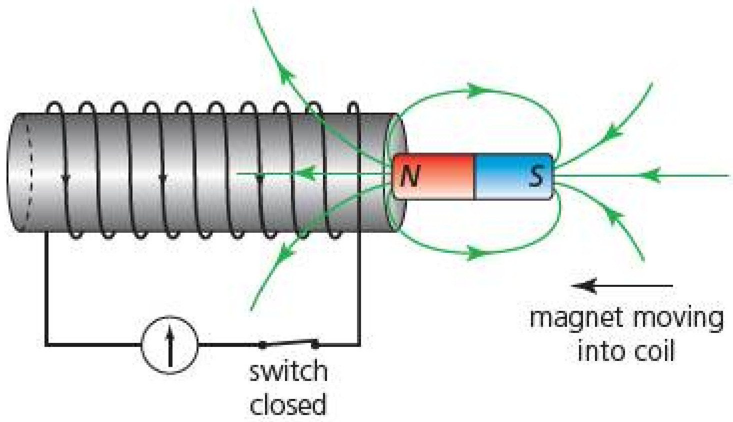


Figure 12.7 Inducing an emf and a current in a coil of wire

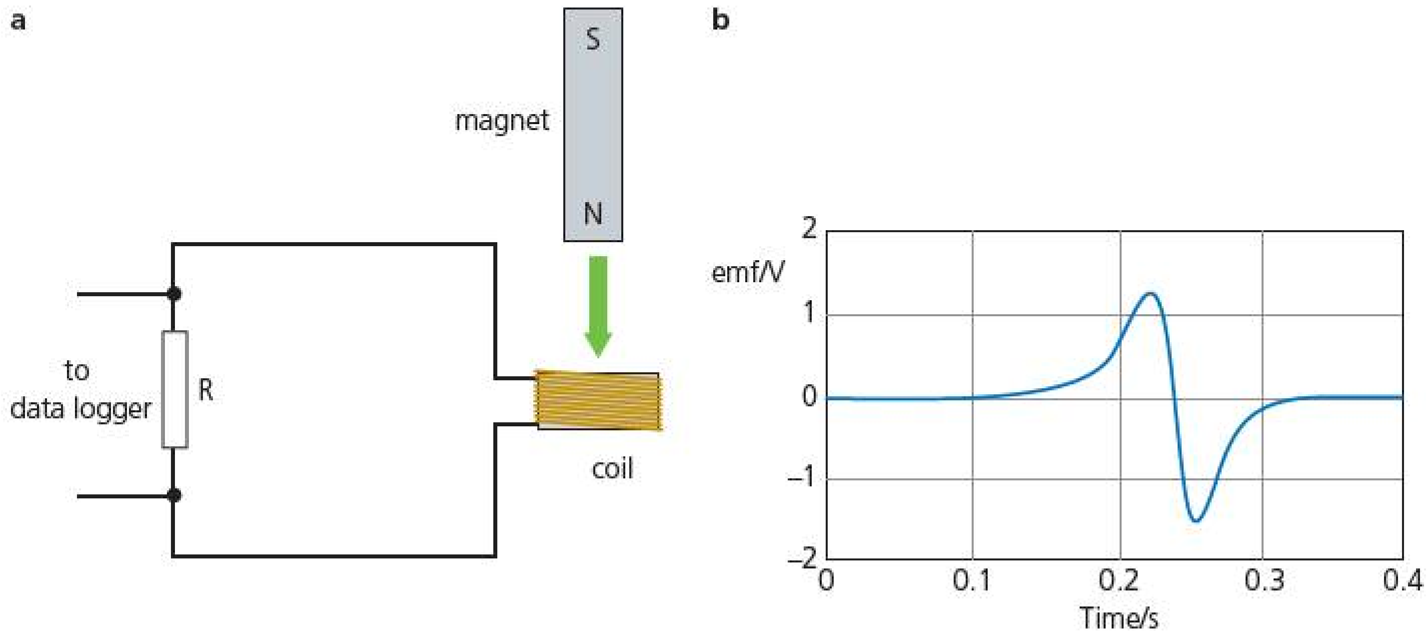


Figure 12.8 Inducing a current by dropping a magnet through a coil

4 In a demonstration of electromagnetic induction similar to that shown in Figure 12.6 the induced current was very small.
   a Suggest two ways of increasing the induced current while still using the same single loop of wire.
   b Give two ways in which the current can be made to flow in the opposite direction around the circuit.

5 Make a sketch similar to Figure 12.7, to show the current direction when the bar magnet comes out of the coil at the other end.

6 Suggest explanations for the shape of the graph shown in Figure 12.8.

An equation for the induced emf when a straight conductor moves across a uniform magnetic field

12.1.2 Derive the formula for the emf induced in a straight conductor moving in a magnetic field.

Figure 12.9 represents a conductor of length l moving perpendicularly across a uniform magnetic field of strength B, with a speed v.

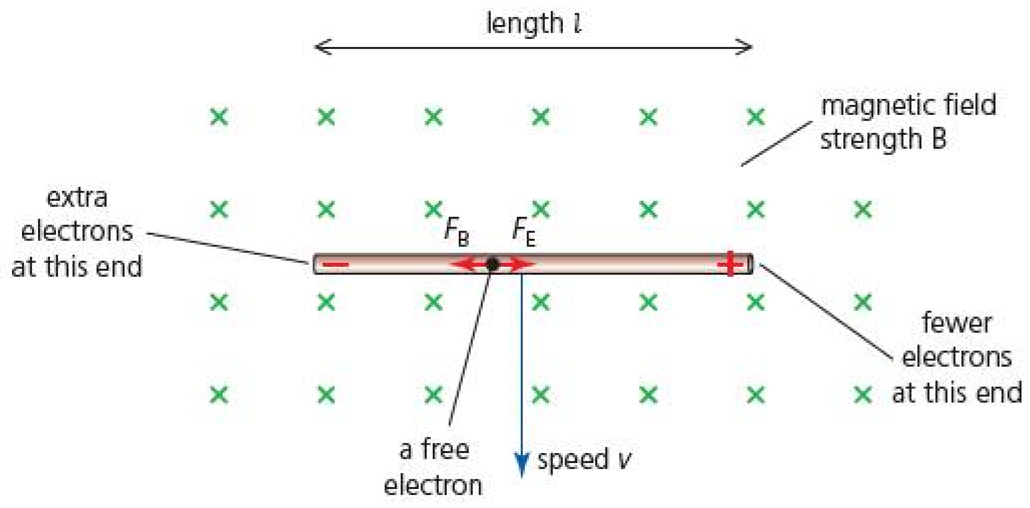


Figure 12.9 You need to know the length of the conductor in the field, the speed at which it is moving and the magnetic field strength to derive the equation for emf

Free electrons in the conductor will each experience a magnetic force, F_B, given by the expression, $F = qvB\sin\theta$ (Chapter 6). In this perpendicular arrangement $\sin\theta = 1$. These forces move free electrons towards the left of the conductor. As more electrons move along the conductor, the increasing amount of negative charge repels the motion of further electrons to that end. The right-hand end of the conductor, which has lost electrons, will become positively charged and act as an attractive force on the electrons.

The charge separation produces an electric field along the conductor, $E = \mathcal{E}/l$, where $\mathcal{E}$ is the induced emf. The maximum induced potential difference will occur when the force on each free electron due to the magnetic field, F_B, is equal and opposite to the force on the electron, F_E, due to the electric field (see Figure 12.9).

$$\text{electric field} = \frac{\mathcal{E}}{l}$$

$$\text{electric force, } F_E = \text{electric field} \times \text{charge} = \frac{\mathcal{E}q}{l}$$

$$F_E = F_B$$

$$\frac{\mathcal{E}q}{l} = qvB$$

So that,

$$\mathcal{E} = Bvl$$

This equation is given in the IB *Physics data booklet.*

This equation shows us that the induced emf is proportional to the speed, magnetic field strength and the length of conductor in the field. This is a useful equation, but it can only be used for a straight wire moving perpendicularly across a uniform magnetic field. If the field is not perpendicular to the wire, the component of the field in that direction has to be used in the calculation.

Worked example

1 a What emf is produced across a 23.0 cm long conductor moving at 98.0 cm s^{-1} perpendicularly across a magnetic field of strength 120 μT?

b If the motion is changed so that it moves at an angle, θ, which is 25° to the field, what is the new emf?

a $\mathcal{E} = Bvl$

$\mathcal{E} = (120 \times 10^{-6}) \times 0.98 \times 0.23 = 2.7 \times 10^{-5}\,\text{V}$

b In this example we need to use the *component* of the magnetic field which is perpendicular to the motion ($B\sin\theta$).

$\mathcal{E} = Bvl \times \sin 25° = 2.7 \times 10^{-5} \times 0.42 = 1.1 \times 10^{-5}\,\text{V}$

12.1.6 Solve electromagnetic induction problems.

7 When a conductor of length 90 cm moved perpendicularly across a uniform magnetic field of strength 4.5×10^{-4} T an emf of 0.14 mV was induced. What was the speed of the conductor?

8 What strength of magnetic field is needed for a voltage of 0.12 V to be induced when a conductor of length 1.6 m moves perpendicularly across it at a speed of 2.7 m s^{-1}?

9 A plane is flying horizontally at a speed of 280 m s^{-1} at a place where the vertical component of the Earth's magnetic field is 12 μT, as shown in Figure 12.10.

a Calculate the emf induced across its wing tips if its wingspan is 58 m.

b Suggest why this voltage would be larger if the plane was flying close to the North, or South Pole.

c Could the induced emf be used to do anything useful on the plane? Explain your answer.

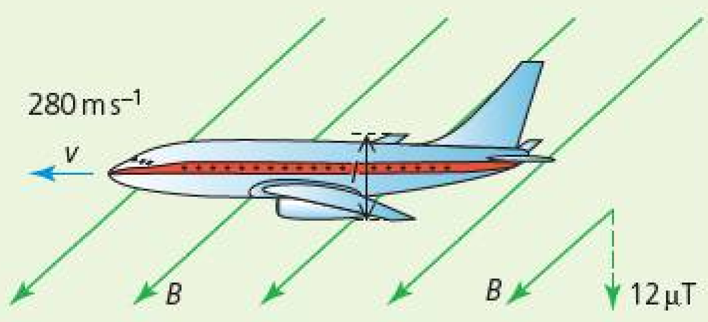


Figure 12.10

Magnetic flux and magnetic flux linkage

12.1.3 Define *magnetic flux* and *magnetic flux linkage*.

Magnetic flux

The size of an induced emf depends not only on the strength of the magnetic field, B, but also the area, A, of the circuit over which it is acting and the angle at which it is passing through the circuit. For these reasons it is useful to introduce and define the important concept of magnetic flux.

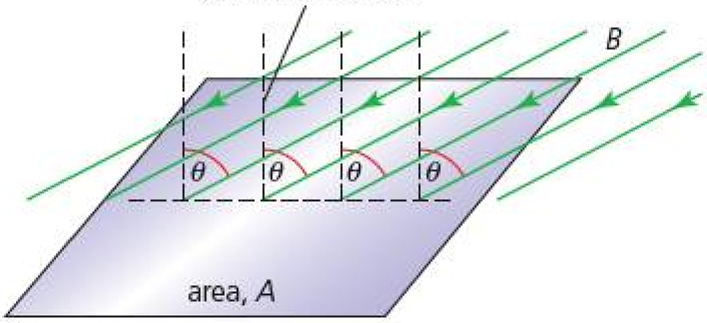


Figure 12.11 Magnetic flux depends on field strength, area and angle

Magnetic flux, Φ, (for a uniform magnetic field) is defined as the product of the area, A, and the component of the magnetic field strength perpendicular to that area, $B\cos\theta$. Figure 12.11 illustrates an example in which the angle θ is measured between the normal to the area and the direction of the field.

$$\Phi = BA\cos\theta$$

This equation is given in the IB *Physics data booklet*.

If the magnetic field is perpendicular to the area, then $\cos\theta = 1$ and the equation simplifies to

$$\Phi = BA$$

The unit of magnetic flux is the weber, Wb. One weber is equal to one tesla multiplied by one metre squared ($1\,\text{Wb} = 1\,\text{T}\,\text{m}^2$).

You can rearrange the equation to give, $B = \Phi/A$, and that is why magnetic field strength is widely known as **magnetic flux density** (flux/area).

Magnetic flux can be a difficult concept to understand and Figure 12.12 may help, as it shows a non-mathematical interpretation of magnetic flux, as the number of magnetic field lines that pass through the area. Flux also has many similarities with ideas in Chapter 8, where the total energy flow through an area depends on the power of the electromagnetic radiation, the area and the angle of incidence.

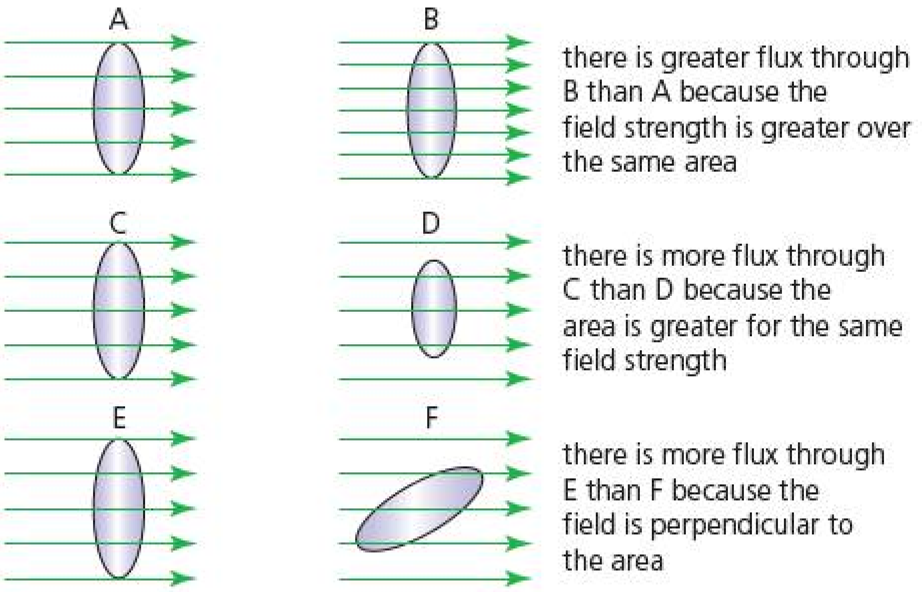


Figure 12.12 Magnetic flux explained in terms of field lines

Worked example

2 a Calculate the magnetic flux in a square coil of wire of sides 6.2 cm when it is placed at 45° to a magnetic field of strength 4.3×10^{-4} T.

b What would the magnetic flux be if only half of the coil was in the magnetic field?

a $\Phi = BA\cos\theta$

$\Phi = (4.3 \times 10^{-4}) \times (6.2 \times 10^{-2})^2 \times \cos 45°$

$\Phi = 1.2 \times 10^{-6}\,\text{Wb}$

b The magnetic flux would be reduced to half because the area used in the calculation is the area of the coil in the magnetic field, not the total area of the coil.

Magnetic flux linkage

Increasing the number of turns, N, in a circuit increases the size of an induced emf and, for this reason, the concept of magnetic flux linkage becomes useful. **Magnetic flux linkage** is defined as the product of magnetic flux and the number of turns in the circuit. It does not have a widely used standard symbol. The units of flux linkage are the same as flux (Wb).

Magnetic flux linkage $= N\Phi$

In Worked example **2 a** above, the *flux* through the coil was calculated to be 1.2×10^{-6} Wb. If the coil had 200 turns, the *flux linkage* would be $200 \times (1.2 \times 10^{-6}) = 2.4 \times 10^{-4}$ Wb.

12.1.6 Solve electromagnetic induction problems.

10 Calculate the magnetic flux in a flat coil of area 48 cm² placed in a field of magnetic flux density 5.3×10^{-3} T if the field is at an angle of 30° to the plane of the coil.

11 A magnetic field of strength 3.4×10^{-2} T passes perpendicularly through a flat coil of 480 turns and area 4.4×10^{-5} m². What is the flux linkage?

12 A flat coil of 600 turns and area 8.7 cm² is placed in a magnetic field of strength 9.1×10^{-3} T. The axis of the coil was originally parallel to the magnetic field, but it was then rotated by 25°. What was the *change* of flux linkage through the coil?

Electromagnetic induction by a time-changing magnetic flux

12.1.4 Describe the production of an induced emf by a time-changing magnetic flux.

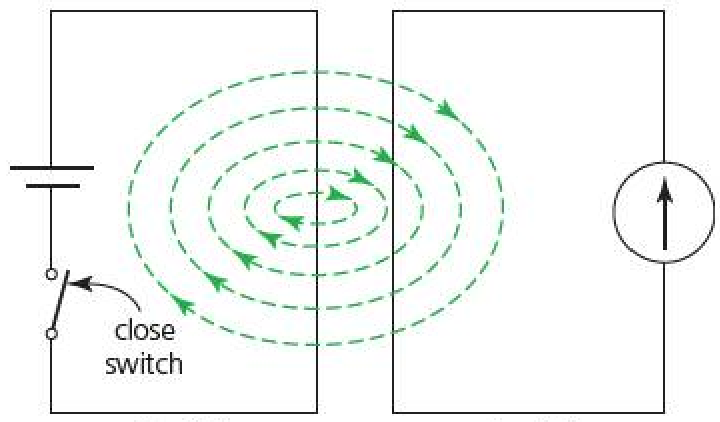


Figure 12.13 When the switch is closed, a magnetic field (flux) passes from circuit A to circuit B

Emfs can also be induced, not by movement, but by changes in the current in one circuit affecting another, separate circuit. Figure 12.13 represents the simplest example, which should be understood in detail. First consider circuit B at a time when the switch in circuit A is open – there is no power source and no magnetic field near it, so there is no current shown by the galvanometer.

When the switch in circuit A is closed, however, a current starts to flow around circuit A and this sets up a magnetic field (flux) around it, which spreads out and passes through circuit B. This sudden *change* of magnetic flux induces an emf and a current, which is detected by the galvanometer in circuit B. The *changing* current produces a *changing* magnetic flux in the same way as moving a magnet.

This induced emf/current only lasts for a moment, while the switch in A is being turned on, because when the current in A is constant there is no longer a *changing* magnetic flux. When the switch is turned off, there is an induced emf/current for a moment in the opposite direction.

This is a very small (but important) effect. However, the induced emf can be greatly increased by winding the conducting wires in *both* circuits into coils of many turns (to increase the flux linkage) and placing them on top of each other with an iron core through the middle. This is shown in Figure 12.14. (Remember from Chapter 6, that iron has high magnetic permeability.)

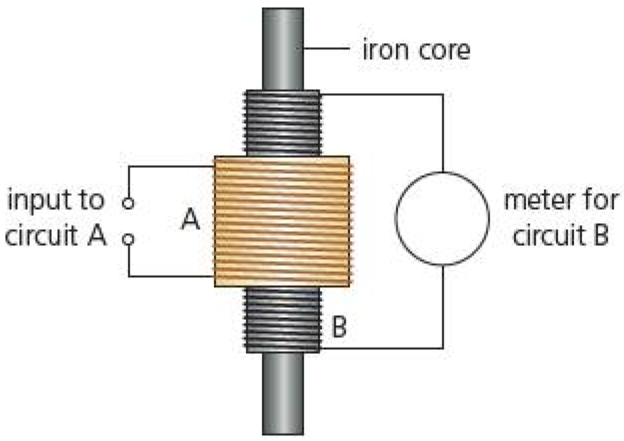


Figure 12.14 Making the induced emf larger by using an iron core and coils of many turns (increasing the flux linkage)

If the voltage source in circuit A is changed from one which gives a direct current (dc) of constant value to a source of alternating current (ac) then the magnetic flux in both circuits will change continuously and an alternating emf will be induced continuously. This effect has a large number of important applications, which are discussed later in this chapter.

13 a Figure 12.15 shows four loops of wire wrapped around a solenoid and connected to a centre-reading galvanometer (as shown on the right). Describe and explain what will be seen on the meter at the moment when:

i the switch is closed

ii the switch is opened.

b What is seen on the meter when a constant current is flowing in the solenoid?

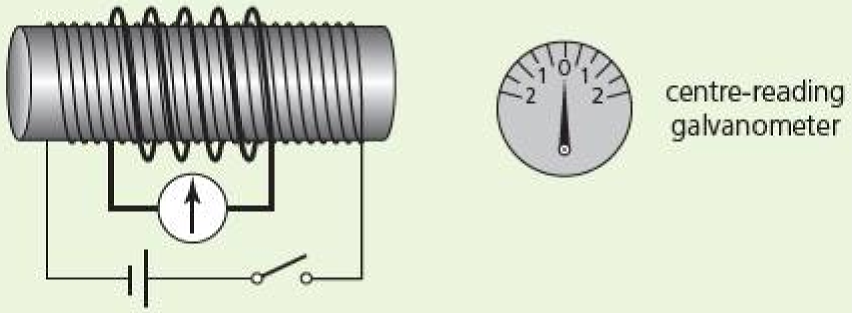


Figure 12.15

14 Figure 12.16 shows two coils wound on cardboard tubes. An alternating current flows through coil P and a changing emf is induced in coil S because there is a changing magnetic flux. Suggest four possible ways of increasing the maximum magnetic flux linkage in coil S.

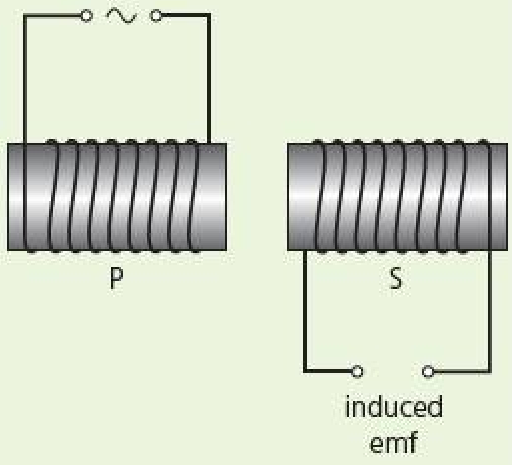


Figure 12.16

Summary: three kinds of electromagnetic induction

In order to induce an emf across a circuit, a *changing* magnetic field must pass through the circuit. There are three different ways in which this can be done:

1 Part of a circuit can be moved through a stationary magnetic field (as shown in Figure 12.1).
2 A magnetic field, for example around a bar magnet, can be moved through a circuit (as shown in Figures 12.6, 12.7 and 12.8).
3 A changing current in one circuit can produce a changing magnetic field which spreads out and passes through another circuit (as shown in Figures 12.13 and 12.14).

(When describing examples of electromagnetic induction generally, the terms magnetic field and magnetic flux are often interchangeable.)

Faraday's law of electromagnetic induction

12.1.5 State Faraday's law and Lenz's law.

Faraday's law

Considering the three kinds of electromagnetic induction, the size of an induced emf may variously depend on such factors as the strength and size of a magnetic field, the speed and direction of motion, the area of a circuit, the number of turns in a circuit, whether the circuit surrounds air or some other material, and the rate at which an electric current is changing.

Michael Faraday (Figure 12.17) was the first to realize that *all* these examples of electromagnetic induction could be summarized by a single law expressing the fact that *any* induced emf depends on the rate of change of magnetic flux linkage.

Figure 12.17 Michael Faraday (1791–1867) is considered to be one of the greatest scientists

Faraday's law of electromagnetic induction states that the magnitude of an induced emf is equal to the rate of change of magnetic flux linkage.

$$\varepsilon = -\frac{N\Delta\Phi}{\Delta t}$$

This equation is given in the IB *Physics data booklet*.

The negative sign is because of the conservation of energy as represented by Lenz's law (see page 416).

Worked example

3 The magnetic flux through a coil of 1200 turns increases from zero to 4.8×10^{-5} Wb in 2.7 ms. What is the magnitude of the induced emf during this time?

$$\varepsilon = -\frac{N\Delta\Phi}{\Delta t}$$

$$\varepsilon = -\frac{1200 \times (4.8 \times 10^{-5})}{(2.7 \times 10^{-3})} = -21\text{V}$$ (The negative sign is often omitted.)

12.1.6 Solve electromagnetic induction problems.

15 A coil of area 4.7 cm² and 480 turns is in a magnetic field of strength 3.9×10^{-2} T.
- **a** Calculate the maximum magnetic flux linkage through the coil.
- **b** What is the average induced emf when the coil is moved to a place where the perpendicular magnetic field strength is 9.3×10^{-3} T in a time of 0.22 s?

16 Consider the circuit shown in Figure 12.14. When the input to circuit A is an alternating current of frequency 50 Hz, the induced emf in circuit B has a maximum value of 4 V. Suggest what would happen to the induced emf in circuit B if the frequency in circuit A was doubled (with the same maximum voltage). Explain your answer.

17 Figure 12.18 shows a coil of 250 turns moving from position A, outside a strong uniform magnetic field of strength 0.12 T, to position B at the centre of the magnetic field in a time of 1.4 s.
- **a** Calculate the change of magnetic flux in the coil when it is moved.
- **b** What assumption did you make?
- **c** What is the change of magnetic flux linkage?
- **d** Calculate the average induced emf.
- **e** Sketch a graph to show how the induced emf changes as the coil is moved at constant speed from A to C (no values needed).

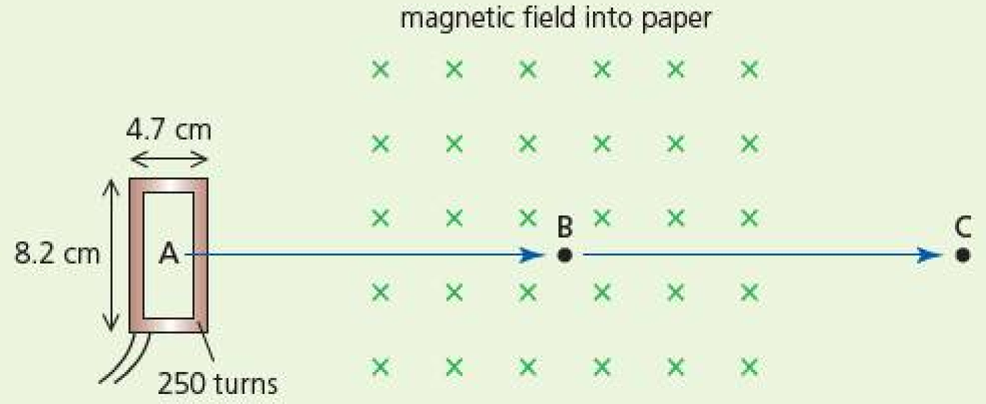


Figure 12.18

18 Imagine you are holding a flat coil of wire. You are in the Earth's magnetic field.
- **a** Make a sketch to show how you would hold the coil so that there is no magnetic flux through it.
- **b** At a place where the magnitude of the Earth's magnetic field strength is 48 μT, what emf would be induced by moving a coil of 550 turns and area 17 cm² from being parallel to being perpendicular to the magnetic field in 0.50 s?

19 A small coil of area 1.2 cm² is placed in the centre of a long solenoid with a large cross-sectional area. A steady current of 0.5 A in the solenoid produces a magnetic field of strength 8.8×10^{-4} T.
- **a** How many turns would be needed on the coil if an induced emf of 2.4 mV was needed when the current in the solenoid was increased to 2.0 A in a time of 0.10 s?
- **b** In what direction would the coil need to be positioned?

20 Refer back to the graph shown in Figure 12.8.
- **a** What quantity is represented by the area under the graph?
- **b** Compare the areas above and below the time axis.

21 Research into the work of Michael Faraday and Nikola Tesla. In your opinion, which of them contributed more to the development of the theories of electromagnetism? Explain your answer.

Energy transfers during electromagnetic induction

If a current is generated from motion by electromagnetic induction, then energy must have been transferred from outside the circuit. (We know this from an understanding of the law of conservation of energy.) The origin of this energy is the kinetic energy of the moving conductor or magnet. The moving object must therefore slow down as it loses some of its kinetic energy (unless there is a force keeping it moving).

For example, the magnets shown in Figures 12.7 and 12.8 must experience forces trying to slow them down (*opposing* their motions) as they move into the coils. This means that work has to be done to move the magnet into the coil, and this energy is transferred to the electromagnetically induced current. To understand why there is a force opposing motion, we need to remember ideas about the magnetic fields produced by currents from Chapter 6. If an (induced) current flows around the coil shown in Figure 12.7, then the coil will behave as a magnet. The direction of the current makes the end near the permanent magnet a north pole (this can be predicted using the right-hand grip rule, as mentioned in Chapter 6). The two north poles repel each other and cause a repulsive force opposing motion. If the induced current flowed the other way, then the force would be attractive and the magnet would gain kinetic energy, which would conflict with the law of conservation of energy.

If the polarity of the magnet moving into the coil was reversed, the induced current would flow the other way around the coil and then two south poles would repel each other. If the magnet is taken away from the coil, the current flows in the opposite direction and a magnetic field is set up that tries to stop the magnet moving away.

If the switch in Figure 12.7 is opened, there will still be an induced emf, but no current can flow. This means that there will be no magnetic field and no force from the coil.

Lenz's law

Because of the law of conservation of energy, we know that in *all* examples of electromagnetic induction, any induced emf and current *must* set up a magnetic field that tries to stop whatever change produced the current in the first place. This is known as **Lenz's law**.

Lenz's law states that the direction of an induced emf is such that it will *oppose* the change that produced it. This is represented mathematically by the negative sign in the Faraday's law equation.

12.1.6 Solve electromagnetic induction problems.

22 Figure 12.19 shows a bar magnet on the end of a spring placed close to a coil of wire.
- **a** If the magnet is pulled down so that its end is inside the coil and then released, sketch a graph to show how the position of the magnet will change with time for a few seconds.
- **b** On the same axes indicate how the emf induced across the coil changes.
- **c** Suggest what differences will be observed if the switch is closed while the magnet is oscillating. Explain your answer.

23 When a strong magnet is dropped vertically down inside a copper tube it takes significantly longer to reach the ground than when it is dropped outside the tube. Explain this observation.

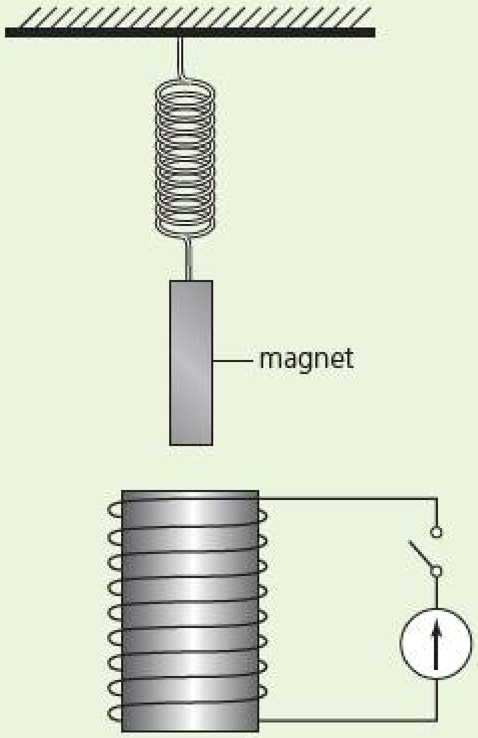


Figure 12.19

24 The loop of wire shown in Figure 12.20 is travelling at constant speed when it enters the magnetic field.
- **a** Explain why the loop decelerates.
- **b** Does the top of the loop act like a north pole or a south pole?
- **c** In which direction does the current flow in the side of the loop first entering the magnetic field?
- **d** Why is there no emf or current induced while the loop is moving through the magnetic field?
- **e** Discuss what happens as the loop leaves the other side of the field.
- **f** What has happened to the kinetic energy transferred from the loop because it reduced speed?

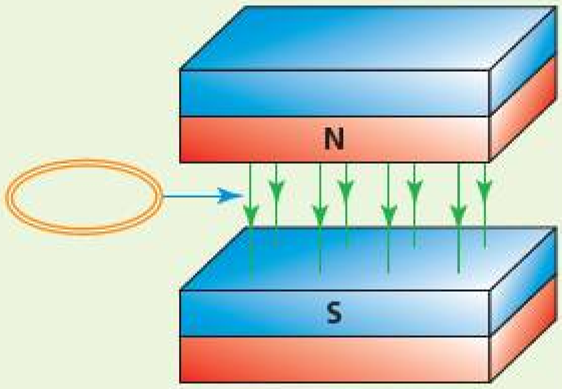


Figure 12.20

Additional Perspectives

Eddy currents and induction cooking

In an induction cooker, like that shown in Figure 12.21a, there are coils of wire below the flat-top surface. When a high frequency current is passed through a coil in the cooker, a strong, rapidly oscillating magnetic field is created which will pass through anything placed on or near the cooker's surface, like a cooking pot. If the material of the pot is a conductor, emfs and currents will be induced in it. Currents circulating within conductors, rather than around wire circuits, are known as **eddy currents**.

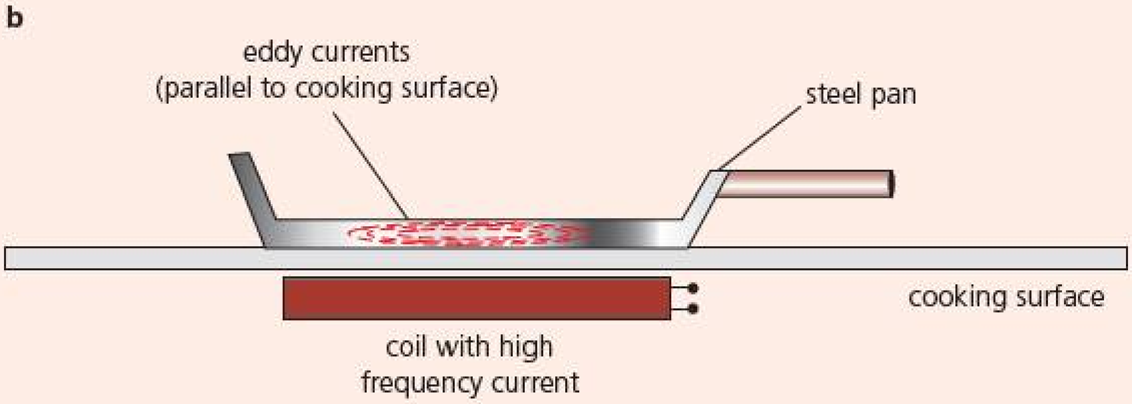


Figure 12.21 a A steel pan on an induction cooker **b** how an induction cooker works

The electrical energy in the eddy currents will be transferred to internal energy and the pan will get hotter. As in any circuit, the power generated can be calculated using $P = I^2R$. Thermal energy is then transferred by conduction to any food which is in the pan.

The choice of metal for the pan is important; for example, copper has a low resistivity and it also conducts thermal energy well. (This is not a coincidence; the ability to conduct thermal energy and the ability to conduct electricity are related. Both depend on the number of free electrons per unit volume in the material.) However, copper would *not* be a good choice for pans to be used with an induction cooker because it does not have good magnetic properties. The pots and pans need to be made of material with high magnetic permeability, that is a ferromagnetic material like iron or steel, so that the strength of the magnetic field in the pan is greatly increased. (Some pans made especially for use with induction cookers use a combination of different metals.)

Induction cookers have at least two significant advantages:

1 The cooker itself does not get noticeably hot (except for any thermal energy transferred from the pan). This means that induction heaters are safe, quick and reasonably efficient ways of cooking.
2 The heating process can be completely controlled, stopped or started by the turn of a switch.

Induction cookers are more efficient at cooking food than conventional electrical hot plates and both are a lot more efficient than natural gas cookers (although this does not allow for the inefficiencies of electrical power production). However, they are not in more wide-scale use because induction cookers are more expensive than gas or electric cookers (largely because of the extra electronics needed) and they also need special cooking vessels. In locations where space is limited, safety is an issue and an oven is not needed, small induction cookers may be the ideal choice.

Questions

1. Suggest why it is not possible to design and make an efficient induction oven.
2. Why do the heating coils in an induction cooker use high frequency currents?
3. Explain why, although induction cookers are much more efficient than gas cookers, they both create roughly the same 'carbon footprint'.
4. Find out how, and where, eddy currents can be used for braking and or flaw/crack detection.

12.2 Alternating current

ac generators

12.2.1 **Describe** the emf induced in a coil rotating within a uniform magnetic field.
12.2.2 **Explain** the operation of a basic alternating current (ac) generator.
12.2.3 **Describe** the effect on the induced emf of changing the generator frequency.

Consider a coil of wire as shown in Figure 12.22 between the poles of a magnet. (For simplicity, only one loop is shown, in practice there will be a large number of turns on the coil.) If the coil is rotating, there will be a changing magnetic flux passing through it and a changing emf will be induced. As side WX moves upwards, the induced emf will make a current flow into the page, if it is connected in a circuit. At the same time any induced current in YZ will flow out of the page, because it is moving downwards. In this way, the current flows continuously around the coil.

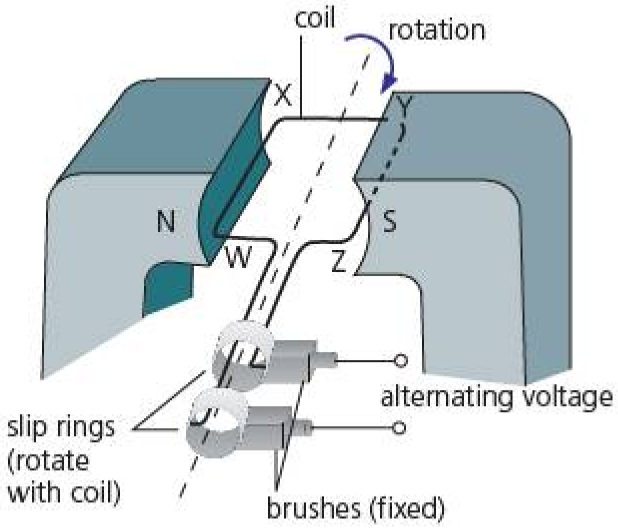


Figure 12.22 A simple ac generator

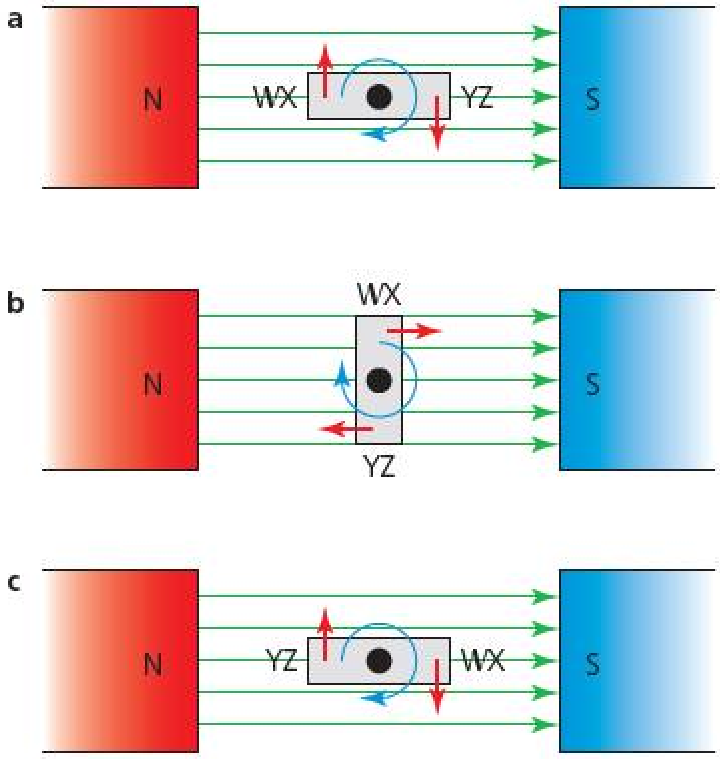


Figure 12.23 The sides of the coil cut the magnetic field at different angles as they rotate, alternating the emf produced

The connection between the coil and the external circuit cannot be fixed and permanent because the wires would become twisted as the coil rotated. Therefore carbon **'brushes'** are used to make the electrical contact with **slip rings** which rotate with the coil, so that the induced current can flow into an external circuit.

Figure 12.23 shows three views of the rotating coil from the side. In Figure 12.23a the plane of the coil is parallel to the magnetic field and, at that moment, the sides WX and YZ are cutting across the magnetic field at the fastest rate, so this is when the maximum emf is induced. In Figure 12.23b the sides WX and YZ are moving parallel to the magnetic field, so that no emf is induced at that moment. In Figure 12.23c the induced emf is a maximum again, but the direction is reversed because the sides are moving in the opposite direction to Figure 12.23a.

The overall result, if the coil rotates at a constant speed in a uniform magnetic field, is to induce an emf that varies **sinusoidally** (with a waveform shaped like a sine wave). This is shown by the green line in Figure 12.24. In positions B and D the plane of the

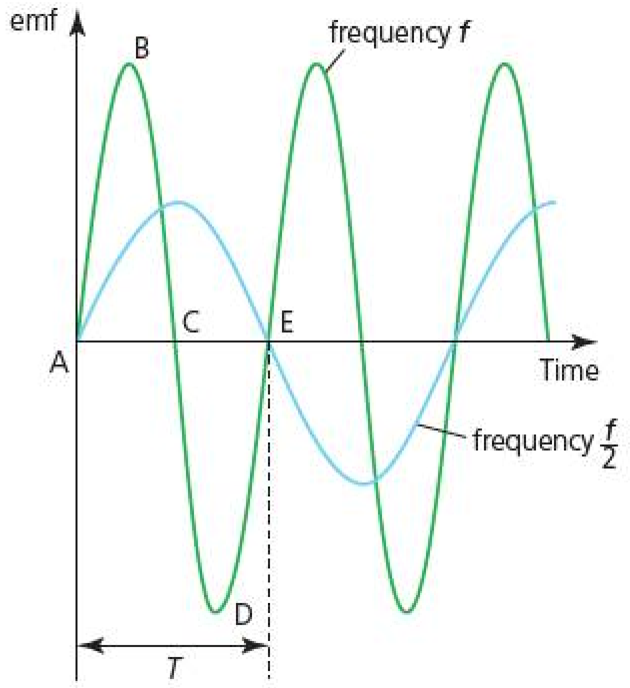


Figure 12.24 Comparing induced emfs at different frequencies

coil is parallel to the magnetic field. At A, C and E the plane of the coil is perpendicular to the field. One complete revolution occurs in time T. Frequency, f, equals $1/T$.

If the coil rotates at a slower frequency (fewer rotations every second), then there will be a smaller rate of change of magnetic flux through it and a smaller emf will be induced. For example, halving the frequency will halve the rate of change of magnetic flux linkage and therefore halve the induced emf. The time period is doubled. This is represented by the blue line in Figure 12.24. If possible, watch a computer simulation of an ac generator as the coil(s) rotate slowly, to aid your understanding. Throughout the world electrical energy is generated in this way using **ac generators**. Turbine blades can be made to rotate by the forces provided by high pressure steam produced from burning fossil fuels or nuclear reactions, or from falling water (as discussed in Chapter 8). These blades are attached to the large coils inside an ac generator. The coils – with many turns and cores with high magnetic permeability – are rotated in strong magnetic fields by the action of the turbine blades.

Electricity can also be generated using the same principle, but with the magnetic field rotating inside the circuit, rather than the other way around. Such devices are commonly used in cars and they are called *alternators*.

25 **a** Consider Figure 12.22. Explain why no emf is induced across the rotating end of the coil, XY.
b List four ways in which the induced emf could be increased.

26 A bicycle ridden at night should have lights on the front and the back. The energy for these is most conveniently provided using an alternator (often, somewhat misleadingly, called a bicycle *dynamo*). Make a simple sketch of the basic parts of a bicycle dynamo (which contains a rotating permanent magnet) and show on the drawing how kinetic energy is transferred to the rotating magnet from the bicycle.

27 An ac generator produces a sinusoidal emf of frequency 30 Hz and maximum value 9.0 V.
a Sketch an emf–time graph to show how the output of the generator changes during a time of 0.1 s.
b Add a second line to your graph to show the output from the same generator if the frequency is reduced to 10 Hz.

Additional Perspectives

Motors and induced emfs

Simple electric motors are similar in their basic design to generators – both consist of coils rotating in magnetic fields. We know from Chapter 6 that when a current is made to flow across a magnetic field it experiences a force. In electric motors, like the simplified version shown in Figure 12.25, forces in opposite directions on opposite sides of a current-carrying coil result in continuous rotation. Compare the simple ac generator in Figure 12.22 with the simple dc motor in Figure 12.25.

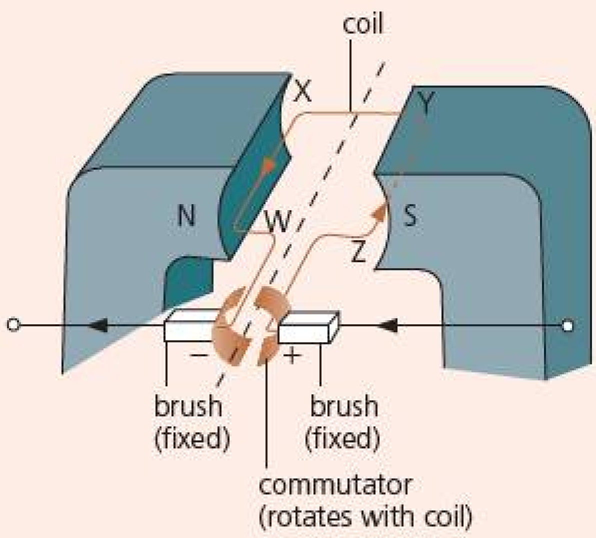


Figure 12.25 A simple dc motor

However, when the coil in a motor is rotating, electromagnetic induction also takes place, and an emf is induced across the coil which will tend to oppose its motion (Lenz's law). In other words, when a potential difference is applied to a coil in a motor to make it rotate, the motion then induces another potential difference in the opposite direction, trying to stop the rotation. This is often called a 'back emf'.

In a well designed, efficient dc motor which is spinning freely without doing any useful work, the back emf can be almost as much as the applied p.d., so that the resultant p.d. is just enough to drive a small current through the coil and provide the necessary low power. But when the motor is doing useful mechanical work, for example lifting a load, the rate of rotation decreases and the back emf is reduced (because of the decreased rate of changing magnetic flux linkage). The resultant p.d. and current increase to provide the extra power needed.

An advantage of a dc motor is that the speed of rotation can be controlled easily, but most motors operate with alternating currents. There are two main types of motor that use alternating currents:

1 A synchronous ac motor uses slip rings (as in the ac generator) and usually rotates with the same frequency as the ac supply.
2 An ac induction motor uses electromagnetic induction to transfer the energy to the coil.

Questions

1 Explain why a split ring commutator and brushes, as shown in Figure 12.25, are needed for a motor to rotate when supplied with a direct current.
2 Find out some basic information on how induction motors work.

Alternating currents and direct currents

When a current is described as **alternating**, it means that the direction of the current changes (usually in a regular way) and this is shown graphically by both positive and negative values. We have seen that the output from ac generators is usually sinusoidal, but other waveforms are possible, like the square and triangular waves shown in Figure 12.26.

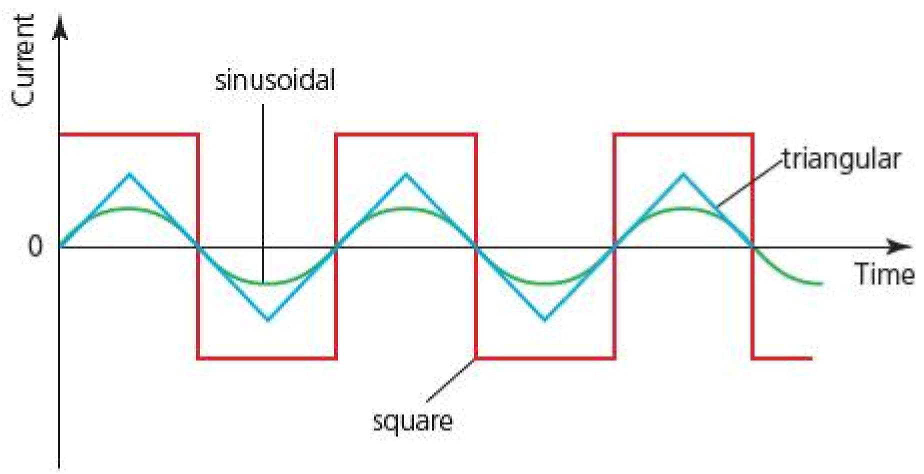


Figure 12.26 Alternating currents shown graphically with different waveforms and the same frequency, but different amplitudes

A **direct** current always flows in the same direction around a circuit (shown graphically by always being positive or always being negative), although its magnitude may vary.

Electrical energy is generated and transmitted using *alternating* currents because of the need to reduce energy losses. As we shall see in Section 12.3, high voltages are needed for transmission and this requires **transformers**, which can change (transform) voltages to higher or lower values. Transformers can only work using alternating currents.

Root mean squared (rms) values

12.2.4 Discuss what is meant by the root mean squared (rms) value of an alternating current or voltage.
12.2.5 State the relation between peak and rms values for sinusoidal currents and voltages.

Figure 12.27 shows the varying voltage and current (and therefore power) in a circuit with an ohmic resistor. For easy comparison all three are shown on the same axes. The peak (maximum) values of voltage and current are called V_0 and I_0. The power ($P = VI$)at any time can be determined by multiplying together values of voltage and current at that moment.

The maximum power is given by:

$$P_{max} = I_0 V_0$$

This equation is given in the IB *Physics data booklet*.

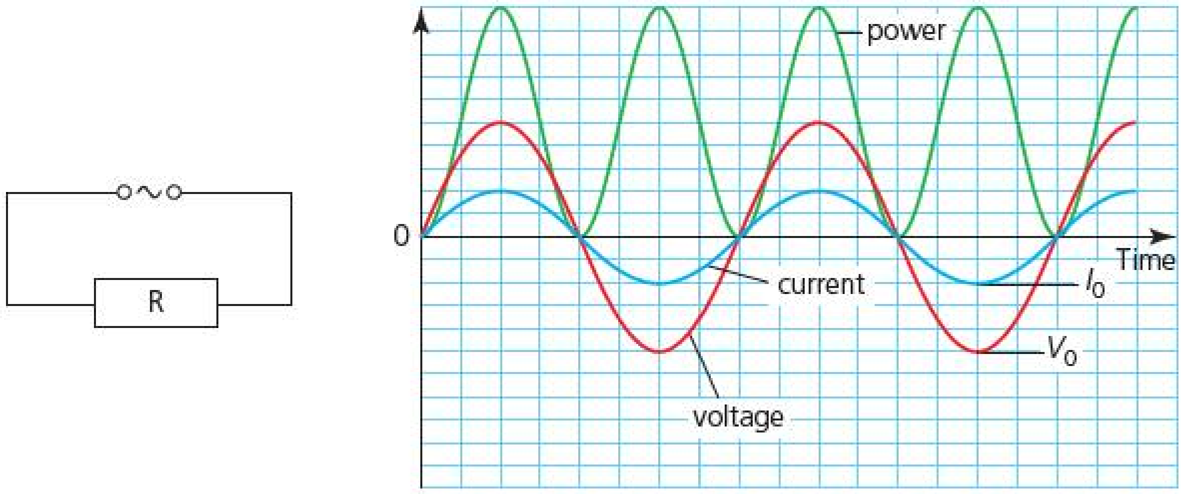


Figure 12.27 Oscillating voltage, current and power for a circuit with a resistor and an ac power supply

The power varies but always remains positive because the voltage and current are either both positive or both negative at all times. (This is only true for circuits which can be considered to be purely resistive.) The power transferred in the resistor does not depend on the direction of current flow.

Both the voltage and current variations are sinusoidal, which means that when they are multiplied together, the power variation ($P = VI$) follows a *sine squared relationship*. For this kind of relationship, the average value is exactly half of the peak value:

$$P_{average} = \frac{1}{2}P_{max} = \frac{1}{2}I_0 V_0$$

This equation is given in the IB *Physics data booklet*.

The **effective** values of the current and voltage are the values that give the average *power*; these are not the average values of the voltage and current, which are both zero, but are less than the peak (maximum) values. The equation below can be used to determine the relationship between peak and effective values for current and voltage:

$$P_{average} = I_{effective} \times V_{effective} = \frac{1}{2}I_0 V_0$$
$$= \left(\frac{I_0}{\sqrt{2}}\right) \times \left(\frac{V_0}{\sqrt{2}}\right)$$

The effective value of an alternating current (or voltage) is therefore its maximum value divided by $\sqrt{2}$. The effective value is called its **root mean square (rms)** value, or its **rating**.

The rms value (rating) given to an alternating current (or voltage) is the same as the value of a direct current (or voltage) that would dissipate power in an ohmic resistor at the same rate.

This means, for example, that 230 V ac and 230 V dc transfer energy in a resistor at the same rate, even though the alternating voltage rises to a maximum of $230 \times \sqrt{2} = 325$ V, while the dc voltage remains constant at 230 V.

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

These equations are given in the IB *Physics data booklet*.

The equations $R = V/I$ and $P = VI = I^2R = V^2/R$ can all be used with ac circuits using maximum values or rms values (but not a mixture), so that:

$$R = \frac{V_0}{I_0} = \frac{V_{rms}}{I_{rms}}$$

These equations are also given in the IB *Physics data booklet*.

Worked examples

4 If the resistor in the circuit shown in Figure 12.27 is 47 Ω and the supply voltage has a peak value of 12 V, what is the rms current in the circuit?

$$I_0 = \frac{V_0}{R} = \frac{12}{47} = 0.26\text{ A}$$

$$I_{rms} = \frac{I_0}{\sqrt{2}} = \frac{0.26}{1.41} = 0.18\text{ A}$$

5 Consider Figure 12.28, which shows the variation with time of a supply of an alternating p.d.

a What is the maximum value of the voltage?

b What is the rating of this supply?

c What are the time period and frequency of the supply?

d If this voltage were connected to a 64 Ω resistor, what would be the peak and rms currents?

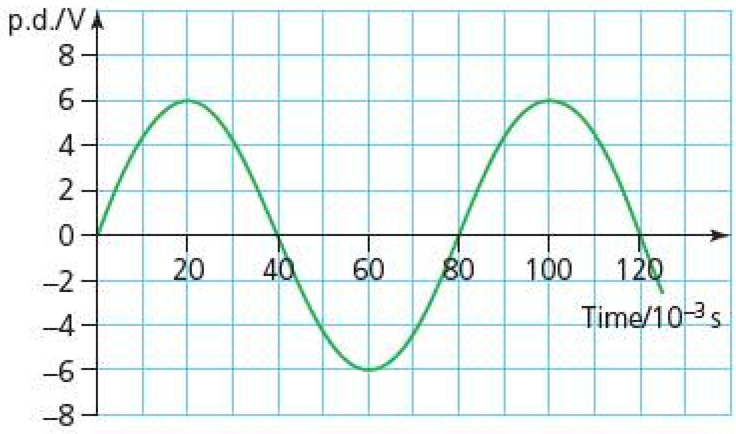


Figure 12.28

a ±6.0 V

b $\frac{6.0}{\sqrt{2}} = 4.2\text{ V}$

c $T = 8.0 \times 10^{-2}\text{ s}$

$$f = \frac{1}{T} = 12.5\text{ Hz}$$

d $I_0 = \frac{V_0}{R} = \frac{6.0}{64} = 0.094\text{ A}$

$$I_{rms} = \frac{I_0}{\sqrt{2}} = \frac{0.094}{1.414} = 0.066\text{ A}$$

Additional Perspectives

Oscilloscopes

Oscilloscopes are one of the most useful pieces of equipment found in a physics or electronics laboratory. They are effectively voltmeters for observing quickly changing and repeating potential differences – they are usually used to plot voltage–time graphs over short intervals of time. But it is not just voltages that can be displayed; any physical quantity which can be *represented* by a voltage can be shown in the same way. A common example would be using the output from a microphone as the input to an oscilloscope to show the pattern of the sound waves. A much simplified representation of the front of an oscilloscope is shown in Figure 12.29.

To begin with, a bright spot is seen on the screen and it can be moved around by voltages that make it move up and down or left and right. Circuits within the oscilloscope are usually used to make the dot move automatically from left to right across the screen at any of a very wide variety of speeds. The dot then moves back to the left of the screen and the process is very quickly repeated. This is known as using the **time base**. Unless the speed is quite slow, the moving dot will appear as a horizontal line.

Like any voltmeter, an oscilloscope is connected across the component whose p.d. is to be investigated, using the *y* input. If, for example, a sinusoidal voltage was applied to the *y* input at the same time as a suitable time base was moving the spot in the *x* direction, then a **trace** such as shown in Figure 12.29 would be seen on the screen, provided the *y* gain (amplification) was adjusted to a suitable value.

Question

1 If the squares shown in Figure 12.29 are each $1\,\text{cm} \times 1\,\text{cm}$ and the settings on the oscilloscope are $50\,\text{V}\,\text{cm}^{-1}$ and $4\,\text{ms}\,\text{cm}^{-1}$, calculate:
  a the peak voltage
  b the rms voltage
  c the time period
  d the frequency.

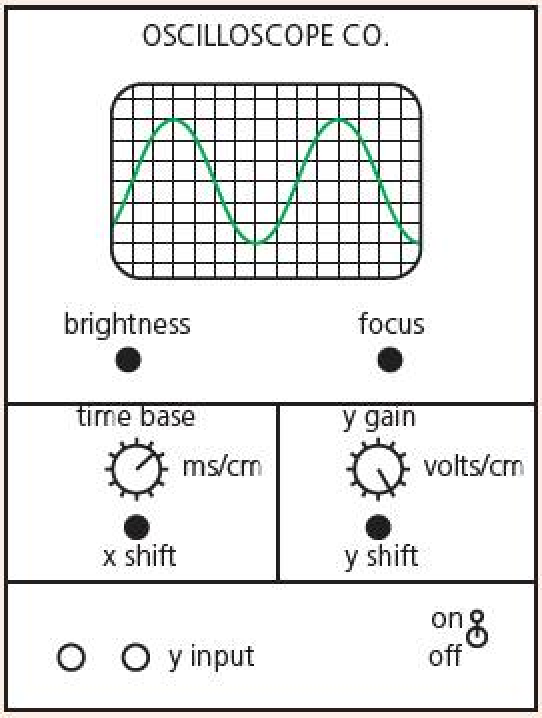


Figure 12.29 A simplified oscilloscope

12.2.6 Solve problems using peak and rms values.
12.2.7 Solve ac circuit problems for ohmic resistors.

28 a If the peak value of the power in a circuit like that shown in Figure 12.27 is 14 W, what is the average power?
b If the rms value of voltage is 25 V, what is the value of the resistance?

29 a In a country where the mains electricity has a rating of 230 V, what is its peak value?
b What is the peak value of the current through an electrical heater rated 2.15 kW in that country?
c What is the rms current through the heater?
d If the same heater was taken to a country where the voltage was rated at 110 V, at what average rate would it transfer energy?

30 What is the rms voltage across a light bulb that has an average power of 60 W when an alternating current of peak value 0.36 A flows through it?

31 a What is the average power dissipated in a 24.0 Ω resistor when an alternating voltage of peak value 150 V is applied across it?
b If the rms value of the voltage was doubled what would be the new power?
c What assumptions did you make?

32 A water heater rated at 2.0 kW is designed to work with a voltage rated at 250 V and frequency 60 Hz.
a Calculate the peak values of power, voltage and current.
b Sketch three graphs on the same axes, to show how voltage, current and power vary over 0.03 s. Include numerical values.

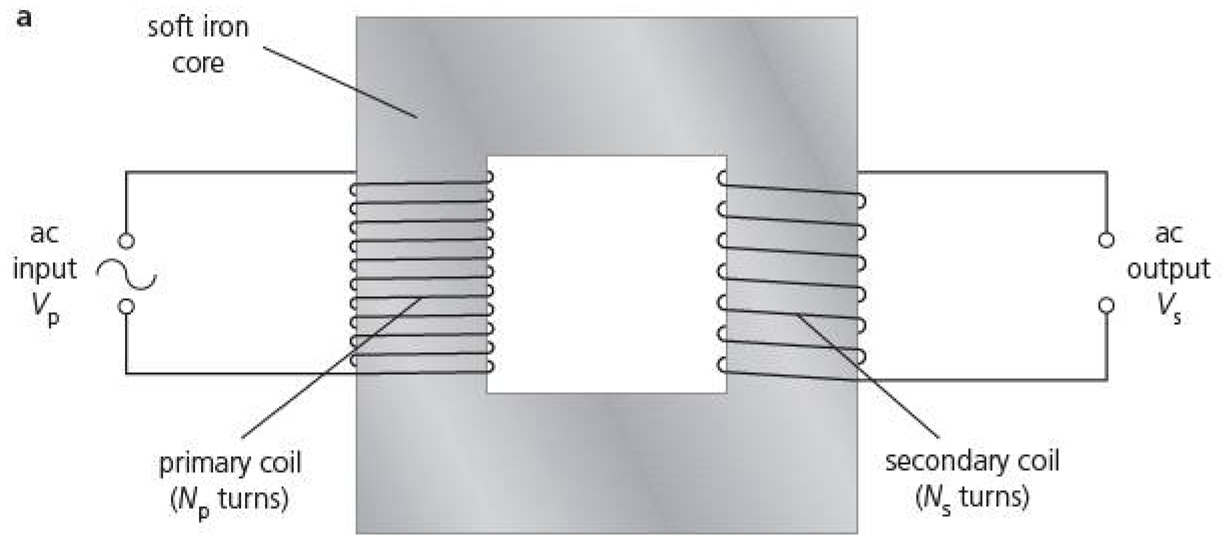
Transformers

12.2.8 Describe the operation of an ideal transformer.

The transfer of electrical energy from power stations to communities requires the use of much higher voltages than is safe for use in homes (or than is produced by the power stations). This means that there is the need to change (transform) voltages between different values, and the devices which do this are called **transformers**. Transformers are also needed in our homes whenever we need a voltage that is different from the mains supply. For example, we may wish to charge a mobile phone at 4 V in a country where electricity is supplied at 110 V.

Transformers use the principle of electromagnetic induction and, because this requires time-changing magnetic fluxes, transformers can only operate with alternating current, not direct current. Figure 12.30 shows the structure of a simple transformer.

The input voltage that we want to transform is connected to a coil of wire known as the **primary coil**. The output voltage is taken from a separate coil known as the **secondary coil**. The two coils are wound on the same **ferromagnetic (soft iron) core** which usually forms a complete loop, even when (as is usual) the coils are wound on top of each other.



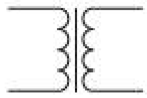
b


Figure 12.30 a The structure of a simple transformer and **b** the circuit symbol for a transformer

An alternating current in the primary coil creates an alternating magnetic flux which then passes through the iron core to the secondary coil and induces an emf.

The greater the number of turns on the secondary coil, the greater the output voltage (for a given arrangement). For an ideal transformer, the input and output voltage and the numbers of turns are linked by the following equation:

$$\frac{V_p}{V_s} = \frac{N_p}{N_s}$$

This equation is given in the IB *Physics data booklet*.

This equation suggests that by choosing a suitable turns ratio (N_p/N_s), an ideal transformer could in theory transform any alternating voltage into any other voltage. A transformer used to increase voltages is called a **step-up** transformer. **Step-down** transformers reduce voltages.

An **ideal transformer** is 100% efficient, so that power into primary coil equals power out of secondary coil.

$$V_pI_p = V_sI_s$$

This can also be expressed as:

$$\frac{V_p}{V_s} = \frac{I_s}{I_p}$$

This equation is given in the IB *Physics data booklet*.

A transformer which steps up the voltage must step down the current. A step-down transformer must have an increased current in its secondary coil. In real transformers, such as that in Figure 12.31, there will be some energy dissipated into the environment.

Figure 12.31 A transformer on a road-side pole

Worked example

6 An ideal transformer has 800 turns on its primary coil and 60 turns on its secondary coil.
a Is this a step-up or a step-down transformer?
b If a voltage of 110 V is applied to the primary coil, what is the output voltage?
c Calculate the current through a 12 Ω resistor connected to the secondary coil.
d Calculate the power in the secondary circuit.
e What is the current in the primary coil?

a A step-down transformer, because there are fewer turns on the secondary coil.

b $\frac{V_p}{V_s} = \frac{N_p}{N_s}$

$\frac{110}{V_s} = \frac{800}{60}$

$V_s = 8.3\text{ V}$

c $I_s = \frac{V_s}{R} = \frac{8.25}{12} = 0.69\text{ A}$

d $P_s = V_sI_s = 8.25 \times 0.69 = 5.7\text{ W}$

e Because it is an ideal transformer we know that the input and output powers are equal.

$I_p = \frac{P_p}{V_p} = \frac{5.7}{110} = 0.052\text{ A}$

Note that the input power and current can only be calculated *after* the power taken from the secondary circuit is known. If no current is taken out from an ideal transformer, then the input current and input power will also be zero.

12.2.9 **Solve** problems on the operation of ideal transformers.

33 A transformer has 600 turns on its primary coil.
a If it transforms 240 V to 12 V, how many turns are on the secondary coil?
b When the secondary coil was connected to a resistance, the current was 480 mA. What was the value of the resistance?
c Calculate the power developed in the secondary circuit.
d What was the current in the primary coil?

34 **a** Draw a fully labelled diagram of a transformer capable of operating a 10 V, 16 W lamp from a 220 V ac supply. Include the output circuit.
b What current flows in the primary circuit?

35 An ideal transformer with 36 turns on its primary coil is needed to step up a voltage from 110 V to 5.0 kV.
a How many turns should be wound on the secondary coil?
b If the input current was measured to be 0.55 A, what was the output current and power?

36 A turn ratio of 10:1 for a step-down transformer could be chosen from 10:1 or 100:10 or 1000:100 or 10000:1000, etc. Suggest factors which might affect the choice of the actual numbers of turns on the transformer coils.

12.3 Transmission of electrical power

Larger power stations tend to be more efficient than smaller ones, which means that most people do not live or work near the power station that supplies them with electrical energy. Therefore, electrical power often has to be **transmitted** significant distances between the place where it is generated and the place where it is used. The cables that transfer electricity around the country are called **transmission lines** (or sometimes **power lines**).

Figure 12.32 Transmission lines transfer electrical power around the country

Power losses in transmission lines and transformers

12.3.1 **Outline** the reasons for power losses in transmission lines and real transformers.

Power loss in transmission lines

We know from Chapter 5 that when a current, I, flows through a conductor of resistance R, some of the electrical energy is transformed to thermal energy. The power loss, P, can be calculated from $P = I^2R$.

Figure 12.33 Part of a power line with separate inner cables

Obviously, every effort must be made to keep the power loss in transmission lines as low as possible. This can be done by keeping the resistance of the cables and the current through them as low as possible.

The metal chosen for the conductor needs to have a reasonably low resistivity. However, resistivity needs to be balanced with economy, and using thicker cables of a cheaper metal rather than using thinner cables of a more expensive metal can cost less overall, assuming that they conduct equally well. (Remember $R = \rho L/A$ from Chapter 5.)

Other properties of the metal will also be important, such as its density, tensile strength and chemical reactivity. Overall, aluminium is usually considered the best choice for transmission lines, usually with a steel core to increase strength (see Figure 12.33).

Power loss in transformers

The transformers described in Section 12.2 were assumed to be ideal (100% efficient). In practice, of course, this is not possible. The main reasons that some of the input energy is dissipated to thermal energy in a transformer are as follows:

- The wires (windings) of the coils have resistance, so power ($P = I^2R$) is generated when currents flow through them.
- Currents are induced in the *core* of the transformer as well as in the coils. These induced currents (eddy currents), swirling around in conductors, will result in resistive heating.
- Energy is also transferred in the core as it is repeatedly magnetized and demagnetized. (This is an example of an effect known as **hysteresis**, in which the properties of the system lag behind the effect producing the changes.)
- *All* of the magnetic flux created in the primary coil will not pass through the secondary coil; there will be some 'leakage' of magnetic flux.

To reduce energy losses, thick wires of a metal with low resistivity (copper) should be used for the windings. The cores are also usually made with layers (laminations) of a ferromagnetic material separated by thin layers of insulation to greatly reduce eddy currents. Larger transformers are usually the most efficient – up to 98% efficient. Even at this efficiency, large transformers may need to be cooled to prevent overheating. For example, a transformer working at a power of 100 kW and efficiency 95% must safely dissipate 5000 J of thermal energy every second.

Most of the world uses electrical power with an alternating current frequency of 50 or 60 Hz. There are a number of conflicting factors affecting this choice, including the efficiency of transformers.

12.3.2 **Explain** the use of high-voltage step-up and step-down transformers in the transmission of electrical power.

Why transformers are needed for the transmission of electrical power

In order to deliver a certain electrical power, P, to a community, in theory it is possible to use any current, I, and any voltage, V, for the transmission lines, as long as $P = VI$. In other words, if the current is to be reduced (to reduce energy dissipation), then the voltage must be increased by the same factor in order to deliver the power required.

Transformers are used to step up the voltages across transmission lines as high as possible, so that the current can be decreased and the thermal energy wasted (I^2R) greatly reduced. Step-up transformers are placed as close as possible to power stations (see Figure 12.34). Decisions on the maximum value of voltages are based largely on a combination of economic and risk factors, and there is no ideal value. The cost of the extra precautions needed to make extremely high voltages safe has to be balanced against the financial and environmental costs of the extra power losses which occur at lower voltages. In theory, stepping up the voltage by a factor of, for example, 1000 can reduce the current by the same factor, and at the same time the power loss (I^2R) is reduced by a factor of $1000^2 = 10^6$.

Figure 12.34 Step-up transformers at a power station

Following transmission, the voltage has to be stepped down again to a value that is reasonably safe for use, for example, in homes and schools. In most countries this is in the range 200–250 V, although industries usually use a higher voltage. Step-down transformers are placed as close as possible to consumers.

(Note that the equation $P = V^2/R$ can be used with transmission lines to calculate the voltage drop *along* a cable. For example, if the power dissipated in a cable of resistance 2 Ω is 10^4 W, then the voltage drop along the cable is 140 V, which is not really significant if very high voltages are being used.)

The need to transform voltages explains why alternating currents are used – only ac currents induce the changing magnetic flux needed for transformers to work.

Figure 12.35 shows a block diagram of a simplified transmission system with some typical voltages. In reality, the voltages are stepped up and down by a series of transformers. The maximum voltage may rise to 750 kV or even higher.

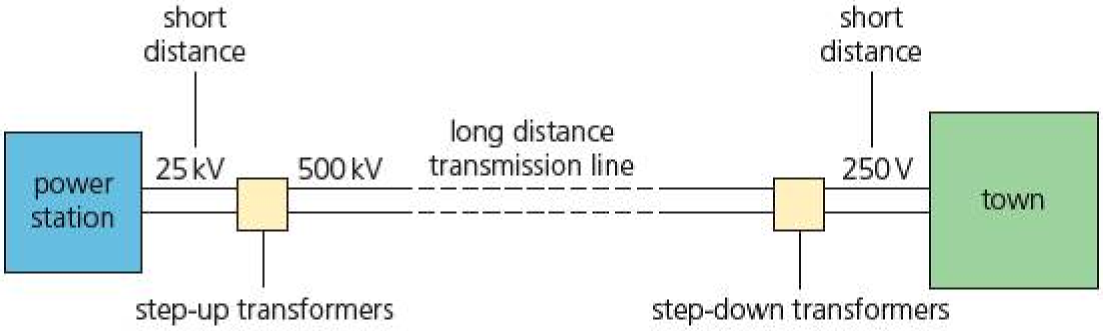


Figure 12.35 A simplified power transmission system

Transmission lines in different parts of a country are usually all connected together in what is known as a **transmission grid**. The grid helps balance supply and demand for electrical energy in different parts of a country, and sometimes also across international borders.

In our homes, many appliances (such as heaters and lights) work equally well if supplied with ac or dc currents, but electronic equipment only works with a dc current, so that the ac current has to be **rectified** (changed to dc).

Worked example

7 a What is the total current supplied to 2000 homes each using an average of 2.4 kW of electrical energy, if it is supplied using a voltage of 230 V?

b Calculate the resistance of a transmission cable made of pure aluminium ($\rho = 2.8 \times 10^{-8}\,\Omega\,m$) that has a radius of 3.0 cm and a length of 15 km.

c At what rate would electrical energy be transferred to thermal energy if the current in **a** flowed through the resistance in **b**?

d Comment on your answer.

a $P = VI$

$2000 \times 2400 = 230 \times I$

$I = 21\,000\,A$

b $R = \rho L/A$

$$R = \frac{(2.8 \times 10^{-8}) \times (1.5 \times 10^{4})}{(\pi \times 0.030^{2})}$$

$R = 0.15\,\Omega$

(This value certainly does not seem like much resistance for a cable 15 km long!)

c $P = I^2R$

$P = 21\,000^2 \times 0.15 = 6.6 \times 10^7\,\text{W}$

d This power loss is far too high. It would be much more than the useful power delivered to the homes, making the process very inefficient. The cable would get very hot and may melt.

12.3.3 Solve problems on the operation of real transformers and power transmission.

37 A small town requires an average of 635 kW of electrical power and is supplied by a transmission line of total resistance 0.76 Ω. Calculate the power loss in the cables if the voltage used is:
a 1000 V
b 250 000 V.

38 Suggest why industries may prefer to be supplied with electrical energy at a higher voltage than homes.

39 Explain why transformers should be placed as close as possible to power stations and to homes.

40 a Suggest why it is desirable for the metal used in transmission lines to have a low density.
b What chemical properties may be needed of the metal?

41 a If the primary windings (of a coil) in a step-up transformer have 120 turns, how many turns are needed on the secondary coil if it is to transform 20.0 kV into 380 kV?
b What voltage is the output from a step-down transformer at the other end of the transmission line, if it has 6800 turns on its primary coil and 80 turns on its secondary coil?
c What assumptions did you make?

42 a Suggest a reason why a transformer might work less efficiently at high frequencies.
b Suggest a reason why a transformer might work less efficiently at low frequencies.

43 A step-up transformer is used as part of transmission system. It has 160 turns in its primary coil and 1920 turns in its secondary coil.
a If the input voltage is 1200 V, what is the maximum theoretical output voltage?
b If a current of 160 A flows in the primary coil, what is the input power?
c The step-up transformer is 96% efficient. Use your answer to **a** to calculate the current that flows through the secondary coil.
d How much power is dissipated to thermal energy in the transformer?
e If the output transmission line has a resistance of 0.18 Ω, what is the power loss in the cables?

12.3.4 Suggest how extra-low-frequency electromagnetic fields, such as those created by electrical appliances and power lines, induce currents within a human body.

Are the electromagnetic fields around power lines and in our homes a health risk?

Absorption of electromagnetic energy in the human body

Any oscillating current in any circuit will create changing electric and magnetic fields which will spread away from the conductor. Mobile phones, wi-fi, TVs and radios, for example, all use this effect to great advantage with circuits and aerials designed to transmit and receive high frequency oscillating electromagnetic fields efficiently (a typical mobile phone frequency is 900 MHz).

Human (and animal) bodies are conductors of electricity. So, if any changing electromagnetic fields produced by circuits that use mains electricity pass into a body, it is scientifically reasonable to expect that very small currents may be electromagnetically induced, much like eddy currents in a piece of metal.

Some electronic equipment used in our homes (for example, computers) involve the use of high frequencies, but most of the electrical appliances in our homes use only the mains frequency of 50 or 60 Hz; the electromagnetic radiation from these can be described as **extra-low-frequency (ELF)**. Because the oscillating electromagnetic fields from mains electricity are ELF, the rate of change of magnetic flux is relatively small, so that any induced currents will also tend to be very small. The corresponding wavelength is very large (confirmed by using

$c = f\lambda$) and this also means that relatively little electromagnetic energy actually spreads away from the low frequency circuits.

Possible health risks

12.3.5 Discuss some of the possible risks involved in living and working near high-voltage power lines.

Much research has gone on around the world into the effects of ELF radiation on the human body. Most scientists believe that the effects are negligible and that there is no evidence of any significant health risk, such as damage to genetic material. Typically, the strength of the magnetic fields a person might experience from electrical devices in their home is much smaller than the Earth's magnetic field strength.

People understandably tend to be more worried about the health risks of being close to transmission lines. Concerns about living near high-voltage power lines include leukemia (especially in children), other cancers, depression and heart disease. However, there is currently no conclusive evidence making a link between these illnesses and living close to power lines.

Of course, it is certainly possible that, despite the current lack of definitive evidence, there *is* a health risk associated with living close to a power line (which could be confirmed by scientific research in the future). However, the fact that extensive research over many years has failed to make any significant connection suggests that any link, as yet unknown, may not be a great cause for concern.

Despite the lack of convincing evidence, a significant number of people still believe that there is a real risk to health, mostly because of reports they have read on the internet or in newspapers, or seen on the television. Such reports often tend to exaggerate any possibility of risks to make an interesting story, or fail to discuss any alternative scientific opinions. Of course, once the possibility of a risk is placed in their minds, many people believe that it is sensible to take precautions, even if that risk is very small or negligible. There are risks associated with everything we do as individuals, or as a society, and before we worry too much about any particular activity, it would be sensible to compare any possible risk with the risks of our other activities like, for example, crossing the road, taking a plane trip, smoking, or regularly eating fast-food.

Research is continuing into the possible effects of tiny induced currents in the human body. If there are any such effects, it would be reasonable to assume that they would be more pronounced with higher currents, higher frequency or a longer time in the field. However, at present, there are no known biophysical processes in the human body which would make scientists expect that exposure to such ELF electromagnetic fields could affect genetic material (DNA) or otherwise be a danger to health.

The energy carried by individual electromagnetic photons of extra-low frequencies is believed to be too small to cause harmful reactions. Similar comments may be applied to the electromagnetic radiation from mobile phones.

44 **a** Calculate the wavelength of electromagnetic radiation of frequency 50 Hz.
b Suggest a reason why electromagnetic wavelengths of this size do not spread away efficiently from household appliances (see Chapter 4).

45 **a** What amount of energy is carried by a single photon of frequency 50 Hz?
b Suggest a reason why electromagnetic radiation of this frequency could be considered to be unlikely to affect the human body, regardless of the intensity or duration of the radiation.
c Questions **44** and **45** use two different and conflicting theories about electromagnetic radiation, which are linked by the concept of frequency. What name is given to this dilemma?

46 Use the website of the World Health Organization (WHO) to find out what their opinion is on the effects of:
a living close to sources of ELF radiation
b using mobile phones for long periods of time.

47 Explain why it would be reasonable to assume that if there were any health risks associated with exposure to ELF radiation, then they would be worse with higher currents, higher frequency or a longer length of time in the electromagnetic radiation.

TOK Link: Correlation and cause

Suppose it was shown in a report on transmission lines that, for people between the ages of 20 and 50 who lived within 200 m of a high voltage overhead line for longer than five years, there was, on average, a 15% greater than average chance of developing a certain kind of cancer. Such information would be understandably very alarming for other people in similar situations and it would certainly receive a lot of publicity. The scientists and statisticians who produced the report may have stated clearly that people should not reach the conclusion that the transmission lines were definitely a health risk, but that information would have certainly received less attention.

Collecting statistical data like this is, by its nature, selective – researchers have to choose what information they are looking for and it is not easy to be completely unbiased, especially if they start with the definite intention of looking for a certain connection. Other, relevant information may have been overlooked, or perhaps the same data could be interpreted in a different way. For example, maybe over a period of 10 years the increased risk was much less than 15%, or maybe there was a decreased risk for people over the age of 50, but such information will receive less attention. Most scientists try to act with integrity and are well aware of the sometimes misleading nature of statistical evidence; there is rarely any intention to mislead.

Even if a connection between two sets of data is firmly established (that is, if there is a definite *correlation*), scientists may still be a long way from proving that one thing (for example, living near to a high voltage power line) *causes* another (for example, an increased cancer risk). There is always uncertainty in the accuracy of any data and statistical data like this is also always subject to unpredictable variations and limitations. So, any correlation may be purely coincidental, although good research and analysis should reduce this possibility.

Even if statistical evidence is convincing, it may still be misleading! Consider the confirmed fact that, *on average*, taller children tend to be better at learning physics than shorter children – in this case, being taller is not the *cause* of this correlation. Growing taller and achieving a better understanding of physics are both consequences of getting older (in children). It is common for a correlation to exist between two facts because they are both caused by something else.

Returning to the transmission line example, there could be other explanations of the possible correlation. It might arise because the people living near the power lines are not typical of the general population in some way, or perhaps there is something else about the environment of the power lines that causes an increased cancer risk, or maybe there is a completely unknown factor having an influence. Such uncertainty means that confirming a definite cause and effect can be very difficult, and society and individuals have to make judgements based on the best evidence available. In Chapter 8, the increasing levels of carbon dioxide were correlated against increasing average global temperatures, leading many scientists to believe that global warming is mainly caused by gases released when fossil fuels are burned, but there are still some people who have doubts about the link.

Additional Perspectives

Overhead or underground?

Almost all long-distance, high voltage transmission lines are overhead – held well above the ground on pylons for safety reasons (see Figure 12.32). The cables themselves are usually bare wires, without any insulation between them and the air, but very good insulators, like glass or ceramics, are needed where the cables are supported by the pylons (see Figure 12.36).

Figure 12.36 Ceramic insulation on a pylon

The breakdown electric field strength for dry air is usually quoted to be about $3300\,\text{V mm}^{-1}$. This suggests that a potential difference of 3.3×10^6 V would be needed for a dangerous spark to jump across from a high voltage transmission cable to a person or object 1 m away, and 6.6×10^6 V at 2 m apart, and so on. However, you are misled if you think that it might be safe to be within a few metres of an overhead cable. This is because the spark risk is unpredictable, depending very much on the shape and nature of the object close to the cable and the humidity of the air.

Many people think that pylons and transmission lines are ugly and spoil the beauty of the countryside, but the cost of burying cables underground is many times more expensive, partly because of the considerable extra insulation needed. When faults occur the repair of underground cables is also slower and more expensive.

Similar considerations apply to the power cables used in towns and cities but, because the voltages are lower, and because people are more concerned about safety and the appearance of their local surroundings, many communities prefer to pay the extra cost of placing cables underground. Alternatively, cables are supported on poles at the side of the road. Although they may not be pretty to look at, they are quickly and easily repaired and extra cables can be easily added.

Figure 12.37 Delivering electrical power to homes

Question

1 Look at Figure 12.37. Make a list of the advantages and disadvantages of delivering electrical power like this, compared with putting the cables underground.

SUMMARY OF KNOWLEDGE

12.1 Induced electromotive force (emf)

- An emf may be induced whenever a changing magnetic field passes through a conductor. This can be demonstrated using a moving magnet or a moving conductor.
- The magnitude of the induced emf increases with the speed of relative motion, the strength of the magnetic field and the number of turns in the circuit. The greatest emf is induced when the conductor is perpendicular to the magnetic field. No emf is induced if the conductor is parallel to the field.
- A voltage is induced because the free electrons within the conductor experience magnetic forces (see Chapter 6) and this results in a charge separation along the conductor.
- The equation for the emf induced in a straight conductor moving perpendicularly across a uniform magnetic field ($\varepsilon = Bvl$) can be derived by equating the electric and magnetic forces acting on free electrons.
- An emf may also be induced when a changing current in one circuit produces a time-changing magnetic field, which then passes through a separate circuit.
- The magnetic flux, Φ, through a circuit is defined as the product of its area, A, and the component of the magnetic field strength perpendicular to that area, $B\cos\theta$. $\Phi = BA\cos\theta$. Magnetic field strength is also called magnetic flux density.
- Magnetic flux linkage is defined as the product of the magnetic flux and the number of turns in the circuit, $N\Phi$.
- Faraday's law of electromagnetic induction states that an induced emf is equal to the rate of change of magnetic flux linkage. $\varepsilon = -N\Delta\Phi/\Delta t$.
- The negative sign shows that an induced emf always opposes the change of magnetic flux which produced it (Lenz's law). In other words, an induced current will always flow in such a direction as to produce a magnetic field which opposes the changing field that produced it.

12.2 Alternating current

- When a coil of wire rotates within a magnetic field, the emfs induced on opposite sides will combine so that a current can be driven continuously around the coil.
- Carbon brushes and slip rings in an ac generator enable a current in a rotating coil to be delivered to an external circuit.
- If the coil rotates at a constant speed, the induced emf and current will have sinusoidal waveforms. Graphs of emf and current against time can be drawn and points on the graph related to positions of the coil.
- If the speed of rotation increases, the induced emf will have a greater amplitude and frequency.
- The rate of transfer of electrical energy to internal energy (power) in a resistor varies continuously during each cycle of an alternating current. $P_{max} = I_0V_0$.
- The average power equals half the peak (maximum) power of a cycle ($P_{av} = ½I_0V_0$). This means that the effective values for the current and voltage are equal to their maximum values divided by $\sqrt{2}$. $I_{rms} = I_0/\sqrt{2}$ and $V_{rms} = V_0/\sqrt{2}$.
- The effective values of current and voltage are called rms (root mean squared) values, or ratings. An rms value is that value of an ac current (or voltage) which dissipates power in a resistor at the same rate as a steady direct current (or voltage) of the same value. Ohm's law can be used with rms values. $R = V_{rms}/I_{rms} = V_0/I_0$.
- A transformer changes an alternating voltage to another value. It uses the electromagnetic induction created by a time-changing magnetic flux. An alternating voltage applied to the primary coil creates an alternating magnetic flux linkage which passes through an iron core to the secondary coil, where it induces an emf.
- For an ideal transformer: $N_p/N_s = V_p/V_s$ ($= I_s/I_p$, because the power in the primary equals the power in secondary).

12.3 Transmission of electrical power

- Electrical energy is transferred around the world using 'transmission lines', which are thick cables made of good metallic conductors. The cables need to have low resistance ($R = \rho L/A$), so they are made of metals of relatively low resistivity. Aluminium is usually used because of its low cost and density. Thinner cables of copper (for example) could equally well be used, but the cost would be higher.
- Significant amounts of energy may still be transferred to internal energy if high currents flow through the cables ($P = I^2R$).
- In order to minimize power losses by keeping the currents in transmission lines low, high voltages must be used if large amounts of power are to be transferred ($P = VI$). Therefore, transformers are needed to step up and step down the voltage at various places along the transmission line. Safety considerations will limit the maximum voltages that can be used.
- Real transformers are not 100% efficient because of power losses: energy is transferred to internal energy when **i** currents flow through the coils; **ii** eddy currents flow through the core; **iii** the core is repeatedly magnetized and demagnetized. Some magnetic flux also escapes from the iron core.
- Because transformers are needed, electrical power is transferred using alternating currents (dc would not produce the necessary changing magnetic flux for the operation of a transformer).
- All alternating currents create changing electromagnetic fields which spread away from the conductors. Oscillating charge is the origin of all electromagnetic radiation. The electromagnetic fields which are created by mains electricity in the home and in transmission lines have 'extra-low-frequency' (ELF) when compared to other parts of the electromagnetic spectrum (50 Hz or 60 Hz).
- When alternating electromagnetic fields pass through the human body it is reasonable to assume that electromagnetic induction will induce very, very small currents. There is no convincing evidence that this is harmful, but there have been some reports suggesting that living near transmission lines has a health risk.

Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 Power losses in transmission lines can be reduced by the use of step-up and/or step-down transformers. Which of the following describes the correct positions for the transformers?

A step-up transformer near the power station; step-down transformer near the consumers
B step-up transformer near the power station; step-up transformer near the consumers
C step-down transformer near the power station; step-down transformer near the consumers
D step-down transformer near the power station; step-up transformer near the consumers

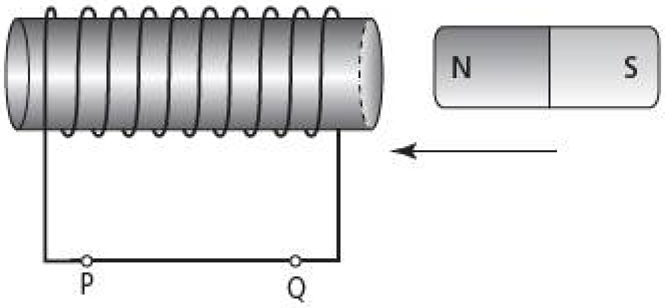
Q2 Which of the following is *not* a possible reason for power losses in a transformer?

A eddy currents in the core
B resistive heating of the windings of the coils
C friction between the coil and the core
D 'leakage' of magnetic flux from the core

Q3 An increasing magnetic flux passes through a metal ring of cross-sectional area A. If the magnetic flux increases in value by F in a time t, the emf induced in the ring is:

A F
B F/t
C Ft
D FAt

Q4 A permanent bar magnet is moved towards a coil of conducting wire wrapped around a non-conducting cylinder. The ends of the coil P and Q are joined by straight piece of wire.



The induced current in the straight piece of wire is

A alternating.
B zero.
C from P to Q.
D from Q to P.

Higher Level Paper 1, May 09 TZ1, Q24

Q5 The equation $\varepsilon = Bvl$ describes electromagnetic induction. Which of the following is *not* true?

A v is the speed of a conductor moving perpendicularly across a magnetic field.
B B is the magnetic flux.
C l is the length of the conductor.
D ε is the magnitude of the induced emf.

Q6 If a current is described as alternating (ac), which of the following *must* be changing?

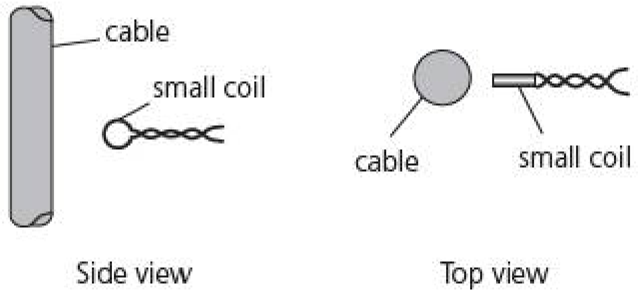
A frequency of the current
B magnitude of the maximum current
C effective value of the current
D direction of the current

Q7 The frequency of rotation of the coil in an ac generator is doubled. The rms voltage output:

A remains the same.
B is multiplied by 2.
C is multiplied by $\sqrt{2}$.
D is divided by $\sqrt{2}$.

Paper 2 IB questions and IB style questions

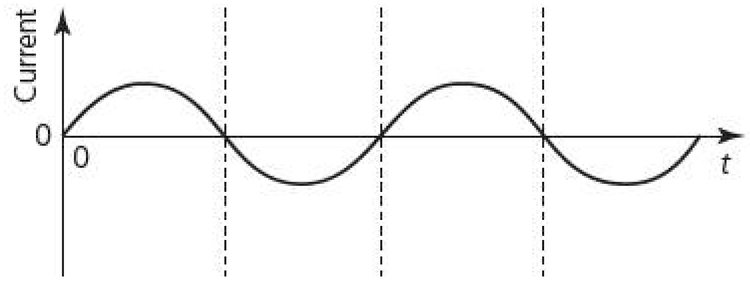
Q1 a In order to measure the rms value of an alternating current in a cable, a small coil of wire is placed close to the cable.



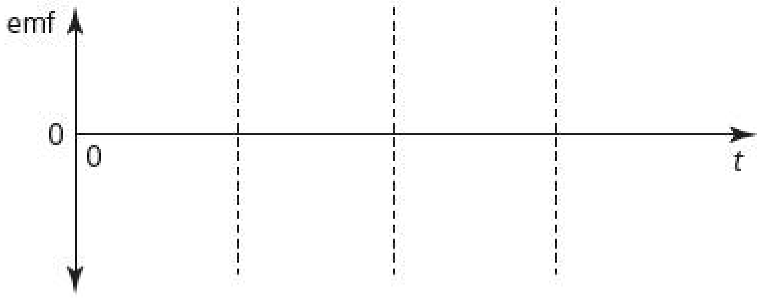
The plane of the small coil is parallel to the direction of the cable. The ends of the small coil are connected to a high resistance ac voltmeter.

Use Faraday's law to explain why an emf is induced in the small coil. [3]

b The graph below shows the variation with time t of the current in the cable.



Copy the axes below. On your copy, draw the variation with time of the emf induced in the small coil. [2]



c Explain how readings on the high resistance ac voltmeter can be used to compare the rms values of alternating current in different cables. [3]

Higher Level Paper 2, Nov 10, QA4

13 Quantum physics and nuclear physics

STARTING POINTS

- According to the principle of conservation of energy, the total energy in an isolated system remains constant.
- Electric charge can be either positive or negative.
- Electric charge is measured in coulombs.
- Like charges repel; unlike charges attract each other.
- The variation with distance of the size of forces between charges follows an inverse square law.
- An electric field is a region of space where a charge experiences a force.
- Using wave theory, the energy of electromagnetic radiation is assumed to propagate in the form of continuous transverse waves.
- Using the 'particle' model electromagnetic radiation travels in the form of photons.
- The energy of a photon is determined by its frequency.
- Some large nuclei are unstable and decay to form more stable nuclei. Ionizing radiation is released during this process.
- The scattering of alpha particles from thin metal foils provides evidence for the nuclear model of the atom.

13.1 Quantum physics

In 1900 it was believed that physics was almost fully understood, although there were a few knowledge 'gaps', such as the nature of atoms and molecules and the ways in which radiation interacted with matter. Many physicists believed that filling in these gaps was unlikely to involve any new theories. However, within a few years the 'small gaps' were seen to be fundamental, and a radically new theory was required. One important discovery was that energy only came in certain *discrete* (separate) amounts known as **quanta** (singular: **quantum**). The implications of this discovery were enormous and collectively they are known as *quantum physics*.

As quantum physics developed, some classical concepts had to be abandoned. There was no longer a clear distinction between particles and waves. The most fundamental change was the discovery that systems change in ways which cannot be predicted precisely; only the probability of events can be predicted.

The quantum nature of radiation

Photoelectric effect

13.1.1 Describe the photoelectric effect.

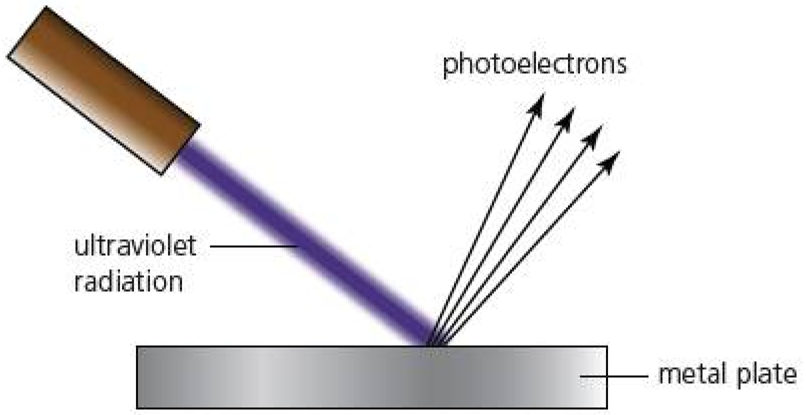


Figure 13.1 The photoelectric effect – a stream of photoelectrons is emitted from a metal surface illuminated with ultraviolet radiation

When electromagnetic radiation is directed onto a clean surface of some metals, electrons are ejected. This is called the **photoelectric effect** (Figure 13.1) and the ejected electrons are known as **photoelectrons**. Under suitable circumstances the photoelectric effect can occur with visible light, X-rays and gamma rays, but it is most often demonstrated with ultraviolet radiation and zinc. Figure 13.2 on page 436 shows a typical arrangement.

Ultraviolet radiation is shone onto a zinc plate attached to a digital coulombmeter (which measures very small quantities

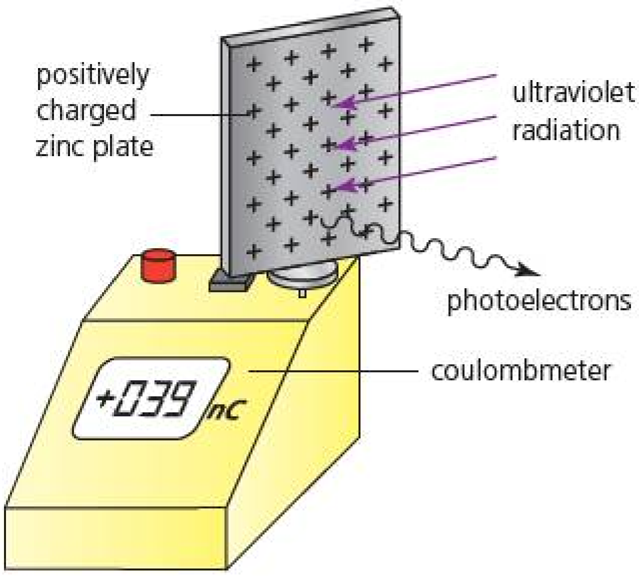


Figure 13.2 Demonstration of the photoelectric effect

of charge). The ultraviolet radiation causes the zinc plate to become positively charged because some negatively charged electrons on the (previously neutral) zinc plate have gained enough kinetic energy to escape from the surface.

Simple investigations of the photoelectric effect show a number of key features.

- If the intensity of the radiation is increased the charge on the plate increases more quickly because more photoelectrons are being released every second.
- There is no time delay between the radiation reaching the metal surface and the emission of photoelectrons. The release of photoelectrons from the surface is *instantaneous*.
- The photoelectric effect can only occur if the frequency of the radiation is above a certain *minimum* value. The lowest frequency for emission is called the **threshold frequency**, f_0. (Alternatively we could say that there is *maximum* wavelength above which the effect will not occur.) If the frequency used is below the threshold frequency, the effect will not occur, *even if the intensity of the radiation is greatly increased*. The threshold frequency of zinc, for example, is 1.04×10^{15} Hz, which is in the ultraviolet part of the spectrum. Visible light will not release photoelectrons from zinc (or other common metals).
- For a given incident frequency the photoelectric effect occurs with some metals, but not with others. This because different metals have different threshold frequencies.

Explaining the photoelectric effect

13.1.2 **Describe** the concept of the photon, and use it to **explain** the photoelectric effect.

If we use the wave theory of radiation to make predictions about the photoelectric effect, we would expect the following. (1) Radiation of *any* frequency will cause the photoelectric effect if the intensity is made high enough. (2) There may be a delay before the effect begins because it needs time for enough energy to be provided (like heating water until it boils).

These predictions are *wrong*, so an alternative theory is needed. Einstein realized that we cannot explain the photoelectric effect without first understanding the quantum nature of radiation.

Photons

The German physicist Max Planck was the first to propose (in 1900) that the energy transferred by electromagnetic radiation was in the form of a very large number of separate, individual amounts of energy (rather than continuous waves). These discrete 'packets' of energy are called quanta. Quanta are also commonly called **photons**. (The concept of photons was introduced in Chapter 7.)

This very important theory, developed further by Albert Einstein in the following years, essentially describes the nature of electromagnetic radiation as being *particles*, rather than *waves*. ('Wave–particle duality' is discussed later in this chapter.) Explaining the photoelectric effect played a central role in the early development of the photon theory.

The energy, E, carried by each photon (quantum) is given by the following relationship:

$$E = hf$$

This equation is in the IB *Physics data booklet*.

where f represents the frequency of the electromagnetic radiation and h represents a constant called **Planck's constant** (Chapter 7).

The value of Planck's constant is 6.63×10^{-34} J s. (A value for this *fundamental constant* is given in the IB *Physics data booklet*.)

Since $c = f\lambda$, this equation is also commonly expressed in term of wavelength as follows:

$$E = \frac{hc}{\lambda}$$

Photons travel at the speed of light and have zero rest mass. However, photons can transfer energy *and* momentum during interactions with subatomic particles, which indicates that photons have momentum themselves.

Using the wave theory of radiation, greater intensity is explained by waves of greater amplitude. Using a photon model of radiation, greater intensity is explained simply by having more photons (of the same energy).

We know that different parts of the electromagnetic spectrum have very different properties. This can be mostly explained by understanding that the amount of energy carried by photons can vary enormously. For example, a gamma ray photon has a frequency about one million times greater than a visible light photon. Therefore, gamma ray photons each transfer one million times as much energy as light photons, so when each individual photon is absorbed it can be dangerous to living organisms.

Within the visible spectrum, blue light has a higher frequency than red light and each of its photons carries more energy than photons of red light.

Worked examples

1 Determine the energy of a photon of light with a wavelength of 500 nm in joules and in electronvolts.

$$E = \frac{hc}{\lambda}$$

$$E = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{500 \times 10^{-9}} = 3.98 \times 10^{-19}\,\text{J}$$

$$E = \frac{3.98 \times 10^{-19}}{1.60 \times 10^{-19}} = 2.49\,\text{eV}$$

2 Electrical power of 60 watts is supplied to a lamp which radiates light of wavelength 590 nm. The lamp is 30% efficient. Determine the number of photons of visible light emitted per second by the lamp.

$$E = \frac{hc}{\lambda}$$

$$E = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{590 \times 10^{-9}} = 3.37 \times 10^{-19}\,\text{J}$$

$$\text{Number of photons emitted per second} = \frac{\text{useful power}}{\text{energy}} = \frac{0.30 \times 60}{3.37 \times 10^{-19}} = 5.3 \times 10^{19}$$

1 Calculate the number of photons emitted every second from a mobile phone operating at a frequency of 850 MHz and at a radiated power of 780 mW.

2 What is the approximate ratio of the energy of a photon of blue light to the energy of a photon of red light?

3 A detector of very low intensity light receives a total of 3.32×10^{-17} J from light of wavelength 600 nm. Calculate the number of photons received by the detector.

4 **a** Calculate an approximate value for the energy of an X-ray photon in joules and electronvolts.
b Suggest a reason why exposure to X-rays of low intensity for a short time is dangerous, but relatively high continuous intensities of visible light cause us no harm.

Einstein model

The Einstein model explains the photoelectric effect using the concept of photons. When a photon in the incident radiation interacts with an electron in the metal surface, it transfers *all* of its energy to that electron. It should be stressed that a *single* photon can only interact with a *single* electron and this transfer of energy is *instantaneous*; there is no need to wait for a build-up of energy. If a photoelectric effect is occurring, increasing the intensity of the radiation only increases the number of photoelectrons, not their energies.

Einstein realized that some of the energy carried by the photon (and then given to the electron) was used to overcome the attractive forces that normally keep an electron within the metal surface. The remaining energy is transferred to the kinetic energy of the newly released (photo) electron. Using the principle of conservation of energy, we can write:

$$\text{energy carried by photon} = \text{work done in removing the electron from the surface} + \text{kinetic energy of photo electron}$$

But, the energy required to remove different electrons is *not* always the same. It will vary with the position of the electron with respect to the surface. Electrons closer to the surface will require less energy to remove them. However, there is a well-defined *minimum* amount of energy needed to remove an electron, and this is called the **work function** of the metal surface. The symbol ϕ is used for work function. Different metals have different values for their work function. For example, the work function of a clean zinc surface is 4.3 eV. This means that *at least* 4.3 eV ($= 6.9 \times 10^{-19}$ J) of work has to be done to remove an electron from zinc.

To understand the photoelectric effect we need to compare the photon's energy, hf, to the work function, ϕ, of the metal:

- $hf < \phi$ If an incident photon has less energy than the work function of the metal, the photoelectric effect cannot occur. This means that the same radiation may cause the photoelectric effect with one metal, but not with another (which has a different work function).
- $hf_0 = \phi$ At the *threshold frequency*, f_0, the incident photon has exactly the same energy as the work function of the metal. We may assume that the photoelectric effect occurs, but the released photon will have zero kinetic energy.
- $hf > \phi$ If an incident photon has more energy than the work function of the metal, the photoelectric effect occurs and a photoelectron will be released. Photoelectrons produced by different photons (of the same frequency) will have a range of different kinetic energies because different amounts of work will have been done to release them.

It is important to consider the situation in which the *minimum* amount of work is done to remove an electron (equal to the work function):

energy carried by photon = work function + *maximum* kinetic energy of photoelectron

Or in symbols:

$$hf = \phi + E_{max}$$

This equation is often called Einstein's photoelectric equation, and it is given in the IB *Physics data booklet*. Since $hf_0 = \phi$, we can also write this as:

$$hf = hf_0 + E_{max}$$

Figure 13.3 shows a graphical representation of how the maximum kinetic energy of the emitted photons varies with the frequency of the incident photons. The equation of the line is $E_{max} = hf - \phi$.

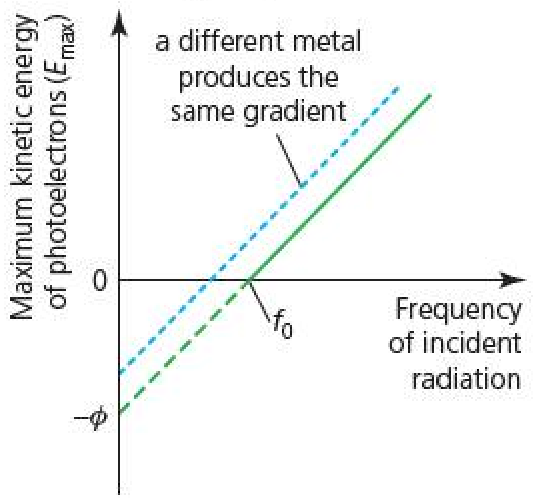


Figure 13.3 Theoretical variation of maximum kinetic energy of photoelectrons with incident frequency (for two different metals)

We can take the following measurements from this graph:

- The gradient of the line is equal to Planck's constant, h. (Compare the equation of the line to $y = mx + c$.) Clearly the gradient is the same for all circumstances because it does not depend on the frequency or the metal used.
- The intercept on the frequency axis gives us the value of the threshold frequency, f_0.
- A value for the work function can be determined from:
  i when $E_{max} = 0$, $\phi = hf_0$; or
  ii when $f = 0$, $\phi = -E_{max}$

Worked example

3 Radiation of wavelength 5.59×10^{-8} m was incident on a metal surface which had a work function of 2.71 eV.
a What was the frequency of the radiation?
b How much energy is carried by one photon of the radiation?
c What is the value of the work function expressed in joules?
d Did the photoelectric effect occur under these circumstances?
e What was the maximum kinetic energy of the photoelectrons?
f What is the threshold frequency for this metal?
g Sketch a fully labelled graph to show how the maximum kinetic energy of the photoelectrons would change if the frequency of the incident radiation was varied.

a $f = \dfrac{c}{\lambda}$

$$f = \frac{3.00 \times 10^8}{5.59 \times 10^{-8}}$$

$f = 5.37 \times 10^{15}$ Hz

b $E = hf$

$E = (6.63 \times 10^{-34}) \times (5.37 \times 10^{15})$

$E = 3.56 \times 10^{-18}$ J

c $2.71 \times (1.60 \times 10^{-19}) = 4.34 \times 10^{-19}$ J

d Yes, because the energy of each photon is greater than the work function.

e $E_{max} = hf - \phi$

$E_{max} = (3.56 \times 10^{-18}) - (4.34 \times 10^{-19}) = 3.13 \times 10^{-18}$ J

f $\phi = hf_0$

$$f_0 = \frac{\phi}{h}$$

$$f_0 = \frac{4.34 \times 10^{-19}}{6.63 \times 10^{-34}} = 6.54 \times 10^{14}\ \text{Hz}$$

g The graph should be similar to Figure 13.3, with numerical values provided for the intercepts.

13.1.4 **Solve** problems involving the photoelectric effect.

5 Repeat Worked example 3 above, but for radiation of wavelength 6.11×10^{-7} m incident upon a metal with a work function of 2.21 eV. Omit part **e**.

6 **a** Explain how Einstein used the concept of photons to explain the photoelectric effect.
b Explain why a wave model of electromagnetic radiation is unable to explain the photoelectric effect.

7 The threshold frequency of a metal is 7.0×10^{14} Hz. Calculate the maximum kinetic energy of the electrons emitted when the frequency of the radiation incident on the metal is 1.0×10^{15} Hz.

8 **a** The longest wavelength that emits photoelectrons from potassium is 550 nm. Calculate the work function (in joules).
b What is the threshold wavelength for potassium? What is the name for this kind of radiation?
c Name one colour of visible light that will *not* produce the photoelectric effect with potassium.

9 When electromagnetic radiation of frequency 2.90×10^{15} Hz is incident upon a metal surface the emitted photoelectrons have a maximum kinetic energy of 9.70×10^{-19} J. Calculate the threshold frequency of the metal.

Experiments to test the Einstein model

13.1.3 Describe and **explain** an experiment to test the Einstein model.

Investigating stopping potentials

To test Einstein's equation (model) for the photoelectric effect, it is necessary to measure the maximum kinetic energy of the photoelectrons emitted under a variety of different circumstances. In order to do this, the kinetic energy must be transferred to another (measurable) form of energy.

The kinetic energy of the photoelectrons can be transferred to electric potential energy if they are repelled by a negative voltage (potential). This experiment was first performed by the American physicist Robert Millikan and a simplified version is shown in Figure 13.4.

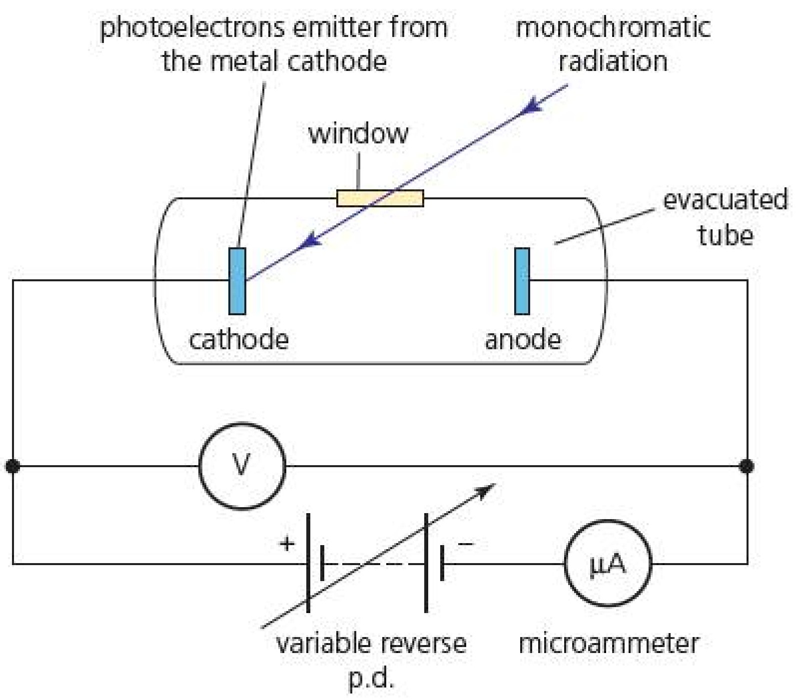


Figure 13.4 Experiment to test Einstein's model of photoelectricity

Ideally *monochromatic* radiations should used, but it is also possible to use a narrow range of frequencies, such as those obtained by using coloured filters with white light.

When radiation is incident upon a suitable emitting surface, photoelectrons will be released with a range of different energies (as explained in the previous section). Because it is emitting negative charge, this surface can be described as a **cathode**. Any photoelectrons which have enough kinetic energy will be able to move across the tube and reach the other electrode, the **anode**. The tube is *evacuated* (the air is removed to create a vacuum), so that the electrons do not collide with air molecules during their movement across the tube.

The most important thing to note about this circuit is that the (variable) source of p.d. is connected the 'wrong way around', it is supplying a **reverse potential difference** across the tube. This means that there is a negative voltage (potential) on the anode which will *repel* the photoelectrons. Photoelectrons moving towards the anode will have their kinetic energy reduced as it is transferred to electric potential energy. (Measurements for positive potential differences can be made by reconnecting the battery around the 'correct' way.)

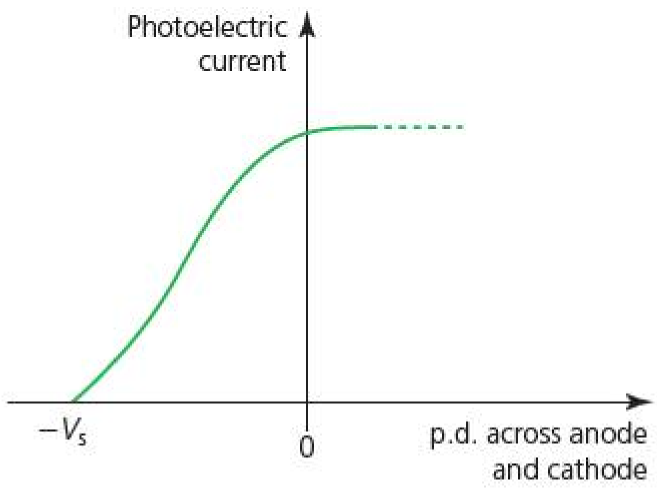


Figure 13.5 Increasing the reverse p.d. decreases the photoelectric current

Any flow of charge across the tube and around the circuit can be measured by a sensitive ammeter (microammeter or picoammeter). When the reverse potential on the anode is increased from zero, more and more photoelectrons will be prevented from reaching the anode and this will decrease the current. (Remember that the photoelectrons have a range of different energies.) Eventually the potential will be great enough to stop even the most energetic of photoelectrons, and the current will fall to zero (Figure 13.5)

The potential on the anode needed to just stop all photoelectrons reaching it is called the **stopping potential**, V_s.

Since, by definition, potential difference = energy transferred/charge,

$$\text{stopping potential, } V_s = \frac{\text{maximum kinetic energy of photoelectrons, } E_{max}}{\text{charge on electron, } e}$$

After measuring V_s we can use this equation to calculate values for the maximum kinetic energy of photoelectrons under a range of different circumstances:

$$E_{max} = eV_s$$

For convenience, it is common to quote all energies associated with the photoelectric effect in electronvolts (eV). In this case, the maximum kinetic energy of the photoelectrons is numerically equal to the stopping potential. That is, if the stopping potential is 3 V, then $E_{max} = 3\,eV$.

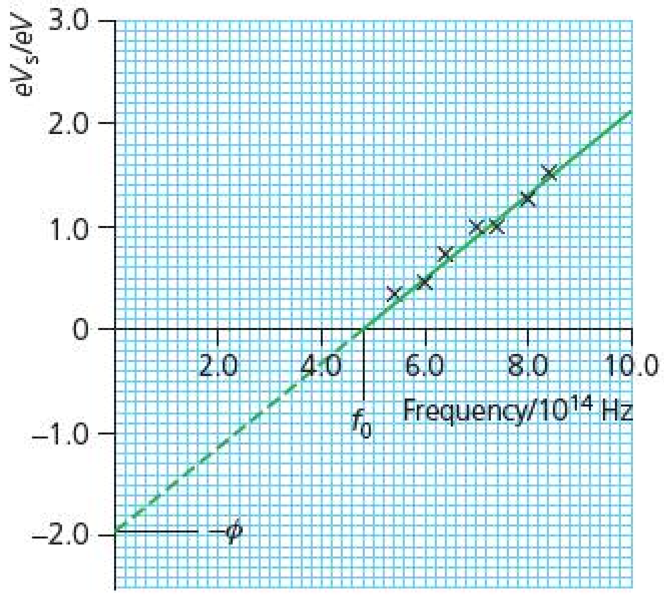


Figure 13.6 Experimental results showing variation of maximum potential energy (eV_s) of photoelectrons with incident frequency

Einstein's equation ($hf = \phi + E_{max}$) can now be re-written as:

$$hf = \phi + eV_s$$

Or, since $\phi = hf_0$,

$$hf = hf_0 + eV_s$$

This equation is given in the IB *Physics data booklet* (although V is used instead of V_s).

By *experimentally* determining the stopping potential for a range of different frequencies the *theory* shown previously in Figure 13.3 can now be confirmed by plotting a graph from actual data, as shown in Figure 13.6.

The threshold frequency, f_0, can be determined from the intercept on the frequency axis.

The work function, ϕ, can be calculated from $\phi = hf_0$, or from the intercept on the eV_s axis.

A value for Planck's constant, h, can be determined from the gradient.

13.1.4 Solve problems involving the photoelectric effect.

10 Calculate the maximum kinetic energy of photoelectrons emitted from a metal if the stopping potential was 2.4 eV. Give your answer in joules and in electronvolts.

11 Make a copy of Figure 13.5 and add lines to show the results that would be obtained with:
- **a** the same radiation, but with a metal of greater work function
- **b** the original metal and the same frequency of radiation, but using radiation with a greater intensity.

12 In an experiment using monochromatic radiation of frequency 7.93×10^{14} Hz with a metal that had a threshold frequency of 6.11×10^{14} Hz, it was found that the stopping potential was 0.775 V. Calculate a value for Plank's constant from these results.

Investigating photoelectric currents

Using the apparatus shown in Figure 13.4, it is also possible to investigate quantitatively the effects on the photoelectric current of changing the intensity, the frequency, and the metal used in the cathode.

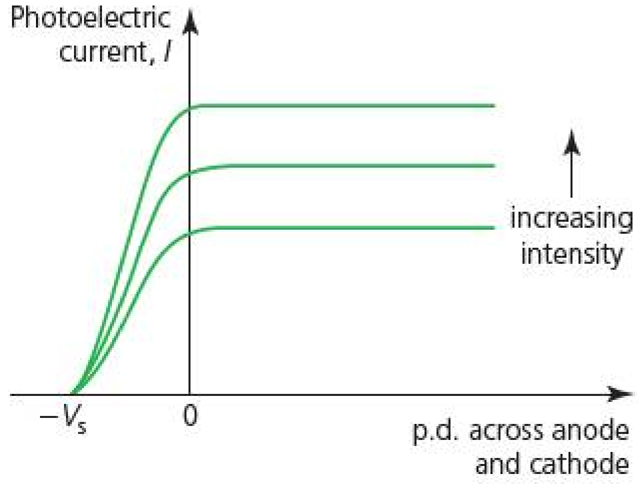


Figure 13.7 Variation of photoelectric current with p.d. for radiation of three different intensities (same frequency)

- **Intensity** – Figure 13.7 shows the photoelectric currents produced by monochromatic radiation of the same frequency at three different intensities.

  For positive potentials, the photoelectric currents remain constant because the photoelectrons are reaching the anode at the same rate as they are being produced at the cathode, and this does not depend on the size of the positive potential on the anode.

  Greater intensities (of the same frequency) produce greater photoelectric currents because there are more photons releasing more photoelectrons (of the same range of energies).

  Since the maximum kinetic energy of photons depends only on frequency and not intensity, all of these graphs have the same value for stopping potential, V_s.

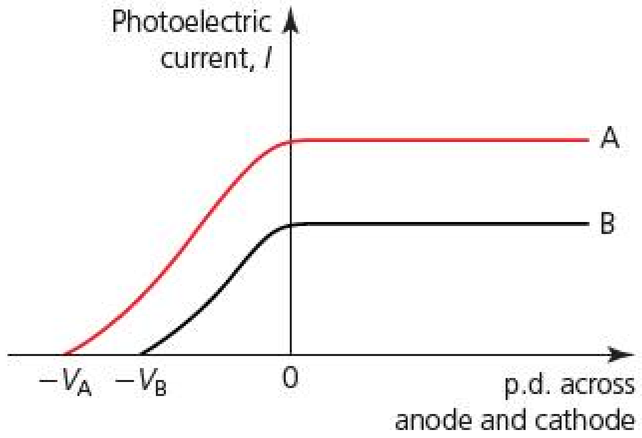


Figure 13.8 Variation of photoelectric current with p.d. for radiation of two different frequencies

- **Frequency** – Figure 13.8 shows the photoelectric currents produced by radiation from two monochromatic sources, of different frequencies, A and B, incident on the same metal.

  The individual photons in radiation A must have more energy (than B) and produce photoelectrons with a greater maximum kinetic energy. We know this because a greater reverse potential is needed to stop the more energetic photoelectrons produced by A.

  No conclusion can be drawn from the fact that the current for A has been drawn higher than for B, because the intensities of the two radiations are not known. In the unlikely circumstances that the two intensities were equal, the maximum current for B would have to be higher than for A because the radiation from B must have more photons, since each photon has less energy than in A.

- **Metal used in the cathode** – Experiments confirm that when different metals are tested using the same frequency, the photoelectric effect is observed with some metals, but not with others (for which the metal's work function is greater than the energy of the photons).

13.1.4 **Solve** problems involving the photoelectric effect.

13 Make a copy of Figure 13.5 and add the results that would be obtained using radiation of greater intensity (of the same frequency) incident on a metal which has a smaller work function.

14 Make a copy of Figure 13.8, line A only. Add to it a line showing the results that would be obtained with radiation of a higher frequency, but with same number of photons every second incident upon the metal.

The wave nature of matter

The de Broglie hypothesis and matter waves

13.1.5 **Describe** the de Broglie hypothesis and the concept of matter waves.

In 1924 the French physicist Louis de Broglie proposed that electrons, which were thought of as particles, might also have a wave-like character. He later generalized his hypothesis to suggest that *all* moving particles have a wave-like nature.

According to de Broglie, the wavelength, λ, of a moving particle is related to its momentum, p, by the equation:

$$\lambda = \frac{h}{p}$$

where h is Planck's constant. This is known as **de Broglie's equation**. Rearranging, this becomes:

$$p = \frac{h}{\lambda}$$

This equation is in the IB *Physics data booklet.*

Worked examples

4 Calculate the momentum of a moving particle which has a de Broglie wavelength of 200 pm (1 pm = 1×10^{-12} m).

$$p = \frac{h}{\lambda}$$

$$p = \frac{6.63 \times 10^{-34}}{2.00 \times 10^{-10}} = 3.32 \times 10^{-24}\,\text{kg m s}^{-1}$$

5 Calculate the de Broglie wavelength of a 1.00 g mass moving at a speed of 1.00 m per year.

$$v = \frac{1.00}{365 \times 24 \times 60 \times 60} = 3.17 \times 10^{-8}\,\text{m s}^{-1}$$

$$p = \frac{h}{\lambda}; \quad \lambda = \frac{6.63 \times 10^{-34}}{1.00 \times 10^{-3} \times 3.17 \times 10^{-8}} = 2.09 \times 10^{-23}\,\text{m}$$

The de Broglie wavelength for moving objects of ordinary size is very small because of their large mass. For this reason, the so-called **matter waves** of objects that are large enough to be seen cannot be detected. However, sub-atomic particles, like electrons, are so small that wave-like properties can be detected experimentally.

Matter waves, like electromagnetic waves, can travel in a vacuum. Unlike electromagnetic waves, matter waves are not produced by accelerated charges. Matter waves are probability waves (see page 452). The greater the amplitude of the wave, the greater the probability that the particle exists at a particular point.

The wavelength of de Broglie waves can also be related to the kinetic energy of the particle:

$$E_K = \frac{1}{2}mv^2$$

$$mE_K = \frac{1}{2}m^2v^2$$

$$m^2v^2 = 2mE_K$$

$$mv = \sqrt{2mE_K}$$

But mv represents momentum, p, so:

$$p = \sqrt{2mE_K}$$

Substituting this value of p into the de Broglie relationship, we obtain the following expression:

$$\lambda = \frac{h}{\sqrt{2mE_K}}$$

If a charged particle carrying a charge of q coulombs is accelerated from rest by applying a potential difference of V volts, then the kinetic energy of the particle is given by:

$$E_K = qV$$

Substituting qV in the equation for wavelength gives:

$$\lambda = \frac{h}{\sqrt{2mqV}}$$

The de Broglie hypothesis is the basis for the Schrödinger model of the hydrogen atom (see page 450). The electron of the hydrogen atom forms a standing wave around the central proton (nucleus).

15 Explain the statement, 'matter and radiation have dual character'.

16 **a** Derive the relationship between the wavelength of the de Broglie wave and the kinetic energy of the particle.
b If you double the kinetic energy of an electron (at speeds well below the speed of light), how does its de Broglie wavelength change?
c What happens to the wavelength if you double the speed of the electron?

17 Construct a spreadsheet that converts values of mass and velocity into momentum and wavelength (via the use of the de Broglie equation).

The Davisson–Germer experiment

13.1.6 Outline an experiment to verify the de Broglie hypothesis.

We have seen that de Broglie's hypothesis suggests that electrons have wave-like properties. They should therefore be able to undergo interference and diffraction.

Experimental confirmation of de Broglie's hypothesis and the wave nature of electrons was first obtained in 1927 when Clinton Davisson and Lester Germer showed that an electron beam

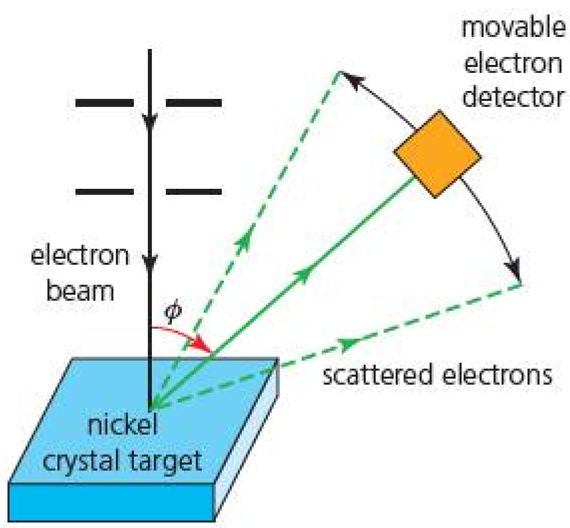


Figure 13.9 The principle behind the Davisson and Germer experiment

could be diffracted by a metal crystal (Figure 13.9). They used a beam of electrons fired at a nickel crystal target and recorded the electrons scattered at different angles. They found that the intensity of the scattered electrons varied with the angle and also depended on their speed (which could be altered by changing the accelerating potential difference).

Work had been done on metal crystals using X-rays, so the spacing between layers of atoms in the nickel crystal was known. The angle at which the maximum scattered intensity was recorded agreed with the angle predicted by constructive interference of waves by layers of atoms in the surface of the metal crystal. Accurate measurements of this diffraction pattern demonstrated that de Broglie's equation was correct.

The diffraction of electrons can be demonstrated in the school laboratory using a high voltage evacuated tube. A beam of electrons is accelerated in an electron gun to a high potential and then directed onto a very thin sheet of graphite (carbon atoms) (Figure 13.10). The atomic spacing of carbon atoms in the graphite is in the same order as the predicted wavelength of the electron (when travelling at very high velocity).

The electrons diffract (Chapter 11) from the carbon atoms and form a pattern on the screen of a series of concentric rings (Figure 13.10). This observed diffraction of electrons is strong experimental evidence for the wave nature of electrons. The characteristic pattern is due to the regular spacing of the carbon atoms in the different layers of graphite. A very thin sample of graphite has to be used to obtain sharp diffraction rings. If the carbon sample is too thick, the electrons lose kinetic energy and their wavelength becomes longer, smearing out the diffraction pattern.

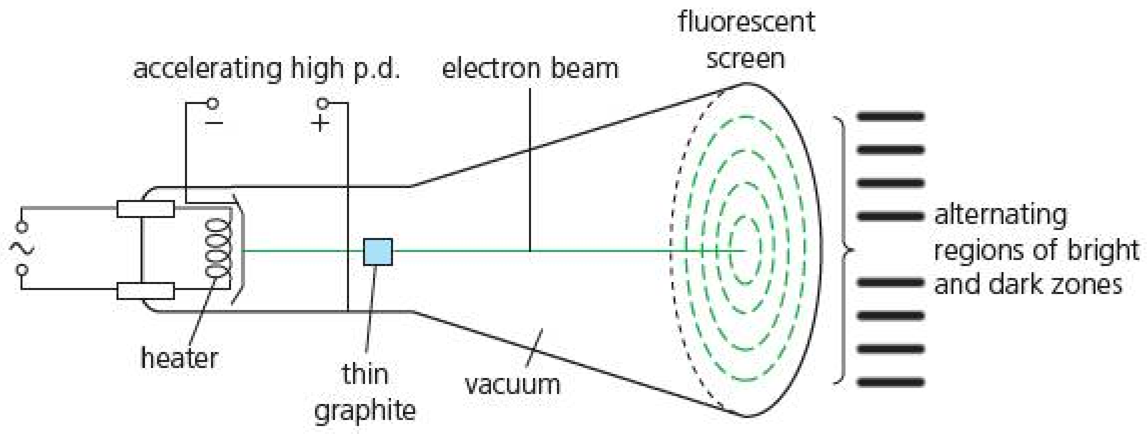


Figure 13.10 An electron diffraction apparatus

18 Describe the Davisson–Germer experiment and explain its importance.

19 Find out about neutron diffraction and its use in establishing the structure of substances, including proteins.

TOK Link: Nature of light

The photoelectric effect clearly demonstrates that light can behave as a stream of particles. However, its ability to undergo diffraction clearly demonstrates its wave-like nature. Is light a wave or particle? Whether a particle or wave model is useful depends on the nature of the experiment, but neither model is a complete description of light behaviour. Physicists often use the wave model when light is travelling through a medium, but use a particle model when light interacts with matter. Neither model is consistent with all observations. This dual nature of light is termed **wave–particle duality**.

Question

1 Try and explain the concept of wave–particle duality to a student who does not study physics. Try and think of some analogies to help convey the concept.

Calculations involving matter waves

13.1.7 Solve problems involving matter waves.

Worked examples

6 Calculate the de Broglie wavelength of an electron travelling with a speed of $1.0 \times 10^7\,\text{m}\,\text{s}^{-1}$. (Planck's constant $= 6.63 \times 10^{-34}\,\text{J}\,\text{s}$; electron mass, $m_e = 9.110 \times 10^{-31}\,\text{kg}$.)

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34}}{(9.110 \times 10^{-31}) \times (1.0 \times 10^7)} = 7.3 \times 10^{-11}\,\text{m}$$

7 Estimate the de Broglie wavelength of a ball of mass 0.058 kg moving with a velocity of $10^2\,\text{m}\,\text{s}^{-1}$.

$p = mv$; $p = 0.058 \times 10^2 = 5.8\,\text{kg}\,\text{m}\,\text{s}^{-1}$

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34}}{5.8} \approx 10^{-34}\,\text{m}$$

To exhibit wave-like properties the ball would need to interact with an object with dimensions of the order of 10^{-34} m (over one million million million times smaller than the nucleus of an atom). Hence, the ball does not exhibit any measurable or detectable wave-like properties.

8 Calculate the de Broglie wavelength of an electron which has been accelerated through a potential difference of 500 V.

The speed of the electrons can be calculated, using the law of conservation of energy:
loss of electrical potential energy = gain in kinetic energy

$$Ve = \frac{1}{2}mv^2$$

$$v = \sqrt{\frac{2Ve}{m}} = \sqrt{\frac{2 \times 500 \times 1.60 \times 10^{-19}}{9.110 \times 10^{-31}}} = 1.32 \times 10^7\,\text{m}\,\text{s}^{-1}$$

$$p = mv = 1.2 \times 10^{-23}\,\text{kg}\,\text{m}\,\text{s}^{-1}$$

$$\lambda = \frac{h}{p} = 5.5 \times 10^{-11}\,\text{m}$$

9 Calculate the de Broglie wavelength of a lithium nucleus $\left({}^7_3\text{Li}\right)$ which has been accelerated through a potential difference of 5.00 MV. (The mass of a lithium-7 nucleus is 1.165×10^{-26} kg.)

$$\frac{1}{2}mv^2 = 3e \times V$$

$$v^2 = \frac{6 \times (1.60 \times 10^{-19}) \times (5.00 \times 10^6)}{(1.165 \times 10^{-26})} = 4.12 \times 10^{14}$$

$$v = 2.03 \times 10^7\,\text{m}\,\text{s}^{-1}$$

$$\lambda = \frac{6.63 \times 10^{-34}}{(1.165 \times 10^{-26}) \times (2.03 \times 10^7)} = 2.81 \times 10^{-15}\,\text{m}$$

20 Calculate the wavelength (in metres) of an electron moving with a velocity of $2.05 \times 10^7\,\text{m}\,\text{s}^{-1}$.

21 A neutron has a de Broglie wavelength of 800 pm. Calculate the velocity of the neutron. The mass of the neutron is 1.675×10^{-27} kg.

22 The mass of an electron is 9.110×10^{-31} kg. If its kinetic energy is 3.0×10^{-25} J, calculate its de Broglie wavelength in metres.

23 Over what potential difference do you have to accelerate electrons for them to have a wavelength of 1.2×10^{-10} m?

24 **a** Which is associated with a de Broglie wavelength of longer wavelength – a proton or an electron travelling at the same velocity? Explain your answer.
b The following (non-relativistic particles) all have the same kinetic energy. Rank them in order of their de Broglie wavelengths, greatest first: electron, alpha particle, neutron and gold nucleus.

25 Explain why a moving airplane has no detectable wave properties.

Atomic spectra and atomic energy states

13.1.8 **Outline** a laboratory procedure for producing and observing atomic spectra.

Emission spectra

Figure 13.11 The flame test for potassium ions from potassium chloride

When a sample of an element is excited by heating (the flame test – Figure 13.11) or by passing an electric current through its gas (at low pressure), the atoms of the element emit electromagnetic radiations of definite frequencies. These characteristic frequencies form the **emission spectrum** of the element. Since the radiations in the spectrum are emitted due to energy changes taking place in the atoms, this spectrum is known as an atomic emission spectrum. An atomic emission spectrum for an element consists of a unique series of coloured lines on a black background.

The instrument used for obtaining and analysing an emission spectrum is known as a **spectrometer**. A simple spectrometer (Figure 13.12) works by passing a beam of light from excited gaseous atoms through a lens via a slit to a second lens and then to a prism (or diffraction grating). This separates the **spectral lines**, which can be observed and measured using an eyepiece (or travelling microscope).

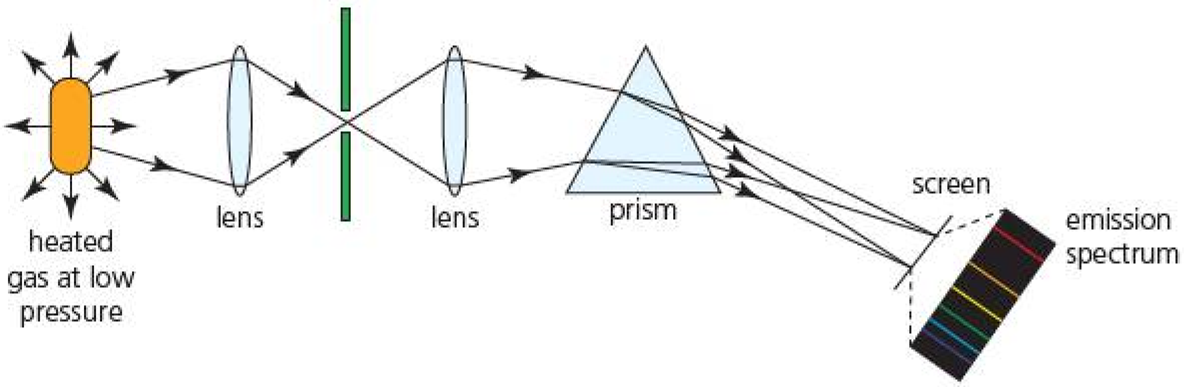


Figure 13.12 Schematic diagram showing the production of an atomic emission spectrum

Different elements have different **energy levels** for their electrons. In a gas excited by heating or by being subjected to electric discharge, electrons in the excited atoms 'fall back' to lower energy levels. They do this by emitting photons of specific frequencies seen in the **line spectrum**. This provides useful information for physicists to identify elements and study their atomic structures.

Absorption spectra

Cool gaseous atoms (and molecules) absorb electromagnetic radiation as well as emitting it. When white light is shone through a sample of a gaseous element (at low pressure and low temperature), the light that emerges has the element's characteristic wavelengths (or frequencies) missing (Figure 13.13). There are dark lines or bands in the discontinuous spectrum.

For each element, the position of these dark bands is the same as the position of the lines in the emission spectrum. An element's characteristic lines or frequencies is called its **absorption spectrum.**

In the formation of an absorption spectrum, the wavelengths that are absorbed are re-emitted, but in all directions, so the original beam of light has a very much reduced (but not zero) intensity of light at these wavelengths.

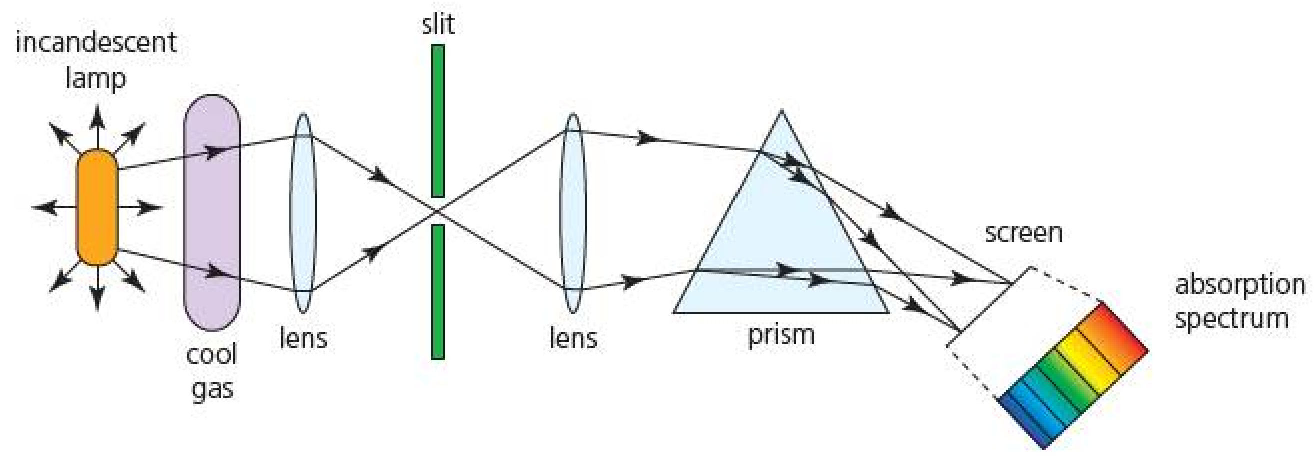


Figure 13.13 Schematic diagram showing the production of an atomic absorption spectrum

26 **a** You are given samples of argon, neon and helium gases. All these elements are colourless gases at room temperature. Outline laboratory procedures for producing and observing emission and absorption atomic spectra for these gases.

b Use the Internet to find the emission and absorption spectra of the three gases. Sketch the visible regions of their emission and absorption spectra.

Quantized energy in atoms

13.1.9 **Explain** how atomic spectra provide evidence for the quantization of energy in atoms.

We have seen that atomic emission and absorption spectra have discrete frequencies corresponding to specific energies. The Bohr model of the hydrogen atom (Chapter 7) can be used to explain how these spectral lines arise as a result of quantized electron energies.

In Bohr's model, electrons can only exist in certain energy levels. Electrons in these energy levels behave like stationary waves fitted into the circumference of the orbit. Energy levels are labelled $n = 1, 2, 3 \ldots$, etc. This is shown in Figure 13.14 for $n = 4$ and $n = 8$. Electrons can only have these energies, but not energies in between.

A 'free' electron at $n = \infty$ has zero energy (the hydrogen atom has become an ion, and is said to be ionized). The energy level at $n = 1$ has the lowest energy: –13.6 eV for a hydrogen atom. A hydrogen atom with its electron in the $n = 1$ energy level is in the **ground state**. All energy values on this scale are negative, indicating a loss of energy as the electron moves closer to the nucleus (with decreasing values of n) (Figure 13.15).

For an electron to move from one energy level to another, energy must be absorbed or given out. Transitions between energy levels can be represented by vertical arrows drawn between the energy levels. (Figures 13.16a and 13.16b illustrate this for the important example of hydrogen.)

When a gas is heated, thermal energy raises some electrons to a higher energy level, resulting in the production of an **excited state**. An excited electron in an upper energy level will fall back to a lower energy level, usually after a very short time interval. The downward electron transition corresponds to the emission of a photon whose energy is the same as the energy *difference* between the energy levels. This is given by the relationship:

$$hf = E_2 - E_1$$

where E_2 is the energy of the higher energy level and E_1 is the energy of the lower energy level.

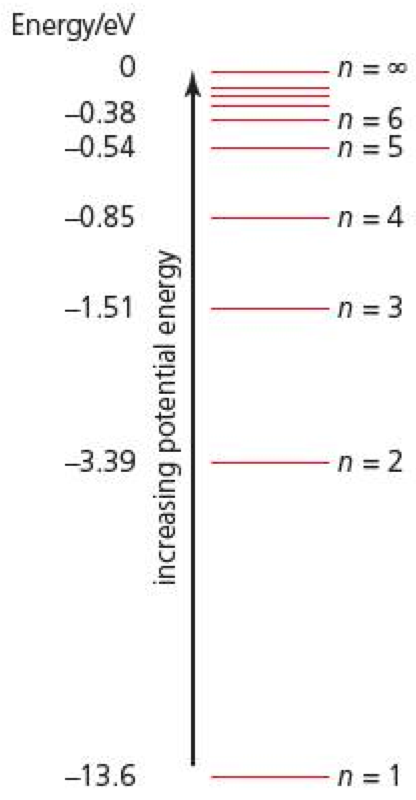


Figure 13.15 Energy levels for the Bohr model of the hydrogen atom

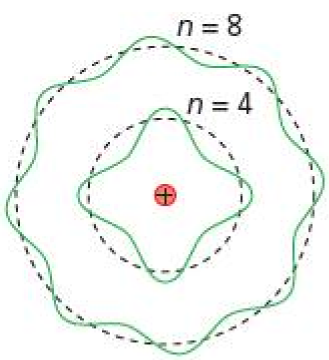


Figure 13.14 Electron waves in the fourth and eighth energy levels of the Bohr model of the hydrogen atom

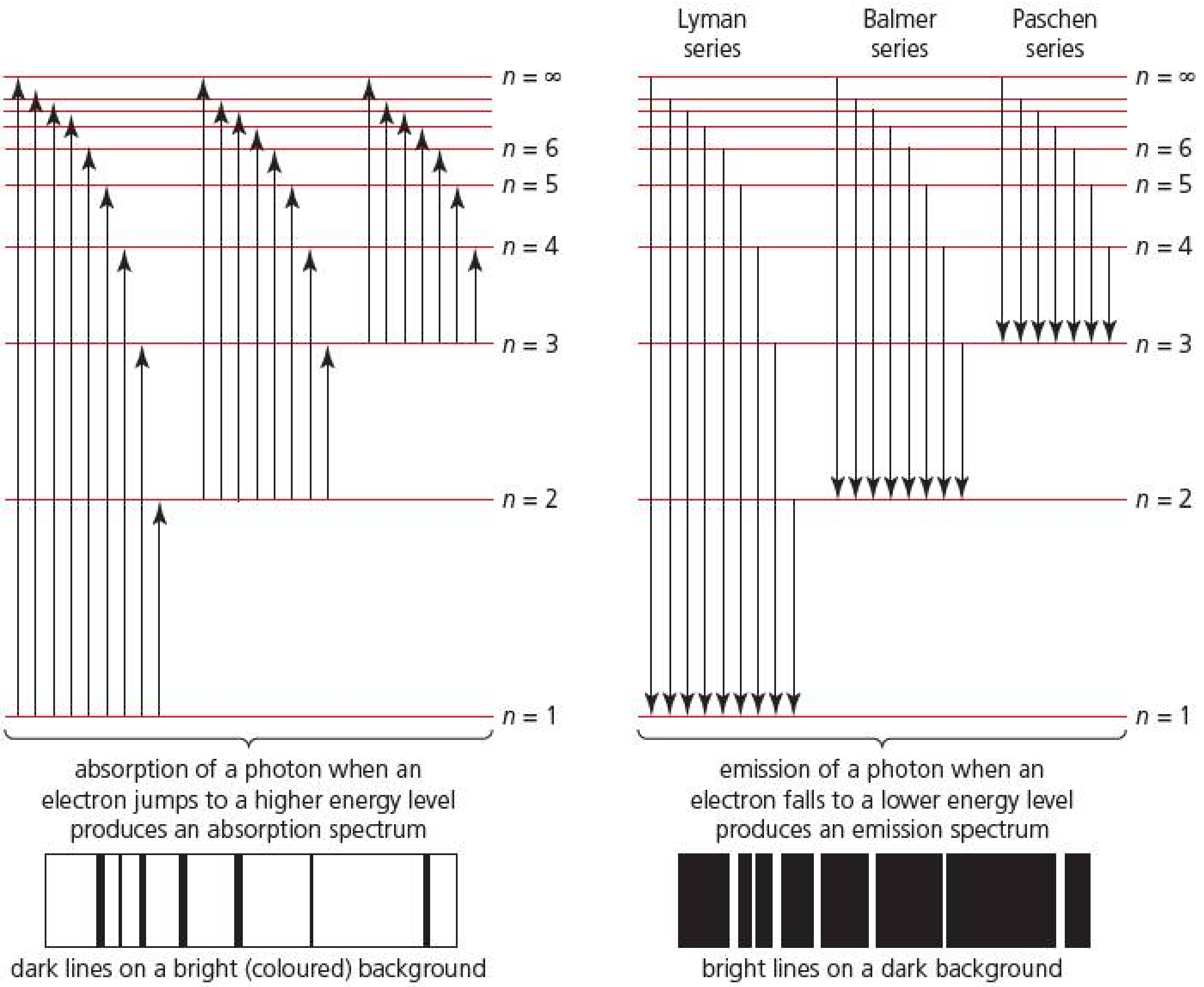


Figure 13.16 a Electron transitions from a lower to a higher energy state in a hydrogen atom **b** Electron transitions from a higher to a lower energy state in a hydrogen atom

In a similar way, photons can only be absorbed if their energy corresponds to the energy difference between two energy levels. So the lines in absorption and emission spectra have the same characteristic frequencies, as they correspond to the same energy differences.

Further away from the nucleus, the energy levels become closer and closer in energy. This arrangement of energy levels is shown in emission and absorption spectra, which show **convergence** of lines at high frequency.

Transitions between the higher energy levels, where the energy difference between levels is smaller, correspond to photons with less energy in the visible and infrared regions of the electromagnetic spectrum. A transition with a large energy difference corresponds to an X-ray photon.

27 Electron energies within atoms and molecules are quantized. Explain what this statement means.

28 a State which feature of emission and absorption spectra (of isolated gaseous atoms) provides strong experimental evidence for electron quantization.

b Account (in general terms) for the production of these features in absorption and emission spectra.

13.1.10 Calculate wavelengths of spectral lines from energy level differences and vice versa.

Calculations involving the interconversion between energy levels and wavelengths

The **Planck relationship**, $hf = E_2 - E_1$, and the wave equation, $v = f\lambda$, can be used to calculate the wavelengths (or frequencies) of spectral lines from energy level differences in atoms, or calculate energy level differences from wavelengths (or frequencies) in a spectrum.

Worked examples

10 Calculate the wavelength of the electromagnetic radiation emitted when the electron in a hydrogen atom makes a transition from the energy level at -0.54×10^{-18} J to the energy level at -2.18×10^{-18} J.
(Planck's constant $h = 6.63 \times 10^{-34}$ J s; speed of light in a vacuum, $c = 3.00 \times 10^{8}$ m s^{-1}.)

$$\Delta E = E_2 - E_1 = -0.54 \times 10^{-18} - (-2.18 \times 10^{-18}) = 1.64 \times 10^{-18}\,\text{J}$$

$$\lambda = \frac{hc}{\Delta E} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{1.64 \times 10^{-18}} = 1.21 \times 10^{-7}\,\text{m} = 121\,\text{nm}$$

11 The shortest wavelength line in the atomic spectrum of hydrogen (in the Lyman series) has a wavelength of 9.17×10^{-8} m. Calculate the energy of the transition in joules.

$$\Delta E = \frac{hc}{\lambda} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{9.17 \times 10^{-8}} = 2.17 \times 10^{-18}\,\text{J}$$

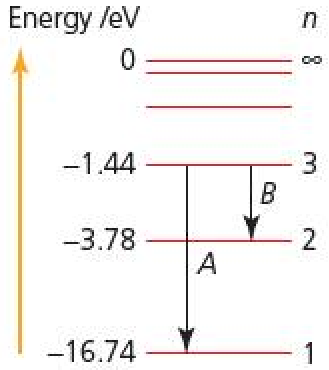


Figure 13.17

29 Figure 13.17 shows two transitions between energy levels of an atom.
a Calculate the energy change (in eV) in the electron transition labelled A.
b Calculate the wavelength of the photons (light) emitted by this transition.
c How would the wavelength for transition A compare with that for transition B? (No calculation is required.) Explain your answer.

30 Explain why the energy levels in the atom are given negative values.

The 'electron in a box' model

13.1.11 Explain the origin of atomic energy levels in terms of the 'electron in a box' model.

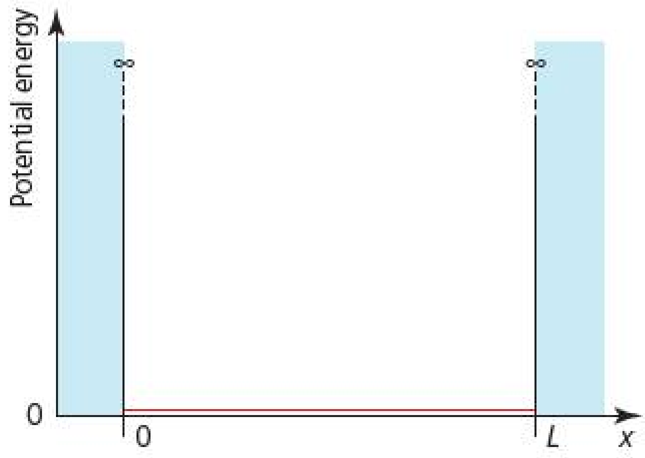


Figure 13.18 The 'electron in a box' model

The 'electron in a box' model uses the idea of the electron behaving as a wave to explain why the energy levels in atoms are quantized. In this model the electron is confined in a small region between two walls and can travel in a straight line in one dimension (along the x-axis) from $x = 0$ to $x = L$ (Figure 13.18). The potential energy of the electron inside the box is zero, but increases to infinity at the walls. The electron is trapped in a *potential well*.

If we think of the electron as behaving as a wave inside the 'box', then it must be subject to *boundary conditions*, like those applied to the waves of a violin string fixed at both ends. The waves are *standing waves* (Chapter 11) generated by the electron reflecting backwards and forwards from the sides of the 'box'. The standing waves have nodes (regions of no vibration) and antinodes (regions of maximum vibration).

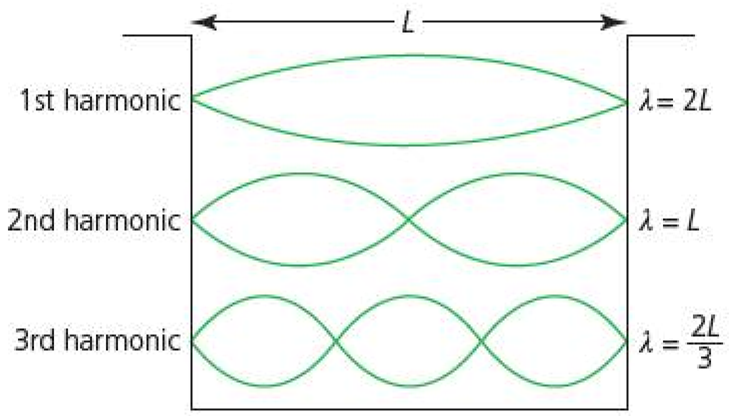


Figure 13.19 Standing electron waves within a box of length L

The walls of the box must always be nodes, so these standing waves must have an *integer* (whole) number of half wavelengths within the 'box'. This results in quantization of energy levels, just as the boundary conditions for a violin string produce particular harmonics. The wavelength of the electron wave is determined by the size of the 'box', the longest possible wavelength being $2L$ as shown in Figure 13.19 (a similar diagram is shown in Chapter 11).

The de Broglie wavelength of the electron of mass m_e is given by:

$$\lambda = \frac{h}{p} = \frac{h}{m_e v}$$

$$\text{so } v = \frac{h}{m_e \lambda}$$

The kinetic energy of the electron is $\frac{1}{2}m_e v^2$. Substituting for v in this expression gives the kinetic energy, E_K, as:

$$E_K = \frac{1}{2}m_e \times \frac{h^2}{m_e^2\lambda^2} = \frac{h^2}{2\lambda^2 m_e}$$

Figure 13.19 shows that the electron in the 'box' can have the following values of wavelength: $2L$, L, $\frac{2L}{3}$, $\frac{L}{2}$, etc. All possible standing electron waves are defined by the following equation:

$$\lambda = \frac{2L}{n}$$ where n represents an integer (1, 2, 3, 4, etc.)

Substituting this expression for the wavelength into the equation for kinetic energy gives:

$$E_K = \frac{n^2h^2}{8m_eL^2}$$

This equation is in the IB *Physics data booklet.*

Because the potential energy is zero inside the box, the total energy is the kinetic energy. As the 'box' gets smaller, the wavelengths decrease and the energy increases.

The electron in a box model is not an accurate model of electrons within an atom, but as we have seen, it clearly explains the quantized (discrete) nature of the electron's energies when the electron is treated as a wave.

Worked examples

12 Calculate the ground state energy ($n = 1$) in eV (to the nearest integer) for an electron confined to a box of length 0.50×10^{-10} m.

$$E_K = \frac{n^2h^2}{8m_eL^2} \text{ where } n = 1$$

$$E_1 = \frac{(1)^2 \times (6.63 \times 10^{-34})^2}{8 \times (9.110 \times 10^{-31}) \times (0.50 \times 10^{-10})^2}$$

$$= 2.4 \times 10^{-17}\,\text{J} = 151\,\text{eV}$$

13 An electron confined to a box has a ground state energy of 20 eV. Calculate the width (in nm) of the box.

$$E_K = \frac{n^2h^2}{8m_eL^2} \text{ where } n = 1$$

$$20 \times 1.60 \times 10^{-19} = \frac{(1)^2 \times (6.63 \times 10^{-34})^2}{8 \times (9.110 \times 10^{-31})L^2}$$

$$L = 1.4 \times 10^{-10}\,\text{m} = 0.14\,\text{nm}$$

31 a Describe the 'electron in a box' model. Explain the origin of the atomic energy levels.
b What are the allowed wavelengths for an electron in this model?

32 Calculate the kinetic energies for an electron ($m_e = 9.110 \times 10^{-31}$ kg) in the first two energy levels ($n = 1$ and $n = 2$) in a one-dimensional box of length 1.0×10^{-10} m.

The Schrödinger model of the hydrogen atom

13.1.12 Outline the Schrödinger model of the hydrogen atom.

In 1926 the Austrian physicist Erwin Schrödinger proposed a quantum mechanical model describing the behaviour of the electron within a hydrogen atom. In Schrödinger's theory (wave equation) the electron in the atom is described by a **wavefunction**, $\Psi(x, t)$, in which the amplitude psi (Ψ) of the wave is a function of position x and time t.

According to quantum mechanics, a particle cannot have a precise path; instead, there is only a probability that it may be found at a particular place at any particular time (this is discussed in more detail in the next section on the Heisenberg uncertainty principle). A wavefunction may be viewed as a blurred version of the path of an electron (Figure 13.20). The darker the area of shading, the greater the probability of finding an electron there.

If the electrostatic forces that act on the electron are known then the **Schrödinger wave equation** can be solved to obtain $\Psi(x, t)$ (Figure 13.21), which allows the energy levels to be calculated.

Figure 13.20 Classical and quantum mechanical descriptions of a moving electron

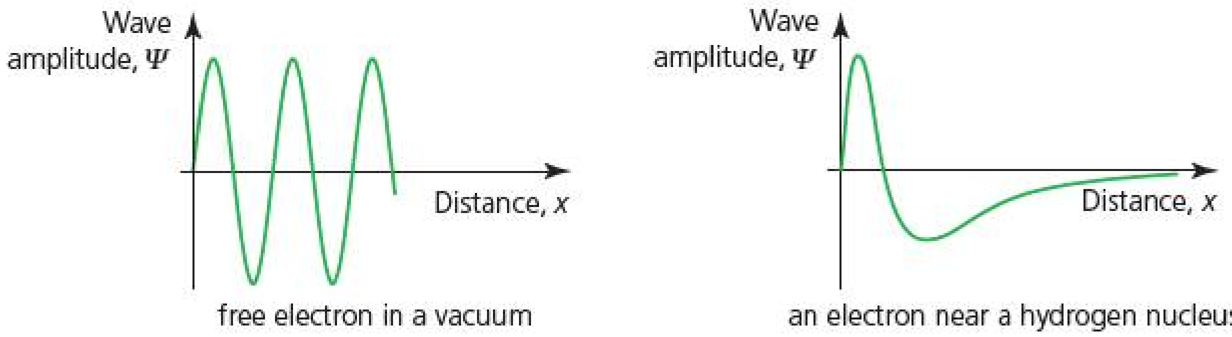


Figure 13.21 The wavefunctions, Ψ, of a free electron in a vacuum and an electron near a hydrogen nucleus (a proton)

A more useful quantity is $|\Psi(x, t)|^2$, the square of the absolute value of the amplitude. This is a measure of the electron density (**probability distribution**) in a region of space, and so gives us the probability that an electron will be found near a point x at time t. The wavefunctions of electrons in atoms are known as **orbitals** and often described as 'electron clouds'.

Solving the Schrödinger wave equation for the single electron in a hydrogen atom gives results similar to the simple 'electron in a box' model described in the previous section. It predicts that the total energy, E, of the electron in the hydrogen atom (the sum of the kinetic energy and potential energy) is given by the expression:

$$E \text{ (in eV)} = -\frac{13.6}{n^2} \quad \text{where } n \text{ is the energy level}$$

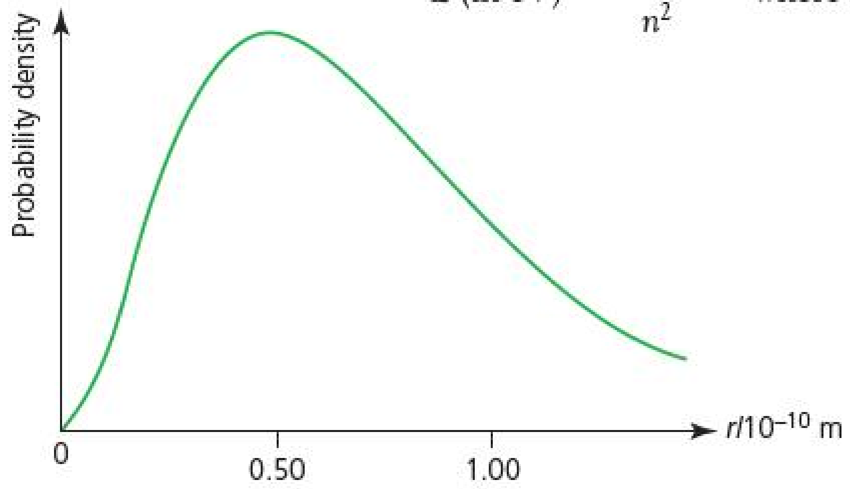


Figure 13.22 Probability distribution function for the variation with distance r from the nucleus for the energy level $n = 1$ (lowest energy, or ground state) of the hydrogen atom

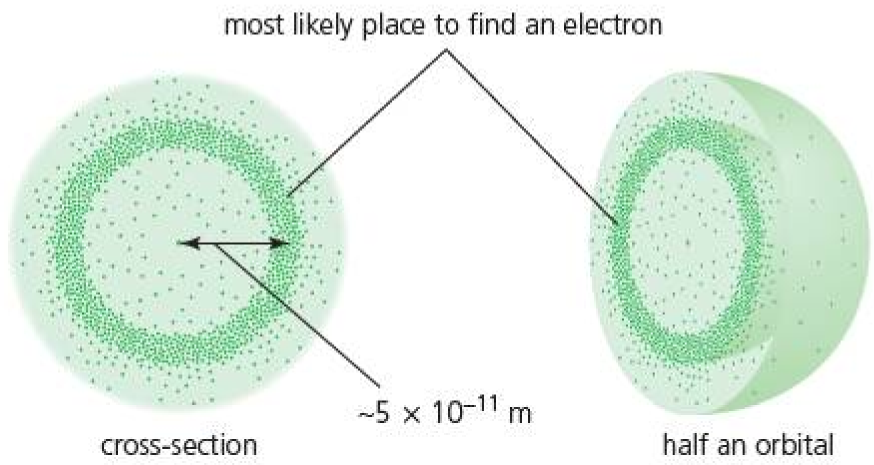


Figure 13.23 Electron cloud for the 1s orbital in the hydrogen atom

Schrödinger's theory therefore predicts that the electron in the hydrogen atom has quantized values of total energy. The electron will be found in one of the energy levels of the hydrogen atom depending on the value of the integer n. These energy levels are similar to those predicted by Bohr theory and electron transitions between energy levels are also predicted. The Schrödinger wave equation can also be applied to atoms other than hydrogen and solved approximately.

The graph in Figure 13.22 shows the *calculated* variation in *electron density* (proportional to $|\Psi|^2$) with radial distance r from the hydrogen nucleus for the lowest energy level $n = 1$ (ground state). The peak indicates the most probable distance of the electron from the nucleus, which is about 0.50×10^{-10} m. This is where the electron density is highest. The ground state orbital of a hydrogen atom is known as the 1s orbital, where 1 refers to the first energy level and s its shape.

The wavefunction exists in three dimensions and often the electron orbital is pictured as an 'electron cloud' (Figure 13.23) of varying electron density. The electron density distribution is like a photograph of the atom with a long exposure time. The electron is more likely to be in the positions where the electron density (represented by dots) is highest. For the first main energy level ($n = 1$) the orbital takes the form of a fuzzy sphere.

33 Summarize the main features of Schrödinger's quantum mechanical model of the hydrogen atom.

34 Calculate the wavelength of the photon emitted in the transition from $n = 2$ to $n = 1$.

35 Find out about the energies and shapes of p, d and f orbitals in atoms.

36 Describe how a Bohr orbit is different from a Schrödinger orbital.

The Heisenberg uncertainty principle

13.1.13 Outline the Heisenberg uncertainty principle with regard to position–momentum and time–energy.

Imagine in a 'thought experiment' recording the measurement of the speed and position of a single electron in a beam of electrons, as shown in Figure 13.24. The position of an electron in the beam can be found by aiming a photon from a light source at it. After interacting with the electron the photon may rebound into a photomultiplier tube, a device for detecting photons. However, when the photon interacts with the electron there will be a transfer of energy, making the precise measurement of the electron's speed impossible.

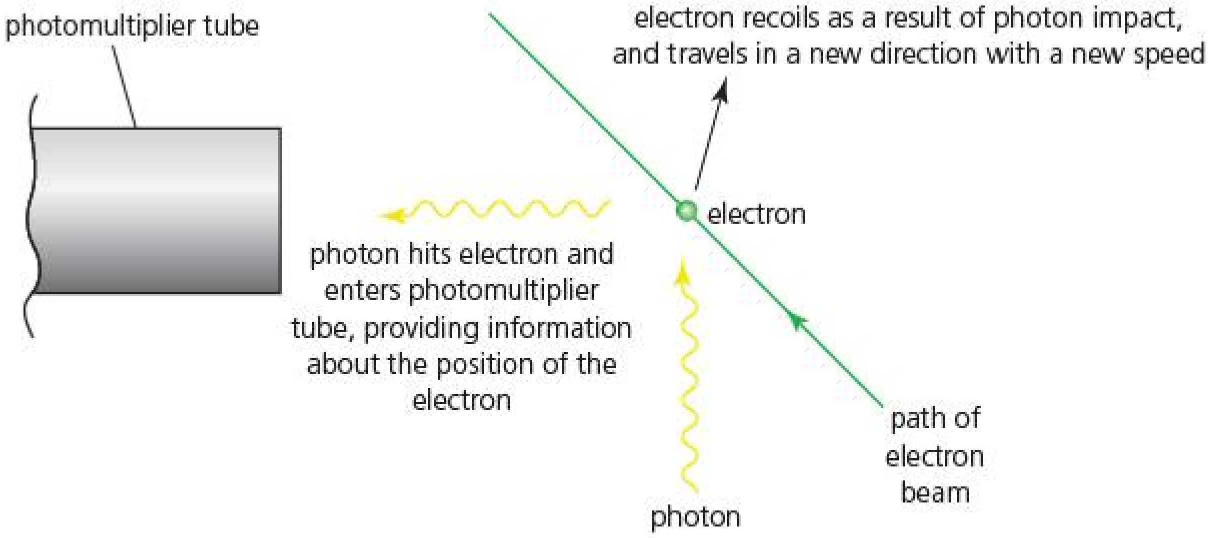


Figure 13.24 When the speed and position of an electron are measured (using photons) these quantities are altered, introducing uncertainties into the measurements

The German physicist Werner Heisenberg put these ideas into a mathematical relationship known as the Heisenberg uncertainty principle.

For the measurement of position the relationship can be written as:

$$\Delta x \Delta p \geq \frac{h}{4\pi}$$

This equation is in the IB *Physics data booklet.*

where Δx represents the uncertainty in the measurement of position, Δp represents the uncertainty in the measurement of momentum (the product of mass and velocity) and h represents Planck's constant.

The Heisenberg uncertainty principle implies that the more accurately we know the speed of the electron (the smaller Δp), the less we know about where it is (the larger Δx), and this uncertainty is significant. Because of this, momentum and position are linked variables and termed **conjugate quantities**. In particular, if one is made zero, the other has to be infinite.

This implies the following link between the de Broglie hypothesis (page 442) and the Heisenberg uncertainty principle: if a particle has a uniquely defined de Broglie wavelength, then its momentum will be known precisely, but there is no knowledge of its position. This can also be applied to the Schrödinger model of the hydrogen atom: if the wavelength of the electron's matter wave is well defined, then the position of the electron is unknown. The uncertainty principle is a general result that follows from wave–particle duality.

Measurements of time and energy are also linked variables and described by the energy–time uncertainty principle:

$$\Delta E \Delta t \geq \frac{h}{4\pi}$$

This equation is in the IB *Physics data booklet.*

where ΔE represents the uncertainty in the measurement of energy and Δt represents the uncertainty in the measurement of time.

Worked examples

14 An electron moves in a straight line with a constant speed ($1.30 \times 10^6\,\text{m s}^{-1}$). The speed can be measured to a precision of 0.10%. What is the maximum possible precision (minimum uncertainty) in its position if it is measured simultaneously?

$p = mv$

$p = (9.110 \times 10^{-31}) \times (1.30 \times 10^6) = 1.18 \times 10^{-24}\,\text{kg m s}^{-1}$

$\Delta p = 0.0010p = 1.18 \times 10^{-27}\,\text{kg m s}^{-1}$

$$\Delta x \geq \frac{h}{4\pi \times \Delta p} = \frac{6.63 \times 10^{-34}}{4 \times \pi \times 1.18 \times 10^{-27}}$$

$\Delta x_{min} = 4.45 \times 10^{-8}\,\text{m}$

Maximum possible precision of position (minimum uncertainty), $\Delta x = 44.5\,\text{nm}$.

15 The position of an electron is measured to the nearest $0.50 \times 10^{-10}\,\text{m}$. Calculate its minimum uncertainty in momentum.

$$\Delta p \geq \frac{h}{4\pi \times \Delta x} = \frac{6.63 \times 10^{-34}}{4 \times \pi \times 0.50 \times 10^{-10}}$$

$\Delta p_{min} = 1.1 \times 10^{-24}\,\text{kg m s}^{-1}$

16 Calculate the uncertainty in velocity for an electron in a $1.0 \times 10^{-10}\,\text{m}$ radius orbit in which the positional uncertainty is 1% of the radius.

$$\Delta p \geq \frac{h}{4\pi \times \Delta x} = \frac{6.63 \times 10^{-34}}{4 \times \pi \times 1.0 \times 10^{-12}} = 5.27 \times 10^{-23}\,\text{kg m s}^{-1}$$

$$\Delta v = \frac{\Delta p}{m} = \frac{5.27 \times 10^{-23}}{9.11 \times 10^{-31}} = 5.8 \times 10^7\,\text{m s}^{-1}$$

37 Find about Schrödinger's hypothetical 'cat in the box' experiment and explain its connection with quantum theory.

38 An electron is located within an atom to a distance of $0.1 \times 10^{-10}\,\text{m}$. What is the uncertainty involved in the measurement of its velocity?

39 A mass of 40.00 g is moving at a velocity of $45.00\,\text{m s}^{-1}$. If the velocity can be calculated with an accuracy of 2%, calculate the uncertainty in the position.

40 The lifetime of a free neutron is 15 minutes. How uncertain is its energy (in eV)?

41 Find out about quantum tunnelling.

TOK Link: Uncertainty principle

The idea of the detached observer of classical or Newtonian mechanics is false, since a completely isolated universe cannot be observed. An observer must always form part of an experiment – otherwise there is no experiment. The uncertainty principle also places a definite limit to the precision of measurements and hence human knowledge. The uncertainty principle is not due to any limitation of the measuring device (Chapter 1) but is a direct consequence of the dual nature of matter (wave–particle duality).

Question

1 Why does the philosophical concept of free will depend on the uncertainty principle?

13.2 Nuclear physics

Investigating the nucleus

This section looks at ways of estimating the size and mass of the nucleus, and evidence for nuclear energy levels.

Rutherford scattering

13.2.1 Explain how the radii of nuclei may be estimated from charged particle scattering experiments.

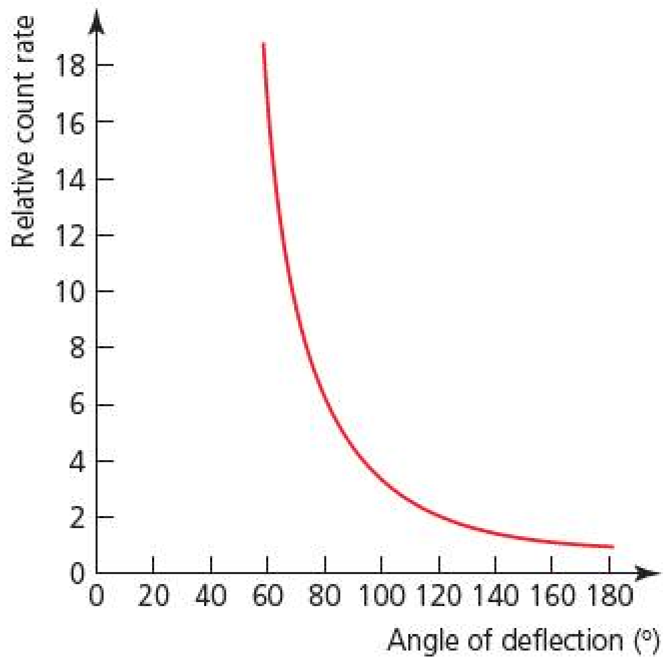


Figure 13.25 The rate of scattering of alpha particles at different angles

In Chapter 7 we looked at the experiment conducted by Geiger and Marsden, under the supervision of Ernest Rutherford, in which alpha particles were scattered by gold foil. The graph of Figure 13.25 shows the typical results obtained from this experiment. The most unexpected observation was that some alpha particles were scattered through large angles. In fact, about 1 in every 10 000 alpha particles incident on the gold foil was deflected by an angle greater than 90°.

In order for an alpha particle to be scattered straight back (with very little loss of kinetic energy), it must have 'collided' with a much larger mass (see elastic collisions in Chapter 2). Because most of the alpha particles are *not* deflected, the mass of each atom must be concentrated in a very small centre (the nucleus), such that most particles do not collide with it. Rutherford further realized that the *pattern* into which the alpha particles were scattered could only be explained by the action of large repulsions between the positive alpha particles and tiny positively charged nuclei. He proposed that each nucleus has a charge much larger than that of an alpha particle and that this is effectively concentrated at a *point* at the centre of the atom. The size of the forces between the charges is described by Coulomb's law ($F = \frac{kq_1q_2}{r^2}$, see Chapter 6), and another name for this effect is 'Coulomb scattering'.

Rutherford was able to estimate the radius of a gold nucleus by calculating how close an alpha particle got to the nucleus during its interaction. At its nearest distance the alpha particle moving directly towards the nucleus stops moving, so all its original kinetic energy is now stored as electrical potential energy. It is as if an 'invisible spring' is being squeezed between the alpha particle and gold nucleus as they come closer and closer. When the alpha particle stops moving all its initial kinetic energy (E_K) is stored in the 'spring' (Figure 13.26).

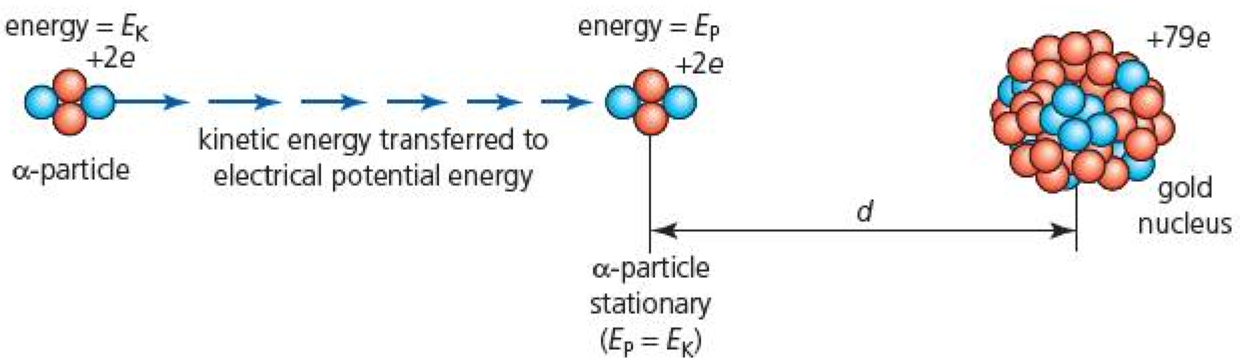


Figure 13.26 At closest approach in a head-on collision, the electrical potential energy stored in the electric field is equal to the initial kinetic energy of the alpha particle, $E_P = E_K$

The equation for electrical potential energy (E_P) depends on the separation r of the two charges:

$$E_P = \frac{q_1q_2}{4\pi\varepsilon_0 r}$$

This expression is one form of Coulomb's law (Chapter 6).

At the distance of closest approach the kinetic energy is transferred to electrical potential energy, so:

$$E_P = \frac{q_1 q_2}{4\pi\varepsilon_0 r} = E_K$$

If the kinetic energy of the alpha particle and the two charges (of the gold nucleus and the alpha particle) are known, then their separation, r, can be calculated:

$$r = \frac{q_1 q_2}{4\pi\varepsilon_0 E_K} = \frac{2e \times Ze}{4\pi\varepsilon_0 E_K}$$

where Z represents the atomic number of the gold nucleus.

If the target nucleus is gold ($Z = 79$) and the incident alpha particles have a kinetic energy of about 4 MeV (which is typical for naturally produced alpha particles), the distance of closest approach is:

$$r = \frac{2 \times 79 \times (1.6 \times 10^{-19})^2}{4\pi \times 8.9 \times 10^{-12} \times 4.0 \times 10^6 \times 1.6 \times 10^{-19}} \approx 10^{-14}\,\text{m}$$

This is the separation of the alpha particle and gold nuclei at closest approach and gives an *upper limit* for the sum of their radii. From this calculation it is assumed that the radius of a gold nucleus is of the order of 10^{-14} m. We cannot be sure that the alpha particle 'touches' the nucleus; a more energetic alpha particle might get closer still. More accurate measurements from the scattering of high energy electrons confirm that the gold nucleus has a radius of 6.9×10^{-15} m.

The behaviour of the alpha particle when it collides with the **potential hill** around the nucleus can be modelled with the apparatus shown in Figure 13.27. It is a gravitational analogue, with a ball bearing representing an alpha particle. The curved surface is constructed so that its height above the bench is proportional to $1/r$, where r represents the distance from its centre. The ball bearings are rolled down a ramp to give them a fixed amount of kinetic energy. As they move over the curved surface which represents the potential hill around the nucleus, they slow down and then accelerate again, coming out in a different direction.

Figure 13.27 A model of the potential hill around a nucleus

Because there is a uniform gravitational field in the laboratory, the ball bearings gain gravitational potential energy in proportion to $1/r$. This means that the force on them varies as $1/r^2$. The Coulomb repulsion between a nucleus and an alpha particle varies in a similar way. The electrical potential energy is proportional to $1/r$ and the electrostatic repulsive force is proportional to $1/r^2$.

42 Outline how the radii of metal nuclei may be estimated from Coulomb scattering experiments.

43 Calculate the velocity at which an alpha particle (mass of 6.64×10^{-27} kg) should travel towards the nucleus of a gold atom (charge $+79e$) so it gets within 2.7×10^{-14} m of it. Assume the gold nucleus remains stationary.

The mass spectrometer

13.2.2 Describe how the masses of nuclei may be determined using a Bainbridge mass spectrometer.

The first direct investigation of the mass of atoms (and molecules) was made possible by the development of the **mass spectrometer**. The first mass spectrometer was built in 1918 by William Aston, a student of J.J. Thomson. It provided direct evidence for the existence of isotopes (Chapter 7). A more accurate mass spectrometer was developed by William Bainbridge in 1932.

This device uses the interaction of charged ions with electric and magnetic fields to measure the relative masses of atoms and to find their relative abundances. Figure 13.28 shows how a simple Bainbridge mass spectrometer works. There is a vacuum inside the machine; no air is allowed to enter. This is to avoid collisions with particles in the air that would disrupt the flight of the ions produced inside the instrument.

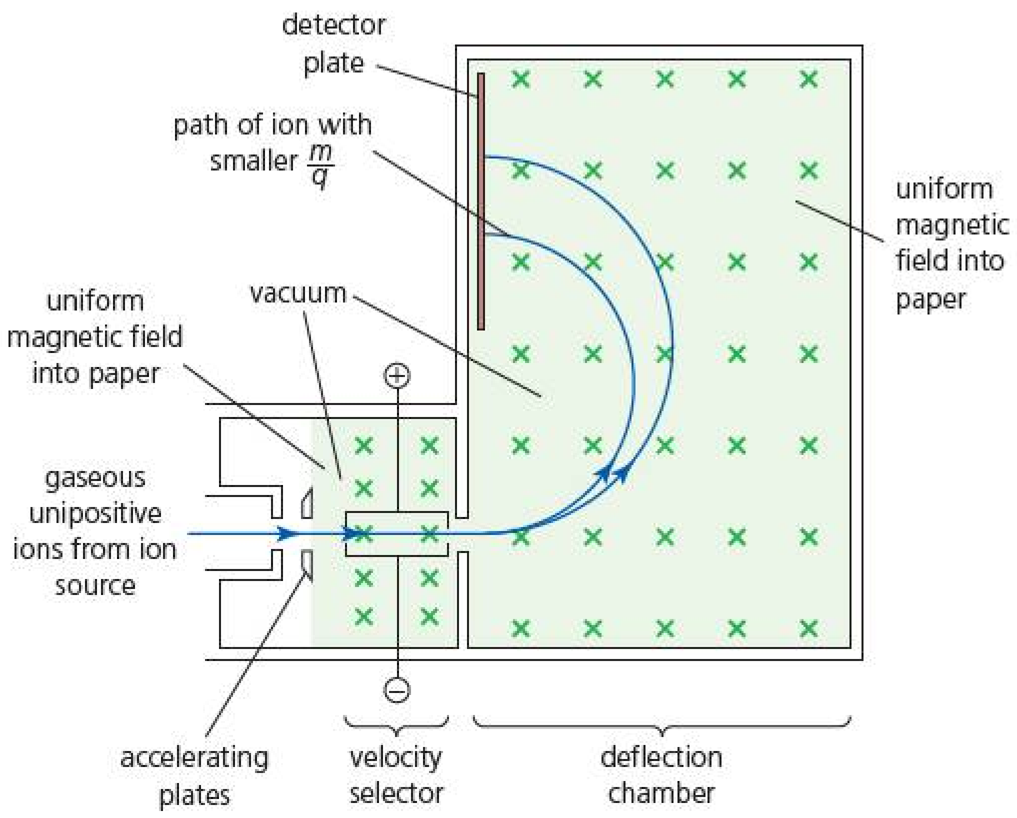


Figure 13.28 Essential features of a Bainbridge mass spectrometer

Positive ions are produced in the mass spectrometer by bombarding gaseous atoms or molecules (M in the following equation) with a stream of high speed electrons:

$$M(g) + e^- \rightarrow M^+(g) + e^- + e^-$$

The ions are accelerated (by a pair of positive plates) and then pass through a **velocity selector**. Here electric and magnetic fields are applied to the ions so that only a narrow beam of ions travelling at the same velocity continue in a straight line and can enter the next chamber. These ions then travel through a uniform magnetic field. This causes the ions to move in a circular path with radius, r, which depends on the mass to charge ratio. For ions with mass m and charge, q, travelling with velocity, v, through the magnetic field, B:

centripetal force = force due to magnetic field

$$\frac{mv^2}{r} = Bqv$$

The radius of the circle for the ion is therefore:

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

If the ions have the same charge, q (usually unipositive, i.e. with a single positive charge), and they are all selected to be travelling at the same velocity, v, then the radius of the circle of each ion's path will only depend on the mass of the ion. An ion with a larger relative mass will travel in a larger circle (for the same magnetic field strength).

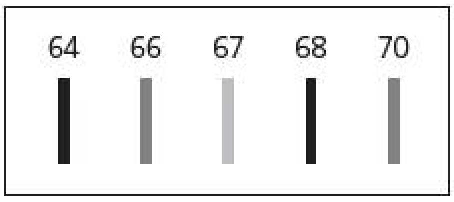


Figure 13.29 A mass spectrum obtained by a mass spectrograph

A number of vertical lines will be obtained on the detector plate, each line corresponding to a different isotope of the same element. The position of a line on the plate will allow the radius, r, to be determined. As the magnetic field strength, B, the charge on the ion, e, and velocity of the ion, v, are all known, the mass of the ion, m, can be easily determined. The older type of Bainbridge mass spectrometer shown is actually a mass spectrograph, since the beam of ions is directed onto a photographic plate (Figure 13.29). The relative intensities of the lines allowed an estimate to be made of the relative amounts of isotopes.

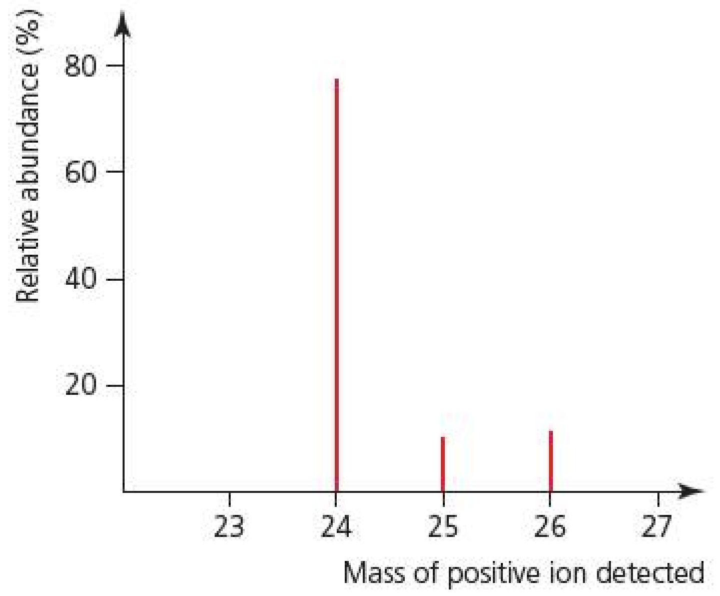


Figure 13.30 The mass spectrum of naturally occurring magnesium atoms showing the percentage abundances of three isotopes with their masses

Modern mass spectrometers neutralize the positive ions with electrons and count the number of ions directly, before amplifying the signal. The results are displayed on a computer screen in the form of a 'bar chart' (Figure 13.30).

Mass spectrometers are sensitive detectors of isotopes based on their masses. A number of satellites and spacecraft are equipped with mass spectrometers to allow them to identify the small numbers of particles intercepted in space. For example, the SOHO satellite uses a mass spectrometer to analyse the solar wind. They are also used in carbon dating (Chapter 7).

44 Describe the structure and operation of a Bainbridge mass spectrometer.

45 Find out how mass spectrometers on missions to Mars have been used to study the planet. What information have they provided to scientists?

46 Unipositive neon-21 ions, $^{21}Ne^{+}$, travelling with a velocity of $2.50 \times 10^{5}\,m\,s^{-1}$, enter a magnetic field of value 0.80T, which deflects them into a circular path. Calculate the radius of the circular path.

Additional Perspectives

The velocity selector

The **velocity selector** uses an electric field, E, and a magnetic field, B, which are at right angles to each other, and also at right angles to the initial direction of motion of the positive ions passing through them. These fields both exert a force on the stream of positive ions. The two forces are arranged to act in opposite directions, as shown in Figure 13.31. The force on a particle of charge q due to the electric field, F_E, remains constant, but the force due to the magnetic field, F_B, varies according to the velocity, v, of the particle.

$$F_E = qE \qquad F_B = Bqv$$

If particles are to pass through the selector undeviated, then:

$$F_E = F_B$$

Therefore $qE = Bqv$, so:

$$v = \frac{E}{B}$$

The velocity that is selected can be varied by changing the ratio of the strength of the magnetic and electric fields.

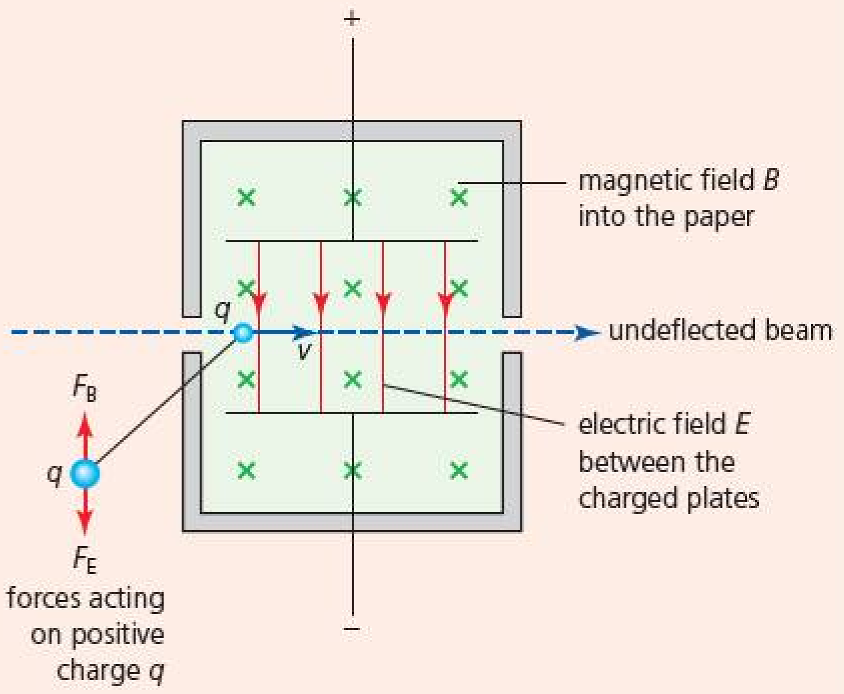


Figure 13.31 Velocity selector (of a mass spectrometer)

Question

1 It is required to select charged ions which have a speed of $3.5 \times 10^{6}\,m\,s^{-1}$. The electric field strength in the velocity selector is $4.5 \times 10^{4}\,V\,m^{-1}$. Calculate the magnetic field strength required in teslas.

Nuclear energy levels

13.2.3 Describe one piece of evidence for the existence of nuclear energy levels.

The nucleus is a quantum system and has discrete energy levels. The observation that the energies of gamma rays are discrete provides strong experimental evidence for the existence of nuclear energy levels. (This is in contrast to beta decays, in which the electron or positron has a continuous range of energies – see the next section on radioactive decay.)

Alpha decay

During alpha decay an alpha particle is ejected from the nucleus. The smaller alpha particle has the higher speed and the parent nucleus moves backwards (recoils) at a much lower speed. This is an example of Newton's third law of motion (Chapter 2). Energy is released during alpha decay. In a simple alpha decay, this energy is carried away as the kinetic energy of the alpha particle and the recoiling nucleus.

For a particular radioactive isotope decaying by simple alpha emission, the products of the decay are always the same for each nucleus – the daughter nuclide and an alpha particle. Plotting the proportion of alpha particles with a particular energy against alpha particle energy produces an energy spectrum. For the simple two-particle decay of radon-220, the graph shows that the energy of the alpha particles is discrete (Figure 13.32). Since these particles have the same mass in each decay, they always share the energy in the same way. Thus the energy involved in this decay process is also discrete.

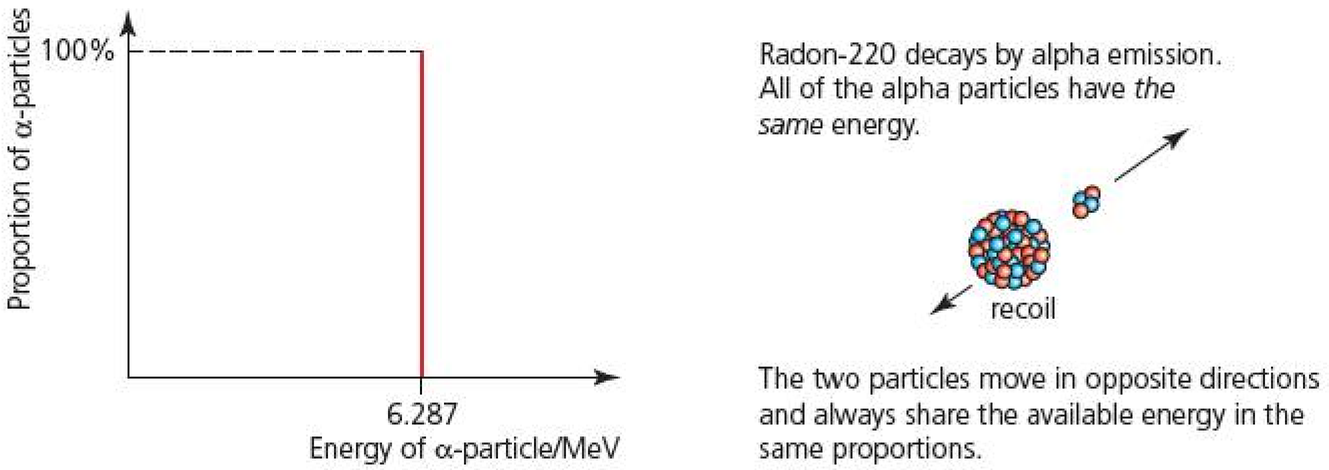


Figure 13.32 Energy spectrum for the alpha decay of radon-220 to plutonium-216

Gamma rays

Evidence for the existence of nuclear energy levels comes from studying gamma ray emissions from radioactive nuclei. These emissions do not change the numbers of protons and neutrons in the nucleus, but remove energy from the nucleus in the form of gamma rays (photons).

The gamma rays have energies that are discrete and distinctive for the nucleus. This suggests that the gamma rays are emitted as a result of **nuclear transitions** from an excited state with a higher energy level to a lower energy level, similar to electronic transitions in the atom (Chapter 7). For example, as shown in Figure 13.33, when carbon-15 decays by beta emission, some of the nitrogen-15 daughter nuclei are left in an excited state. These excited nuclei release gamma rays of a particular frequency, and hence energy, to reach the lowest energy level (the ground state).

The photons emitted by radioactive nuclei carry more energy, and therefore have higher frequencies, than photons emitted as a result of transitions involving electrons. This means that the energy changes involved in nuclear processes are much larger than those involved in transitions of electrons.

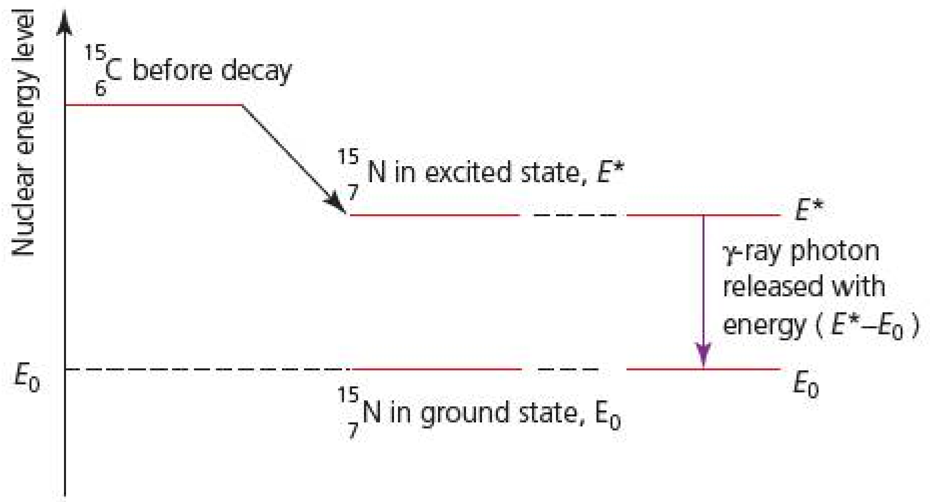


Figure 13.33 The gamma rays released when nitrogen-15 nuclei fall from an excited state to the ground state have a characteristic frequency

47 Figure 13.34 shows the lowest two energy levels for the nucleus of a heavy atom.

a How do nuclear energies compare with electron energies?

b Calculate the wavelength of the electromagnetic radiation emitted in a nuclear change involving a transition from the first excited state to the ground state.

c To what part of the electromagnetic spectrum does this photon belong?

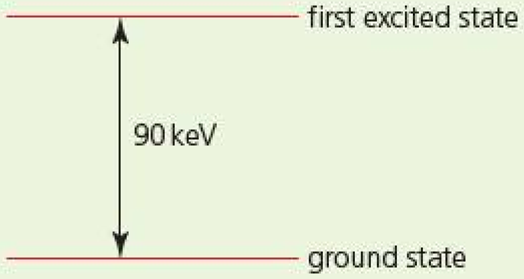


Figure 13.34

Radioactive decay

Beta energy spectra

13.2.4 Describe β^+ decay, including the existence of the neutrino.

Calculations using the masses of nuclei suggest that beta particles emitted from a radioactive source should all have the same energy, in a similar way to alpha particles. However, measurements show that although the *maximum* kinetic energy of the beta particles is characteristic of the beta source, the particles are emitted with kinetic energies that vary continuously, from zero to the maximum (Figure 13.35).

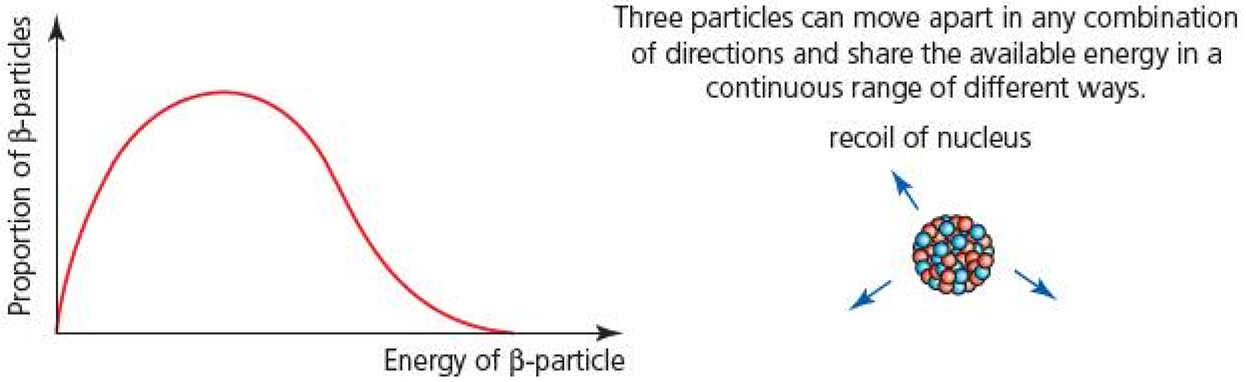


Figure 13.35 Typical energy spectrum for beta decay

This unexpected observation was first recorded in 1928 and suggested that energy and momentum might not be conserved. However, in 1933 the Austrian physicist Wolfgang Pauli suggested that another undetected particle was also involved in beta decay. This particle was hypothesized to be neutral (otherwise the electrical charge would not conserved), have little or no mass (otherwise the energy curve of the beta particles would be of a different shape), and be able to carry the kinetic energy not carried away by the nucleus or beta particle. This new particle was named the **neutrino** (and its antiparticle, the antineutrino). Because the neutrino interacts weakly with matter, it was not until 1955 that evidence for its existence was obtained.

There are two kinds of beta decay; (i) **beta-negative decay** (see Chapter 7) in which a neutron changes into a proton with the emission of an electron and an antineutrino; (ii) **beta-positive decay**, in which a proton changes into a neutron with the emission of a positron (positive electron) and a neutrino.

The full beta decay equations then become:

Beta-negative decay	*Beta-positive decay*
$n \rightarrow p + e^- + \overline{\nu}_e$	$p \rightarrow n + e^+ + \nu_e$

The neutrinos carry away the energy not carried away by the beta particle (positron or electron). It can be shared in any ratio, explaining the continuous spectrum of beta particle energies. The three emissions can vary in relative directions.

48 Describe beta decay (negative and positive) and beta energy spectra and explain why the neutrino was postulated to account for these spectra.

Additional Perspectives

Quarks

Protons and neutrons are not fundamental particles. They are composed of smaller particles called quarks. The six different types of quarks are known as *flavours*. The two most stable quarks are the *up* and *down* quarks. A proton is composed of two up quarks and one down quark; a neutron is composed of two down quarks and an up quark. Figure 13.36 summarizes what is happening inside protons and neutrons when they undergo beta decay (the production of neutrinos is ignored).

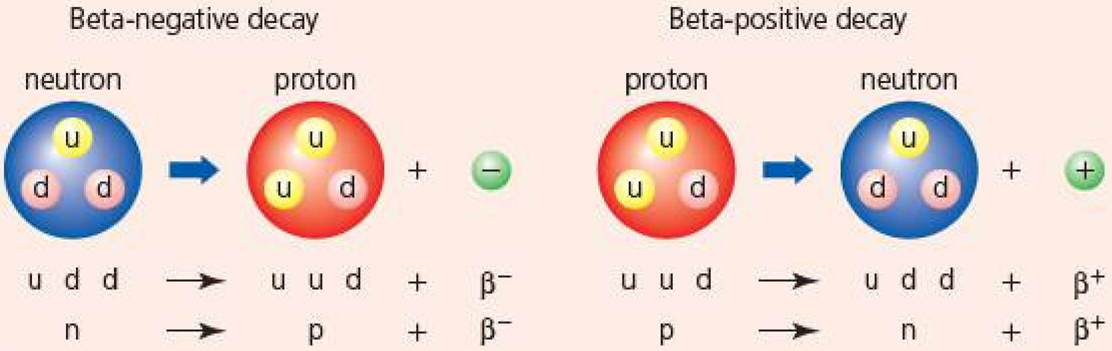


Figure 13.36 Beta decay can be explained by quarks changing their flavours. In beta-negative decay, a down quark in a neutron decays to form an up quark, and an electron is emitted. In beta-positive decay, an up quark in a proton decays to form a down quark, and a positron is emitted

Question

1 Find out why free quarks will never be observed.

Activity and decay constant

13.2.5 State the radioactive decay law as an exponential function and **define** the *decay constant*.

In the dice experiment in Chapter 7, increasing the number of dice thrown increases the number of sixes that appear (this experiment provides a simple analogy for radioactive decay). Similarly, if the decay of a radioactive material is investigated, for example, by using a Geiger–Müller tube, then it is found that the greater the number of radioactive nuclides in the sample, the greater the rate of decay.

This can be described mathematically by the following expression:

$$-\frac{\Delta N}{\Delta t} \propto N$$

where delta, Δ, represents 'change in' and N represents the number of undecayed nuclides in the sample.

Therefore $\frac{\Delta N}{\Delta t}$ represents the rate at which the number of nuclides in the sample is changing, and hence $-\frac{\Delta N}{\Delta t}$ represents the rate of decay.

Introducing a constant of proportionality, λ, we get:

$$-\frac{\Delta N}{\Delta t} = \lambda N$$

The constant λ (lambda) is known as the **decay constant**. It has units of reciprocal of time (s^{-1}).

The decay constant is defined as the probability per unit time that any particular nucleus will undergo decay.

$$\lambda = \frac{-\Delta N}{N \Delta t}$$

In Chapter 7 we introduced the **activity**, A, of a radioactive source as the number of nuclei decaying per second. Activity is the same as the rate of decay, therefore:

$$-\frac{\Delta N}{\Delta t} = A$$

and so

$$A = \lambda N$$

Both of these equations are given in the IB *Physics data booklet.*

Activity is measured in **becquerels** (Bq), where 1 becquerel is 1 decay per second. The becquerel is named after Henri Becquerel, who shared a Nobel Prize for Physics with Pierre and Marie Curie for their work in discovering radioactivity.

Worked examples

17 The activity of a radioactive sample is 2.5×10^5 Bq. The sample has a decay constant of $1.8 \times 10^{-16}\,s^{-1}$. Determine the number of undecayed nuclei remaining in the sample at that time.

$A = \lambda N$

$2.5 \times 10^{-5} = (1.8 \times 10^{-16}) \times N$

$N = \frac{2.5 \times 10^5}{1.8 \times 10^{-16}} = 1.4 \times 10^{21}$

18 A radioactive sample emits alpha particles at the rate of 2.1×10^{12} per second at the time when there are 5.0×10^{20} undecayed nuclei left in the sample. Determine the decay constant of the radioactive sample.

$\lambda = \frac{A}{N} = \frac{2.1 \times 10^{12}}{5.0 \times 10^{20}} = 4.2 \times 10^{-9}\,s^{-1}$

19 A sample of a radioactive nuclide initially contains 2.0×10^5 nuclei. Its decay constant is $0.40\,s^{-1}$. What is the initial activity?

$A = \lambda N$

$A = 0.40 \times 2.0 \times 10^5 = 8.0 \times 10^4\,s^{-1}$

The equation $-\frac{\Delta N}{\Delta t} = \lambda N$ is important because it relates a quantity that can be measured $\left(-\frac{\Delta N}{\Delta t}\text{, the rate of decay or the activity}\right)$ to a quantity which cannot in practice be determined (N, the number of undecayed nuclei).

The equation $\frac{\Delta N}{\Delta t} = -\lambda N$ has the following solution:

$$N = N_0 e^{-\lambda t}$$

This equation is in the IB *Physics data booklet*.

In this equation N_0 represents the initial number of undecayed nuclei in the sample, and N represents the number of undecayed nuclei after time t.

Since A is proportional to N, the equation can be expresed as

$$A = \lambda N = \lambda N_0 e^{-\lambda t}$$

This equation is in the IB *Physics data booklet*.

Worked example

20 A radioactive element decays $\frac{7}{8}$ of its original mass in 12 days. Calculate the fraction of radioactive mass that is left after 24 days.

After 12 days $\frac{N}{N_0} = \frac{1}{8}$

therefore $\frac{N}{N_0} = \frac{1}{8} = e^{-\lambda(12)}$

$8 = e^{12\lambda}$

$\lambda = \frac{\ln 8}{12} = 0.173$

After 24 days $\frac{N}{N_0} = e^{-0.173(24)} = 0.0157 = \frac{1}{64}$

49 State the radioactive decay law as an exponential function and define the decay constant.

50 A sample of radium contains 6.64×10^{23} atoms. It emits alpha particles and has a decay constant $1.36 \times 10^{-11}\,s^{-1}$. How many atoms are left after 100 years?

51 A radioactive nuclide has a decay constant of $0.0126\,s^{-1}$. Initially a sample of the nuclide contains 10 000 nuclei.

a What is the initial activity of the sample?
b How many nuclei remain undecayed after 200 s?

Additional Perspectives

Exponential graphs

The graph of N against t for the relationship $N = N_0 e^{-\lambda t}$ has the shape known as an exponential decay curve. Any relationship where the rate of change of a quantity is proportional to the quantity will give a graph with the same shape. As well as radioactive decay, many other physical effects, for example the discharge of a capacitor (Chapter 14), can be modelled to a good approximation by a negative exponential function. This is why e and logarithms to the base e (ln) are useful in physics.

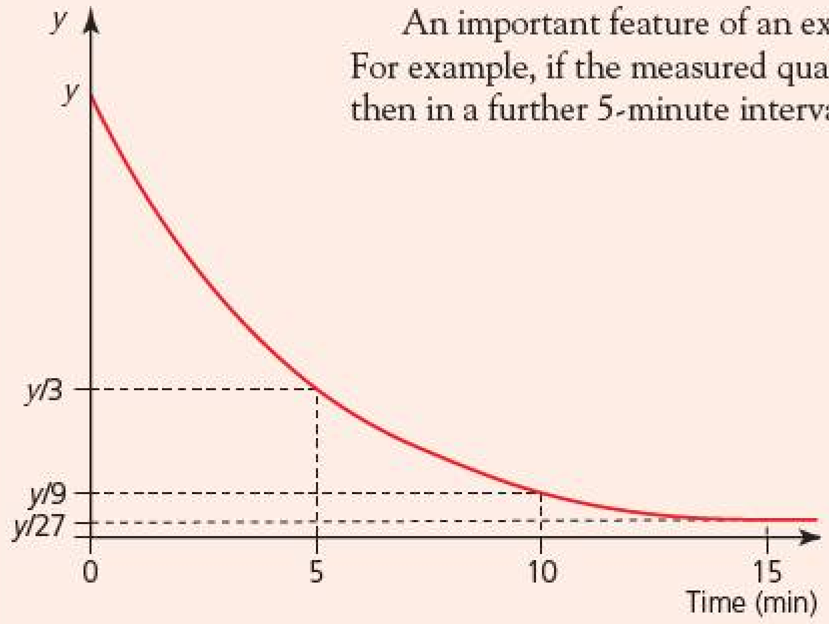


Figure 13.37 An exponential decay curve

An important feature of an exponential relationship is that the fractional change is constant. For example, if the measured quantity reduces to 1/3 of its original value in a 5-minute interval, then in a further 5-minute interval it will reduce to 1/3 of the value at the start of that interval, to 1/9 of the original value, and so on (Figure 13.37). This is true for any chosen fraction. Half-life is the time interval for which the constant fractional change is 1/2.

If the relationship between a quantity y and time is exponential, a plot of $\ln y$ (or $\log y$) against time will always produce a straight line.

Question

1 Construct a spreadsheet that plots an exponential decay curve where the user can adjust the half-life and decay constant.

Decay constant and half-life

13.2.6 Derive the relationship between decay constant and half-life.

Using the equation $N = N_0 e^{-\lambda t}$, we can derive an equation which relates the half-life, $T_{1/2}$, to the decay constant, λ.

For any radioactive nuclide, the number of undecayed nuclei after one half-life is, by the definition of half-life, equal to $N_0/2$, where N_0 represents the original number of undecayed nuclei. Substituting this value for N in the radioactive decay equation at time $t = T_{1/2}$ we have:

$$\frac{N_0}{2} = N_0 e^{-\lambda T_{1/2}}$$

Dividing each side of the equation by N_0:

$$\frac{1}{2} = e^{-\lambda T_{1/2}} \quad \text{or} \quad 2 = e^{\lambda T_{1/2}}$$

Taking natural logarithms (to the base e):

$$\ln 2 = \lambda T_{1/2}$$

So that:

$$T_{1/2} = \frac{\ln 2}{\lambda}$$

This equation is in the IB *Physics data booklet*.

or

$$T_{1/2} = \frac{0.693}{\lambda}$$

Worked examples

21 A radioactive sample gives a count rate of $100\,s^{-1}$ at a certain instant of time. After 100 s the count rate drops to $20\,s^{-1}$. The background count rate is measured to be $10\,s^{-1}$. Calculate the half-life of the sample. Assume that the count rate is a measure of the activity.

Initial count rate due to sample $= 100\,s^{-1} - 10\,s^{-1} = 90\,s^{-1}$
Count rate due to sample after 100 s $= 20\,s^{-1} - 10\,s^{-1} = 10\,s^{-1}$

$$A = A_0 e^{-\lambda t}$$

$$10 = 90 e^{-100\lambda}$$

$$\lambda = \frac{-\ln \frac{10}{90}}{100}$$

$$T_{1/2} = \frac{100 \ln 2}{-\ln \frac{1}{9}} = 32\,s$$

22 The radioactive element A has 6.4×10^{11} atoms and a half-life of 2.00 hours. Radioactive element B has 8.0×10^{10} atoms and a half-life of 3.00 hours. Calculate how much time will pass before the two elements have the same number of radioactive atoms.

For A, $\lambda_A = \frac{0.693}{2.00} = 0.347\,h^{-1}$ for B, $\lambda_B = \frac{0.693}{3.00} = 0.231\,h^{-1}$

At time t: $N_A = (6.4 \times 10^{11})e^{-0.347t}$ $N_B = (8.0 \times 10^{10})e^{-0.231t}$

For $N_A = N_B$: $(6.4 \times 10^{11})e^{-0.347t} = (8.0 \times 10^{10})e^{-0.231t}$

$$8 = \frac{e^{-0.231t}}{e^{-0.347t}}$$

$8 = e^{0.116t}$

$\ln 8 = \ln(e^{0.116t})$; $2.079 = 0.116t$

$t = 18\,h$

After 18 hours the two elements will have the same number of active atoms.

52 Radioactive carbon in a leather sample decays with a half-life of 5770 years.
a What is the decay constant?
b Calculate the fraction of radioactive carbon remaining after 10 000 years.

Measurement of half-life

13.2.7 Outline methods for measuring the half-life of an isotope.

The method used to measure the half-life of a radioactive element depends on whether the half-life is relatively long or short. If the activity of the sample stays virtually constant over a few hours then it has a relatively long half-life. However, if its activity decreases during a few hours then the radioactive element has a relatively short half-life.

Isotopes with long half-lives

$$\frac{\Delta N}{\Delta t} = -\lambda N$$

If the activity ($\Delta N/\Delta t$) of a source can be determined, then the decay constant (and therefore the half-life) can be calculated if the number of undecayed atoms, N, is known. Theoretically this is straightforward, but determining practically the number of atoms of a particular isotope in a sample containing a mixture of isotopes is not easy and requires sophisticated equipment like a mass spectrometer.

However, for a *pure* sample of mass m, the number of atoms of the isotope can be determined from the relative atomic mass (A_r) and Avogadro's contant, N_A as follows:

$$N = \frac{mN_A}{A_r}$$

Therefore if the activity ($\Delta N/\Delta t$) is measured we can calculate the half-life, $T_{1/2}$, from the equation:

$$\frac{\Delta N}{\Delta t} = \frac{-\lambda m N_A}{A_r} = \frac{-0.693 m N_A}{T_{1/2} A_r}$$

53 The experimentally determined activity from 1.0 gram of radium-226 is 1.14×10^{18} alpha particles per year. Calculate its half-life.

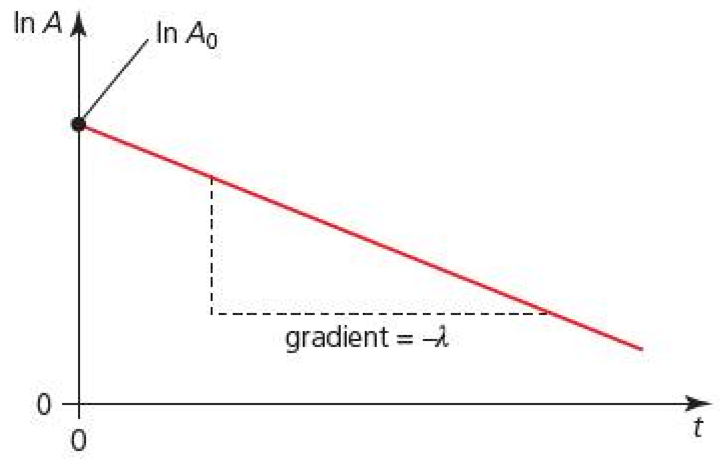


Figure 13.38 A logarithmic–linear graph to show exponential decay of a radioactive nuclide

Isotopes with short half-lives

If an isotope's half-life is less than several days then it can be measured directly. This is done by measuring the count rate over a short period of time at regular intervals. The measured count rate is assumed to be proportional to activity. A graph of activity, A (or count rate) against time, t, is plotted and the half-life obtained directly from the graph (Chapter 7). (In practice the graphs will be count rates, not activities.)

Alternatively, a graph can be plotted of natural logarithm of the activity ($\ln A$) against time, t. This will give a straight line with a gradient of $-\lambda$ (Figure 13.38). Since $T_{1/2} = \frac{0.693}{\lambda}$, the half-life can be calculated. This is a better method because the reliability of the data can be more readily assessed by seeing how close to the straight line of best fit the data points lie.

This approach relies on transforming the equation $A = A_0 e^{-\lambda t}$ which describes the activity in radioactive decay.

$$A = A_0 e^{-\lambda t}$$

Taking natural logarithms:

$$\ln A = \ln A_0 - \lambda t$$

The equation can be compared to the equation for a straight line ($y = mx + c$):

$$\underbrace{\ln A}_{y} = \underbrace{\ln A_0}_{c} - \underbrace{\lambda}_{m}\underbrace{t}_{x}$$

Thus the gradient is equal to $-\lambda$.

In accurate experiments where measurements of count rate are taken, an allowance should be made for the effect of background radiation. The two graphs in Figure 13.39 show the effect of allowing for the background count. The corrected count rate is assumed to be proportional to the activity.

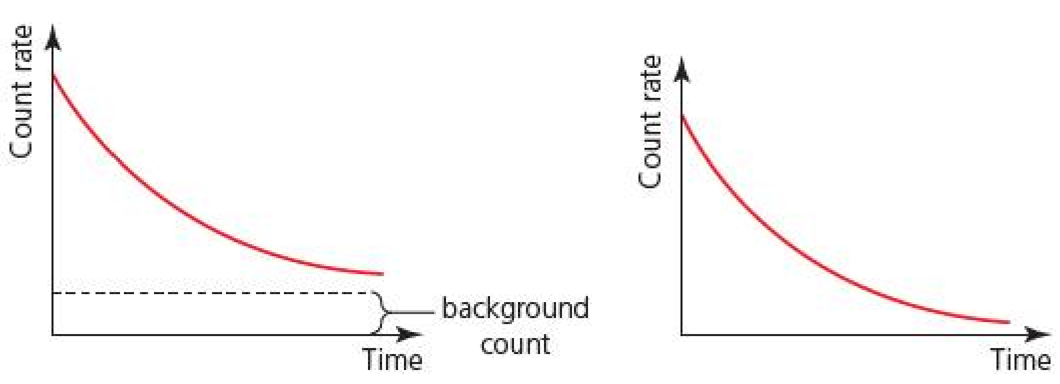


Figure 13.39 Uncorrected and corrected count rates for a radioactive isotope

When the half-life of the isotope is very small, less than a second, then both of these methods are unsuitable. Such half-lives may be found from tracks in a cloud or bubble chamber.

13.2.8 Solve problems involving radioactive half-life.

54 Radium-226 has a long half-life (>1000 years); radium-227 has a half-life of 42 minutes. Outline how the two half-lives of these radium isotopes can be determined experimentally.

55 The following data were obtained from the decay of caesium-130:

Time/s	Activity of ^{130}Cs/disintegrations s^{-1}
0	200
500	165
1500	113
2500	79
3500	54
4500	38
5500	26

Use a spreadsheet to plot a graph of ln *A* against *t* to determine the decay constant and hence the half-life (to the nearest integer).

56 A sample contains the atoms of radioactive element A and another sample contains the atoms of a radioactive element B. After a fixed length of time, it is found that $\frac{7}{8}$ of atoms A and $\frac{3}{4}$ of atoms B have decayed.

Calculate the value of the ratio $\frac{\text{half-life of element A}}{\text{half-life of element B}}$.

SUMMARY OF KNOWLEDGE

13.1 Quantum physics

- Electrons may be emitted from cleaned metal surfaces if the metal is illuminated by electromagnetic radiation, usually ultraviolet radiation. This phenomenon is called photoelectric emission.
- Electron energies are often measured in electronvolts (eV), where $1\,\text{eV} = 1.6 \times 10^{-19}\,\text{J}$.
- For photoelectric emission to occur, there is a threshold frequency (energy) below which no electrons are released. Above the threshold frequency photoelectrons are released at a rate proportional to the intensity of the light. The energy of the emitted electrons is independent of the intensity of the incident radiation.
- Photoelectric emission cannot be explained by the wave model of light. It is necessary to apply quantum theory, in which electromagnetic radiation is described as consisting of packets of energy called photons.
- The energy of a photon, E, is given by the Planck relationship: $E = hf$, where h represents Planck's constant and f represents the frequency of the electromagnetic radiation.
- The work function, ϕ, of a metal is the minimum energy needed to free an electron from the surface of a metal. Different metals have different work functions.
- The Einstein photoelectric equation is: $hf = \phi + \frac{1}{2}m_e v_{max}^2$ or $hf = \phi + E_{max}$
- The threshold frequency for photoelectric emission, f_0, is given by: $hf_0 = \phi$
- The maximum kinetic energy of the photoelectrons is determined by the frequency of the electromagnetic radiation and not by the intensity.
- If a reverse potential is applied the kinetic energy lost by photoelectrons = electric potential energy gained by those electrons: $\frac{1}{2}m_e v^2 = eV$. Therefore:

$$hf = hf_0 + eV \qquad \text{or} \qquad V = \left(\frac{h}{e}\right)f - \frac{hf_0}{e}$$

- All moving particles, in principle, show wave-like properties. However, only electrons and atoms, which have very small mass, will show their wave nature (diffraction and interference) in experiments.
- The de Broglie wavelength, λ, is given by $\frac{h}{p}$, where p represents the momentum of the electron and h represents Planck's constant.
- If a charged particle carrying a charge of q coulombs is accelerated by applying a potential difference of V volts, then the de Broglie wavelength of the particle is given by the following relationship:

$$\lambda = \frac{h}{\sqrt{2m_e qV}}$$

- de Broglie waves are matter waves. The amplitude of the matter wave at a point represents the probability of the particle being at that position.
- The Davisson–Germer experiment involved directing a beam of electrons at a thin metal foil. The electrons were diffracted, providing experimental evidence for de Broglie's hypothesis.
- At certain angles there were peaks in the intensity of the scattered electron beam. These peaks indicated wave behaviour for the electrons, (and could be interpreted to give values for the lattice spacing in the metal crystal.
- Light is regarded as behaving as a stream of particles (quanta) or as a wave, depending on the phenomenon being described.
- A continuous spectrum has a complete range of wavelengths.
- An emission spectrum can be produced by passing light from a sample of excited gas through a slit to generate a narrow beam which is then directed onto a prism or diffraction grating, which causes the light to undergo dispersion.
- An emission spectrum is produced when matter emits electromagnetic radiation. In the visible region it consists of a series of sharp coloured lines on a black background.
- An absorption spectrum is produced using the same apparatus (spectrometer) except that a beam of white light is passed through a sample of cool gas (maintained at low pressure).
- An absorption spectrum is produced when electromagnetic radiation travelling through matter is absorbed. In the visible region it consists of a series of sharp black lines on a coloured background.
- An atom is a quantized system and has definite electron energy levels.
- The lowest energy level is the ground state, and all the other energy levels are excited states. The highest possible energy level occurs when the atom is ionized.
- The values of the energy levels are governed by the principal quantum number, n.
- The energy level at ionization is given the value of 0 eV; all other energy levels have negative values.
- An electronic transition is the process in which an atom changes its quantum state by absorbing or emitting a certain discrete amount of energy.
- The amount of energy, ΔE, absorbed or emitted in an electronic transition is given by:

$$\Delta E = E_2 - E_1 = hf = \frac{hc}{\lambda}$$

- Spectra can be interpreted in terms of the transition of electrons in atoms between different energy levels.
- A line in an absorption spectrum is formed when an electron moves from a low energy level to a higher energy level. This is known as excitation.
- A line in an emission spectrum is formed when an electron moves from a high energy level to a lower energy level.

- For a photon to be absorbed it must have exactly the correct amount of energy to raise an electron from a lower energy level to a higher energy level.
- Absorption and emission spectra support the concept that atoms have quantized energy levels.
- The 'electron in a box' is a simple quantum mechanical model which shows how boundary conditions that a wavefunction must satisfy lead to the quantization of energy.
- The system consists of an electron with a mass m in a one-dimensional region of space of length L: the potential is zero for $0 \le x \le L$ and infinite elsewhere.
- The de Broglie waves associated with the electron will be standing waves of wavelength $\frac{2L}{n}$ where L is the length of the box and n is a positive integer.
- The kinetic energy of the electron in the box is given by: kinetic energy $(E_K) = \frac{n^2h^2}{8m_eL^2}$, where m_e is the mass of the electron and n is the energy level.
- The Schrödinger wave model builds on the concept of de Broglie's matter waves.
- The electron is described in terms of a wavefunction, Ψ, where at any instant of time, the wavefunction has different values at different points in space.
- A wavefunction is a solution of Schrödinger wave equation and is a mathematical description of an electron as a wave.
- The probability of finding the electron at any point in space is given by the square of the absolute amplitude of the wavefunction at that point.
- Orbitals are three dimensional regions in space where electrons are likely to be found. Orbitals have different energies and shapes.
- The wavefunctions for electrons in different energy levels can be determined by solving the Schrödinger wave equation.
- The Heisenberg uncertainty principle states that values cannot be assigned, with full precision, for position and momentum, or for energy and time, for a particle.
- If a particle has a well-defined de Broglie wavelength, then its momentum is known precisely, but there is no knowledge of its position.
- Measurements of time and energy, and position and momentum are linked variables and described by the energy–time and position–momentum uncertainty principles:

  $$\Delta x \Delta p \ge \frac{h}{4\pi} \quad \text{and} \quad \Delta E \Delta t \ge \frac{h}{4\pi}$$

  where ΔE represents the uncertainty in the measurement of energy, Δx represents the uncertainty in the measurement of position, Δp represents the uncertainty in the measurement of momentum and Δt represents the uncertainty in the measurement of time.

13.2 Nuclear physics

- When a fast-moving alpha particle approaches a gold nucleus head-on, it will be directed straight back along the same path. It will get close to the nucleus but not collide with the nucleus, owing to the action of repulsive electrostatic forces. This effect is known as Coulomb scattering.
- An alpha particle approaching a gold nucleus slows down as it gains electrical potential energy and loses kinetic energy.
- At the closest approach, the alpha particle is temporarily stationary and all its energy is electrical potential energy.
- Applying Coulomb's law at closest approach:

  $\frac{q_1q_2}{4\pi\varepsilon_0 r^2} = E_K$, where q_1 is the charge on an alpha particle, q_2 is the charge on the gold nucleus and r is an upper estimate for the radius of the gold nucleus.
- The $1/r$ hill is a gravitational model showing how the electrical potential varies round a charged particle. The elevation of the hill above the bench top represents the potential and the steepness of the hill represents the field.

- A Bainbridge mass spectrometer can be used to measure the specific charge to mass ratio of a positively charged particle and to detect isotopes of an element and measure their percentage abundances.
- The principle of the mass spectrometer is to use a magnetic field to deflect moving positive ions (in the gaseous state). If a moving ion (accelerated by an electric potential) enters a magnetic field of constant strength it will follow a circular path.
- The magnetic force provides the centripetal force required:

  $$Bqv = \frac{mv^2}{r} \qquad \text{or} \qquad r = \frac{mv}{Bq}$$

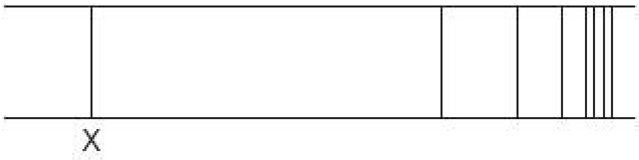
  where B is the magnetic field strength, q the charge on the positive ion, r is the radius of the ion's path, m is the mass of the ion and v is its velocity.
- If the ions have the same charge, q, and they are all selected to be travelling at the same velocity, v, then the radius of the circle will depend on the mass of the ion. An ion with larger mass will travel in a circle with a larger radius.
- The nucleus is a quantum system and its energy values are quantized into discrete energy levels.
- When an alpha particle or a gamma photon is emitted from the nucleus only discrete energies are observed. These energies correspond to the difference between two discrete nuclear energy levels.
- Beta energy spectra are continuous – beta particles are emitted with a range of kinetic energies.
- Neutrinos are small particles with no charge and negligible mass. They pass through matter with little interaction. They carry variable amounts of energy when formed during beta decay.
- Beta positive decay causes no change to the nucleon number of the parent nuclide but causes an increase of one in the proton number.
- The half-life, $T_{1/2}$, of a radioactive nuclide is the time taken for the number of undecayed nuclei to be reduced to half the original number. This is constant for a given isotope.
- The activity of a radioactive source is the number of decays per unit time.
- The activity $\frac{\Delta N}{\Delta t}$ of a radioactive source is related to the number N, of undecayed nuclei by the equation: $\frac{\Delta N}{\Delta t} = -\lambda N$, where λ is the decay constant.
- The decay constant is defined as the probability of decay per unit time of a nucleus. The larger the value of the decay constant the more rapid the radioactive decay.
- The number N of undecayed nuclei in a radioactive sample after time t is given by the equation $N = N_0 e^{-\lambda t}$, where N_0 is the number of undecayed nuclei at the start of timing.
- Since the activity, A, of a source is directly proportional to N, it follows that $A = A_0 e^{-\lambda t}$, where A_0 is the initial activity of the source. Taking logarithms to the base e gives $\ln A = \ln A_0 - \lambda t$.
- A graph of $\ln A$ against t is a straight line, with the gradient equal to $-\lambda$.
- The half-life of a short-lived isotope can be found by measuring the corrected count rate (proportional to activity) over a period of time. The decay constant can be found from a plot of $\ln A$ against t.
- The half-life, $T_{1/2}$, and the decay constant, λ, are related by the equation:

  $T_{1/2} = \frac{0.693}{\lambda}$, which is derived from $N = N_0 e^{-\lambda t}$.
- If the half-life is long, then the activity will be effectively constant over a period of time. The number of nuclei can be calculated from the mass of a pure sample, its atomic mass and Avogadro's constant. The decay equation can then be used to calculate the half-life.

Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 The diagram shows part of a typical line emission spectrum. This spectrum extends through the visible region of the electromagnetic spectrum into the ultraviolet region.



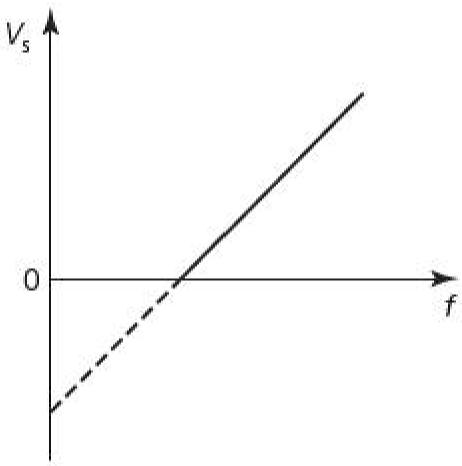
Which statement is true for line X of the emission spectrum?

A It has the highest frequency and is at the ultraviolet end of the spectrum.
B It has the longest wavelength and is at the ultraviolet end of the spectrum.
C It has the shortest wavelength and is at the red end of the spectrum.
D It has the lowest frequency and is at the red end of the spectrum.

Q2 In the photoelectric effect, light incident on a clean metal surface causes electrons to be ejected from the surface.
Which statement is correct?

A The de Broglie wavelength of the ejected electrons is the same as the wavelength of the incident light.
B Electrons are ejected only if the wavelength of the incident light is greater than some minimum value.
C The maximum energy of the electrons is independent of the type of metal.
D The maximum energy of the electrons is independent of the intensity of the incident light.

Q3 Ultraviolet light is shone on a zinc surface and photoelectrons are emitted. The sketch graph shows how the stopping potential V_s varies with frequency f.



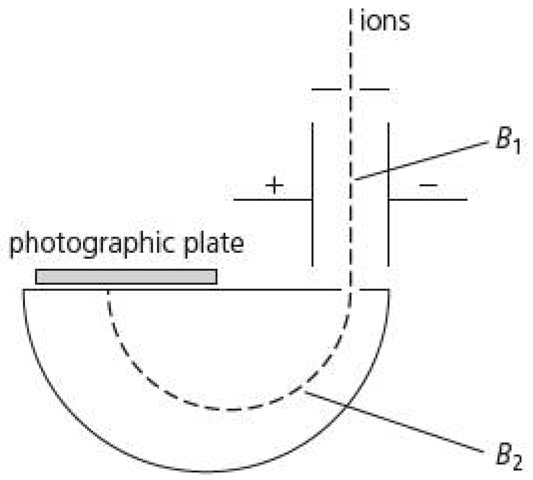
Planck's constant may be determined from the charge of an electron e multiplied by

A the y-intercept.
B the x-intercept.
C the gradient.
D the area under the graph.

Q4 In the Schrödinger model of the hydrogen atom, the probability of finding an electron in a small region of space is calculated from the

A Heisenberg uncertainty principle (time–energy).
B de Broglie hypothesis.
C (amplitude of the wavefunction)^2.
D square root value of the wavefunction.

Q5 The diagram is a schematic representation of the Bainbridge mass spectrometer.
Positive ions are injected between the plates of the speed selector.

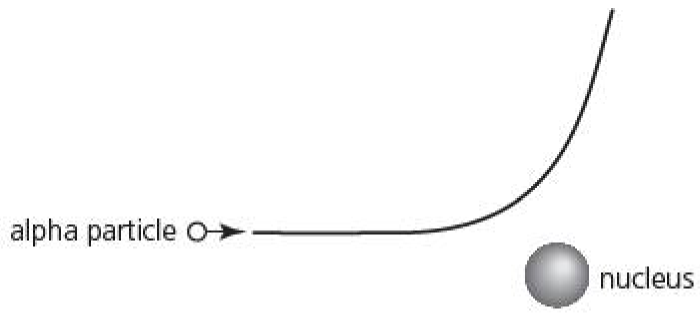


Which of the following correctly shows the direction of the magnetic fields B_1 and B_2?

	B_1	B_2
A	out of the page	out of the page
B	into the page	into the page
C	out of the page	into the page
D	into the page	out of the page

Higher Level Paper, May 09 TZ2, Q30

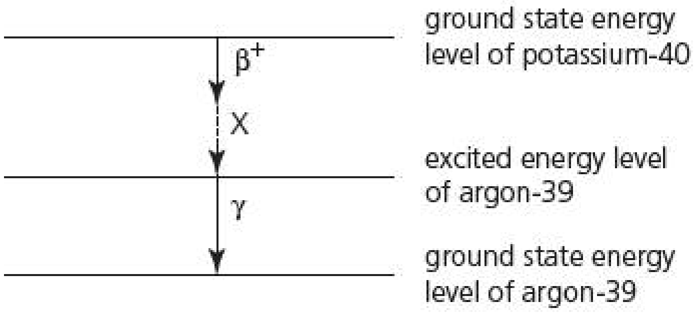
Q6 The diagram shows the path followed by an alpha particle in the vicinity of the nucleus of a gold atom.



Which of the following is correct for the alpha particle?

A The force acting on it is smaller than that acting on the nucleus.
B The force acting on it changes direction.
C Its kinetic energy is constant.
D Its potential energy is constant.

Q7 A nucleus of potassium-40 undergoes β^+ decay to an excited state of a nucleus of argon-39. The argon-39 then reaches its ground state by the emission of a γ-ray photon. The diagram represents the β^+ and γ energy level diagram for this decay process.



The particle represented by the letter X is

A an antineutrino.
B a neutrino.
C an electron.
D a photon.

Higher Level Paper, May 07 TZ1, Q39

Paper 2 IB questions and IB style questions

Q1 a State the de Broglie hypothesis. [2]

b A tennis ball of mass 0.05 kg is thrown with speed 15 m s^{-1} through a gap of width 0.5 m. Discuss with an appropriate calculation, whether the tennis ball will exhibit noticeable wave-like properties in this situation. [3]

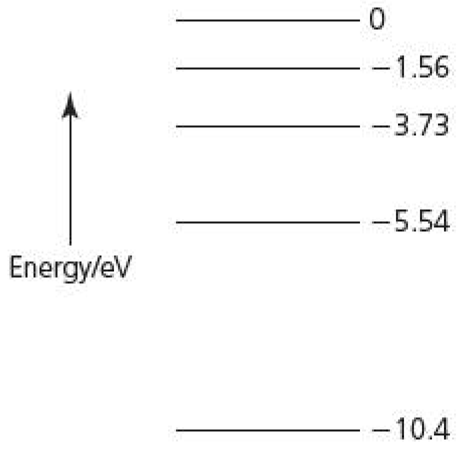
c An electron is accelerated from rest through a potential difference V. Deduce that the de Broglie wavelength λ of the accelerated electron is given by

$$\lambda = \frac{h}{\sqrt{2meV}}$$

where m is the mass of the electron and e its electric charge. [3]

d Calculate the de Broglie wavelength of an electron that has been accelerated from rest through a potential difference of 54 V. [1]

Q2 a The diagram represents some of the energy levels of the mercury atom.



Photons are emitted by electron transitions between the levels. On a copy of the diagram, draw arrows to represent the transition, for those energy levels that gives rise to

i the longest wavelength photon (label this L) [1]
ii the shortest wavelength photon (label this S). [1]

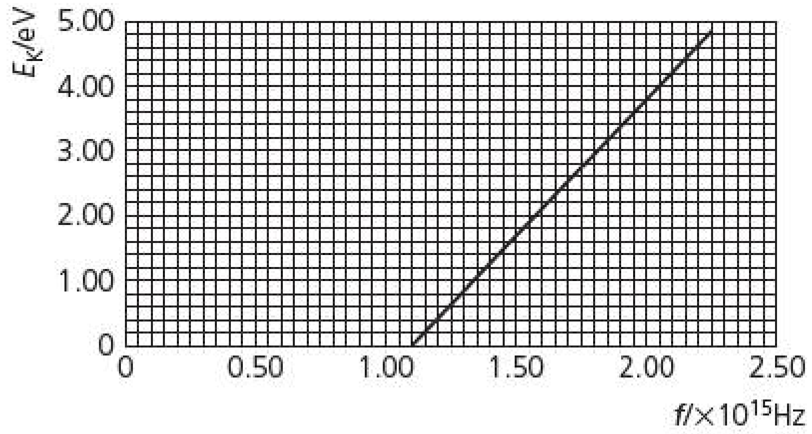
b Determine the wavelength associated with the arrow you have labelled S. [3]

c A nucleus of the isotope bismuth-212 undergoes α-decay into a nucleus of an isotope of thallium. A γ-ray photon is also emitted.
Draw a labelled nuclear energy level diagram for this decay. [2]

d The activity of a freshly prepared sample of bismuth-212 is 2.80×10^{13} Bq. After 80.0 minutes the activity is 1.13×10^{13} Bq. Determine the half-life of bismuth-212. [2]

Higher Level Paper 2, May 09 TZ1, QB1 (Part 2)

Q3 A metal is placed in a vacuum and light of frequency f is incident on its surface. As a result, electrons are emitted from the surface. The graph shows the variation with frequency f of the maximum kinetic energy E_K of the emitted electrons.



a The graph shows that there is a threshold frequency of the incident light below which no electrons are emitted from the surface. With reference to the Planck constant and the photoelectric work function, explain how Einstein's photoelectric theory accounts for this threshold frequency. [4]

b Use the graph in **a** to calculate the

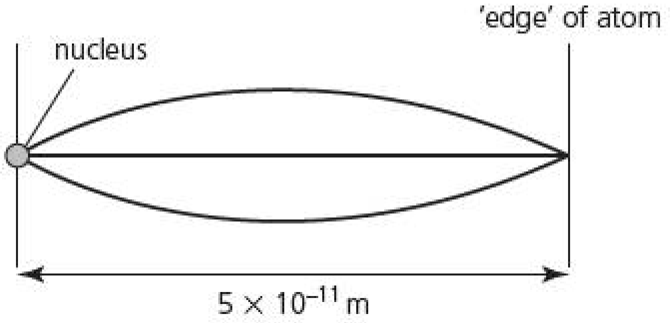
i threshold frequency [1]

ii Planck constant [1]

iii work function of the metal. [2]

Higher Level Paper 2, Specimen Paper 09, QB4 (Part 2)

Q4 In a simple model of the hydrogen atom, the 'size' of the atom determines the kinetic energy of the electron. Its de Broglie wavelength is equal to the wavelength of the standing wave bounded by the nucleus and the 'edge' of the atom, as shown below.



The 'edge' of the atom is 5×10^{-11} m from the nucleus.

a i State the de Broglie wavelength of the electron. [1]

ii The 'edge' of the hydrogen atom is moved closer to the nucleus. Describe what changes occur in the kinetic energy of the electron. [2]

A different model of the hydrogen atom takes into account the fact that the electrical potential energy of the electron depends on its distance from the nucleus.

b i Explain the variation with the distance from the nucleus of the electrical potential energy of the electron. [3]

ii Use your answer to **b i** to explain the variation with distance from the nucleus of the kinetic energy of the electron. [2]

iii Use your answer to **b ii** to suggest how the wavelength of the standing wave of the electron varies with distance away from the nucleus. [3]

Higher Level Paper 2, Nov 05, QB3 (Part 2)

14 Digital technology

STARTING POINTS

- A flow of electrons is an electric current.
- Light travels in straight lines from a source in the form of progressive waves.
- Light waves are described by their amplitude (brightness or intensity), wavelength (colour), frequency and speed.
- The frequency of a wave is the number of waves that pass a point in one second.
- The wave equation relates the speed, frequency and wavelength of a wave. Speed = frequency × wavelength. $v = f\lambda$
- Light waves can exhibit the properties of reflection and interference.
- Light (and other electromagnetic radiations) can be regarded as a stream of energy 'packets' known as photons.
- The energy of a photon is directly proportional to its frequency: $E = hf$
- Interference patterns may be produced by the superposition of two or more waves.
- Constructive interference occurs when two sets of waves meet at positions where the waves arrive in phase; crests/troughs of the waves meet and produce larger crests/troughs.
- Constructive interference occurs at positions where the path difference is equal to 0, λ, 2λ, etc., where λ represents the wavelength.
- Destructive interference occurs when two sets of waves meet at positions where the waves arrive exactly out of phase; crests of a wave meet the troughs of the other and cancel each other out.
- Destructive interference occurs where the path difference is an odd multiple of $\frac{\lambda}{2}$.
- Diffraction of a wave is greatest when it passes through a gap comparable to its wavelength.
- The angles at which waves passing through an aperture are diffracted depend on the ratio of wavelength to aperture size.
- Resolution is the ability to distinguish the images of two separate objects from one another.

Introduction

In earlier chapters we met examples of uses of electricity for lighting, heating, cooling (air conditioners and refrigerators) and in electric motors. More recently, electricity has also been vital to digital technology, which has made huge changes to the way people study, communicate, work and relax.

Digital technology is also the basis for digital devices, such as cameras and video recorders, MP3 players, CD players, games machines, mobile phones, tablets and computers. Digital devices also include input, output and communication devices that allow computer access to the vast amount of information and data that can be found on the Internet (the World Wide Web).

14.1 Analogue and digital signals

When we communicate, we send each other information. Information comes in many forms, for example:

- text (letters, numbers and symbols)
- still pictures, for example, drawings, paintings and photographs
- speech, sound and music
- video and animation.

The transfer of information needs an agreed and recognized code. For example, the sounds of words, smoke signals, written symbols and the sequence of coloured lights at traffic signals are all examples of codes. The transfer of information between modern electronic devices requires the use of standard codes (e.g. the ACSCII code). Standard codes are very important because they allow digital devices from different manufacturers to share data. The use of electronic technology for the storage, processing, validation (checking) and transfer of information involves converting the original information into a different form, usually an electrical signal, using a coding process, as shown in Figure 14.1. The electrical signal can be converted back when needed. This can be done using digital or analogue techniques.

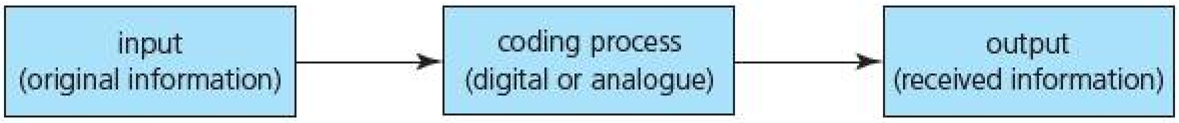


Figure 14.1 The coding of information

Any information signal that has the same variations with time as the information itself is known as an **analogue signal**. For example, the voltage produced by a microphone in a recording studio varies in the same way as the sound waves that are detected by the microphone. The voltage variation is due to the loudness or amplitude of sound detected by the microphone. The voltage output from the microphone is an analogue signal. The graph in Figure 14.2 represents an analogue signal from a microphone that varies continuously within the range +6 V and –6 V.

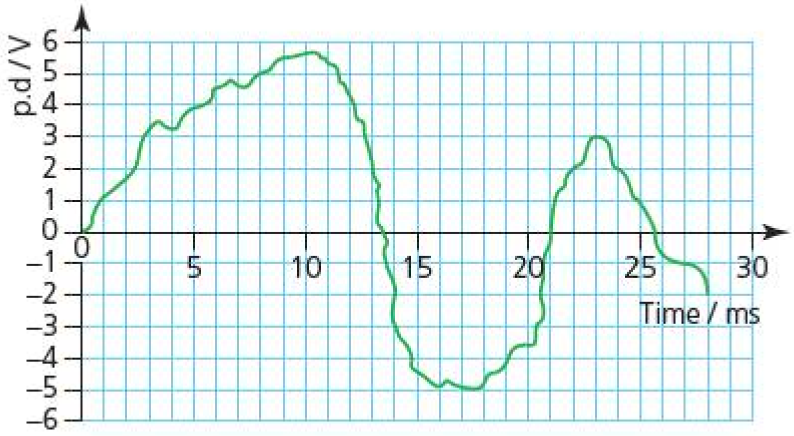


Figure 14.2 An analogue signal

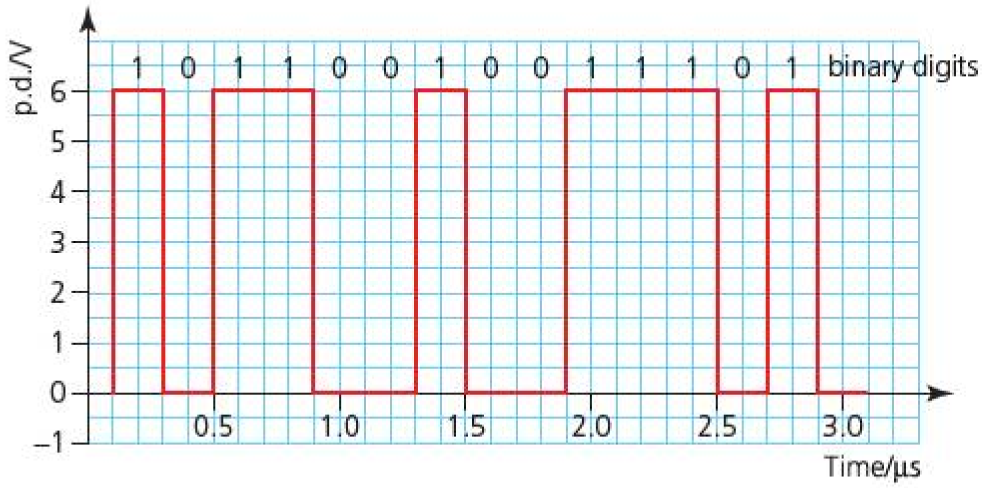


Figure 14.3 A digital signal

A **digital signal** consists of a series of 'highs' and 'lows' with no values between the 'highs' and 'lows'. The discrete data in the signal is transmitted as a series of binary ones (1s) and binary zeros (0s). The binary number system (base 2) represents numbers using two symbols, 0 and 1. The signal in Figure 14.3 is a digital signal which has two values only: +6 V or 0 V, representing respectively a binary one or binary zero. Each '1' or '0' lasts for 0.2 microseconds.

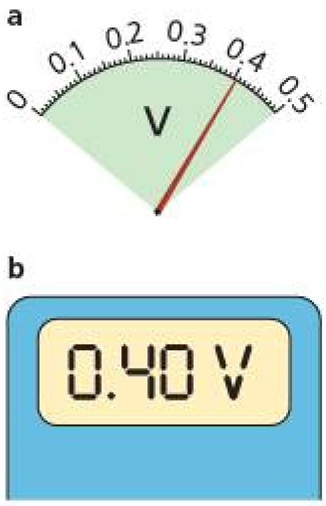


Figure 14.4 Two different voltmeters reading the same value of potential difference:
a analogue voltmeter
b digital voltmeter

Figure 14.4 shows an analogue and a digital voltmeter reading the same value of potential difference. The analogue voltmeter shows *all* values of potential difference from the minimum to the maximum value on the scale. Intermediate values can be found by interpolating (reading between) the gradations (marks) of the scale. The digital voltmeter is accurate to ± 0.01 V. The digital voltmeter reading is said to be quantized in steps of 0.01 V.

There are many advantages in using digital electronic circuits to store, process and transmit data rather than analogue circuits. Today virtually all communication systems and electronic devices are digital.

Analogue signals can be converted to digital signals by an analogue-to-digital converter (ADC). This encodes or changes the analogue signal into a digital form via a process known as quantization.

Computers can only process and store digital signals. Figure 14.5 shows how a computer can be used to log, convert and display data from a temperature sensor.

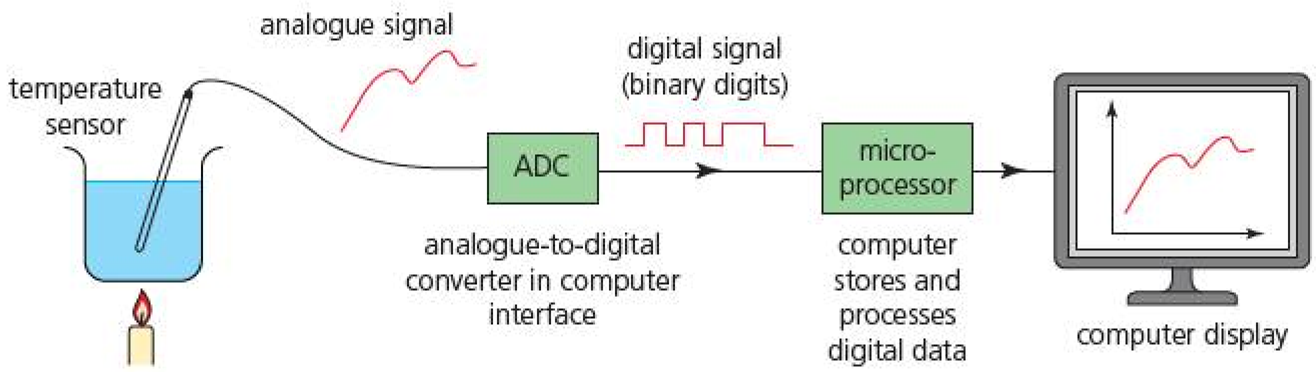


Figure 14.5 The use of a temperature sensor and a data logger

1. Research the use of semaphore flags and Morse code in coding text messages.
2. Give examples of analogue and digital measurements from your practical investigations.
3. Find out about the Google projects to digitally store, by scanning, thousands of works of literature and famous artwork from libraries and art galleries all over the world.

Binary numbers

14.1.1 Solve problems involving the conversion between binary numbers and decimal numbers.

Table 14.1 Decimal and four-bit binary numbers

Decimal number	Binary or digital number
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111

A binary number is a base 2 number; a decimal number is a base 10 number. Table 14.1 shows some decimal numbers and their equivalents in base 2 or binary notation. A binary number is formed from a number of binary digits, or **bits**. All the binary numbers shown in Table 14.1 are four-bit numbers. Zeros are added as extra bits to the left of any number that is less than four bits. Binary numbers may also be called digital numbers. Digital signals are often four-bit binary numbers, or multiples of them: eight-bit, 16-bit, 32-bit, or even 64-bit.

Larger numbers would require digital numbers with more bits. When reading a digital number, the bit on the left-hand side of the digital number is the **most-significant bit** (MSB). This binary digit or bit has the highest value. The binary digit or bit on the right-hand side has the least value and is known as the **least-significant bit** (LSB).

When the LSB is 1 and all other bits are 0, this corresponds to decimal number 1. When the second bit is binary 1, and all other bits are 0, this corresponds to the decimal number 2. When successive bits show binary 1 they correspond to decimal numbers 4, 8, 16, 32, 64, etc. (Table 14.2).

Table 14.2 Decimal numbers and their five-digit binary equivalents showing the powers to the base 2

Base 10	2^4	2^3	2^2	2^1	2^0	Base 2
0	0	0	0	0	0	00000
1	0	0	0	0	1	00001
2	0	0	0	1	0	00010
3	0	0	0	1	1	00011
4	0	0	1	0	0	00100
5	0	0	1	0	1	00101
6	0	0	1	1	0	00110
7	0	0	1	1	1	00111
8	0	1	0	0	0	01000
9	0	1	0	0	1	01001
10	0	1	0	1	0	01010
11	0	1	0	1	1	01011
12	0	1	1	0	0	01100
13	0	1	1	0	1	01101
14	0	1	1	1	0	01110
15	0	1	1	1	1	01111
16	1	0	0	0	0	10000
17	1	0	0	0	1	10001
18	1	0	0	1	0	10010
19	1	0	0	1	1	10011
20	1	0	1	0	0	10100

The binary number 1101 corresponds to the decimal number 8 + 4 + 0 + 1, or 13. The decimal number 11, which equals 8 + 0 + 2 + 1, corresponds to the binary number 1011.

Figure 14.6 shows how the place value you are familiar with in base 10 is used in base 2.

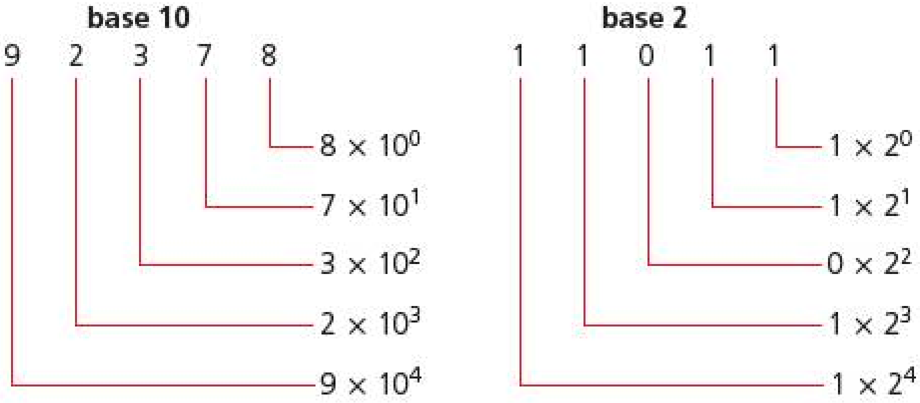


Figure 14.6 A decimal number and a binary number showing base 10 and base 2 place values

A binary signal containing a series of binary 1s and 0s is known as a **word**. An eight-bit binary number or word occupies one **byte** of computer memory. Values stored on a compact disc (CD) are represented by a two-byte number (16 bits). Computer memory is measured in kilobytes (KB), megabytes (MB) and gigabytes (GB). 1 KB = 1024 (2^{10}) bytes; 1 MB = 1048 576 bytes (2^{20}) and 1 GB = 1 073 741 824 (1024^3 or 2^{30}) bytes.

Additional Perspectives

ASCII code

A particular number of binary digits (bits) represents a fixed number of different values. Each additional bit doubles the number of different possible values that can be represented (Table 14.3).

English text contains a mixture of letters (combined to form of words), numbers, symbols and punctuation marks. Additional characters are also needed to represent formatting, such as new line and new paragraph. Computers and the Internet use a code known as ASCII (American Standard Code for Information Interchange). It is an eight-bit code representing 256 different possible characters and formatting codes. For example, 1000000 represents the 'at' symbol @ and 1000011 represents the letter C.

ASCII has been superseded by Unicode, a double byte (16-bit) character system designed to store and display a much wider range of letters and symbols. The extras include foreign languages, such as Mandarin, and mathematical/scientific symbols, plus space for future expansion.

Question

1 Research the problems that Unicode has in representing Chinese characters.

Table 14.3 Number of binary digits with the associated number of different possible digital values

No. of bits used	Bits used	No. of different possible values
1	0, 1	2
2	00, 01, 10, 11	4 (2^2)
3	000, 001, 010, 011, 100, 101, 110, 111	8 (2^3)
4	etc.	16 (2^4)
5	etc.	32 (2^5)
6	etc.	64 (2^6)
7	etc.	128 (2^7)
8	etc.	256 (2^8)
9	etc.	512 (2^9)
10	etc.	1024 (2^{10})

Worked examples

1 What is the least number of binary digits needed to express each of the decimal numbers (base 10) 10 and 2038 as a binary number?

10 – The *smallest* power of 2 that exceeds 10 is 4, since $2^4 = 16$ and is greater than 10. Hence four bits must be used in the binary representation of 10 (1010).

2038 – The smallest power of 2 that exceeds 2038 is 11, since $2^{11} = 2048 > 2038$. Hence 11 bits must be used in the binary representation of 2038 (11111110110).

2 How many decimal numbers can be represented with eight-bit words?

With eight bits the maximum number is $2^8 = 256$. These are the numbers from 0 to 255.

3 Express the six-bit binary number 11101 as a decimal number.

The binary number is shown below. The correct power of 2 has been placed under each binary digit.

1	1	1	0	1
2^4	2^3	2^2	2^1	2^0

$$\begin{aligned} \text{So } 11101 &= (1\times 2^4) + (1\times 2^3) + (1\times 2^2) + (0\times 2^1) + (1\times 2^0) \\ &= 16 \quad + \quad 8 \quad + \quad 4 \quad + \quad 0 \quad + \quad 1 \\ &= 29 \end{aligned}$$

An alternative presentation of the working is shown in Figure 14.7.

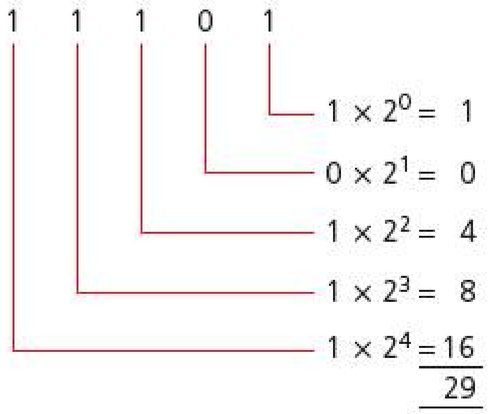


Figure 14.7 Converting binary 11101 into a decimal number

4 Write the decimal number 62 as a binary number using eight binary digits (bits).

You are told to use eight bits, so the highest power of 2 is 2^7 (128).

Draw up a table of powers of 2 up to 2^7. Then subtract the largest possible power of 2 from 62, and keep subtracting the next largest possible power from the remainder, marking 1s in each column in the table where this is possible and 0s where it is not (see Figure 14.8).

$$\begin{array}{r} 62 \\ -32 \\ \hline 30 \\ -16 \\ \hline 14 \\ -8 \\ \hline 6 \\ -4 \\ \hline 2 \\ -2 \\ \hline 0 \end{array}$$

128	64	32	16	8	4	2	1
0	0	1	1	1	1	1	0

Figure 14.8 Converting decimal 62 (base 10) to binary (base 2)

This gives you the number in terms of its coefficients and powers of 2:

$$62 = (0 \times 2^7) + (0 \times 2^6) + (1 \times 2^5) + (1 \times 2^4) + (1 \times 2^3) + (1 \times 2^2) + (1 \times 2^1) + (0 \times 2^0)$$
$$(0 \times 128) + (0 \times 64) + (1 \times 32) + (1 \times 16) + (1 \times 8) + (1 \times 4) + (1 \times 2) + (0 \times 1)$$
$$= 00111110$$

5 The letter A is typed every second into a text document. The letter A is stored as an eight-bit binary digit (ASCII code). How long would it take in days (to the nearest integer) to fill a 2 GB thumb drive?

$2\,GB = (2\,GB \times 1024\,MB \times 1024\,kB \times 1024\text{ bytes} \times 8)\text{ bits} = 1.7179869184 \times 10^{10}$ bits

One 'A' uses eight bits, so (171 798 691 840 / 8) = 2 147 483 648 bytes.

This will take $\dfrac{2147483648}{(60 \times 60 \times 24)} = 24855$ days

4 Convert the following decimal numbers into eight-bit binary numbers:
a 8 **b** 14 **c** 17 **d** 68 **e** 125

5 Convert the following six-bit binary numbers to base 10 (decimal) numbers:
a 000110 **b** 100101 **c** 110010 **d** 111111

6 The hexadecimal system (base 16) is particularly important in computer programming since four binary digits can be easily expressed using a single hexadecimal digit. Describe the hexadecimal system.

7 Table 14.4 shows part of the ASCII code. Write the word CAB in ASCII code.

8 Find out how multiplication and division are performed in binary (base 2).

9 Find out about the history and development of counting systems and numerals. Ensure that your research includes the work of the Babylonians, Egyptians, Arabs, Indians and the Maya.

Table 14.4

Letter	ASCII code
A	0100 0001
B	0100 0010
C	0100 0011
D	0100 0100
E	0100 0101

10 Construct a spreadsheet using Excel that will interconvert decimal and binary numbers. Use the BIN2DEC and DEC2BIN functions. You will need to make sure the Analysis Toolpak is installed.

11 Find out about Grid computing and the processing and storage of digital data at CERN in Geneva, where high-energy particle physics experiments are performed.

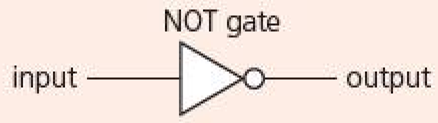
12 Make a list of digital devices that you use regularly and sort the devices into work/learning/study and leisure/fun/entertainment. Suggest what new digital devices may be developed in the future.

Additional Perspectives

Logic gates

Logic gates are switching circuits found in computers and other electronic devices. They 'open' and give a 'high' voltage (a signal, 6 V, represented by binary 1) depending on the combination of voltages at their inputs. There are three basic types (NOT, OR and AND) and the behaviour of each is described by a *truth table* showing what the output is for all possible inputs. 'High' (e.g. 6 V) and 'low' (e.g. 0 V) outputs and inputs are represented by a binary 1 and 0 respectively, and are referred to as logic levels 1 and 0.

The NOT gate is the simplest logic gate, with only one input and one output. It produces a 'high' output if the input is 'low'. Whatever the input signal, the gate *inverts* the signal. The symbol and truth table are given in Figure 14.9.



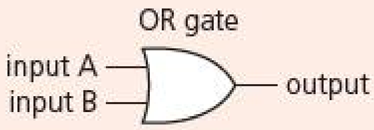
Input	Output
0	1
1	0

Figure 14.9 NOT logic gate symbol and truth table

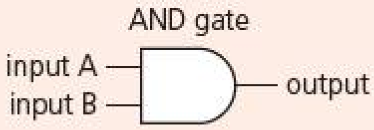
The AND and OR logic gates have two inputs and one output. They behave according to the following statements:

OR	output is 1 if input A OR input B OR both are 1
AND	output is 1 if input A AND input B are 1

The truth tables and diagrams used for the OR gate as well as the AND gate are shown in Figure 14.10.



Input A	Input B	Output
0	0	0
0	1	1
1	0	1
1	1	1



Input A	Input B	Output
0	0	0
0	1	0
1	0	0
1	1	1

Figure 14.10 Symbols and truth tables for an OR and an AND gate

Logic gates can be used as processors in electronic control systems. The block or system diagram that might be used by a jeweller to protect an expensive gold watch is shown in Figure 14.11. The gold watch is placed on a push switch (pressure sensor) which sends a binary 1 to the NOT gate *unless* the gold watch is lifted, in which case a binary 0 is sent. If a binary 0 is sent, then the output from the NOT gate is a binary 1, which rings the buzzer.

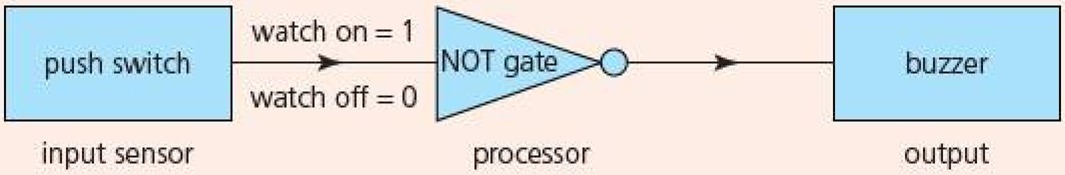


Figure 14.11 A simple alarm system

Question

1 Deduce the truth table for an AND gate whose output is connected to a NOT gate.

Storage of information

14.1.2 Describe different means of storage of information in both analogue and digital forms.

Analogue and digital data must be stored in a permanent 'machine-friendly' form. This then allows the data to be accessed and used many times. Data storage can be physical – the shape of a vinyl record groove or the tracks of pits on a CD or DVD – or it can be magnetic, as in the magnetic patterns on a cassette tape or computer disk.

LPs

Vinyl records, gramophone records or LPs (long players) and cassettes (Figure 14.12) were popular formats for storing recorded music until compact discs (CDs) were introduced in the 1980s. The analogue musical information in an LP is stored in a groove cut into the plastic (vinyl) record by a sharp stylus (needle). The shape of the wiggles in the groove is an analogue representation of the musical signal. Loud sounds are stored as large wiggles, and quiet sounds as small wiggles. The closer together the wiggles are, the higher the frequency of a sound. The average LP contains approximately 470 metres of groove on each side.

The LP is played back by placing another stylus into the groove and then rotating the record at constant speed on a turn table (Figure 14.13). The groove in the LP spirals inwards so the needle moves through the groove faster when it is playing music on the edge of the LP. This means that the groove is more squashed in the centre of the LP than at the outer edge to compensate for the decrease in speed as the record continues to play.

Figure 14.12 Audio cassette tape and LP (vinyl record)

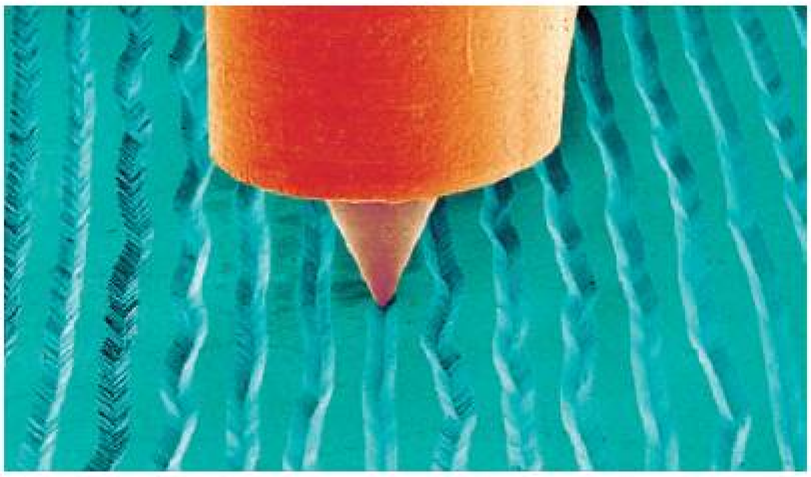


Figure 14.13 Retrieval of analogue information from an LP record

The stylus vibrates in the groove and produces an electrical signal (due to the presence of a piezoelectric crystal), which is played back through an amplifier and speakers. Every time the LP is played, the stylus slightly damages and changes the shape of the groove. In addition, dust and smoke particles settle in the groove. The distorted and dust-filled groove causes unintended vibrations of the stylus and so unwanted electronic 'noise' is added to the signal. The noise is heard as a background hissing sound when the record is played. A relatively large LP may also warp (bend), which distorts the recording.

Cassette tapes

Music (or video) can be recorded on a long thin plastic tape coated with a fine powder of magnetic material (oxides of iron or chromium) which becomes magnetized during recording. A chain of tiny permanent magnets is produced on the tape in a pattern which represents the original sound (or picture). Figure 14.14 shows the simple magnet patterns for single 'high' and 'low' frequencies.

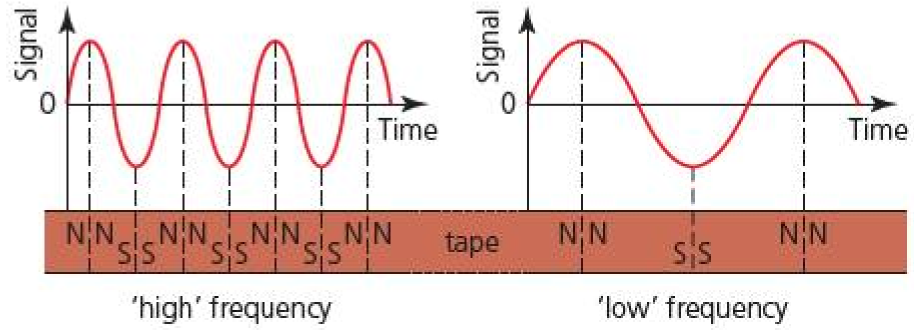


Figure 14.14 Magnetization of audio or video tape

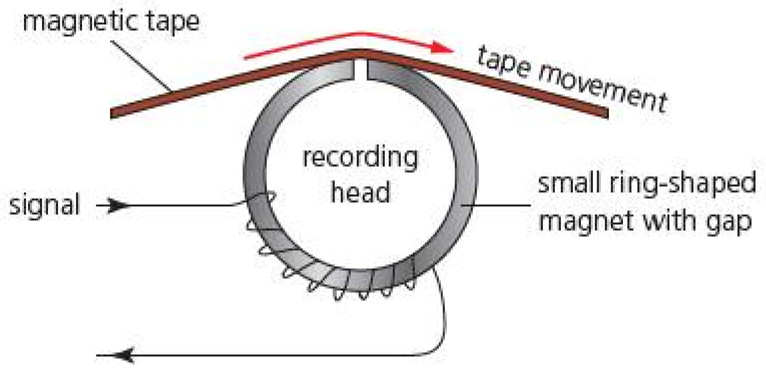


Figure 14.15 The recording head in a cassette tape recorder

The magnetizing is done by the recording head (Figure 14.15), which consists of a small electromagnet. The information to be recorded is sent as a small, changing magnetic field which magnetizes the tape particles in the same pattern as the current in the coil.

To retrieve the information from the tape, the tape is pulled past the electromagnet (which also acts as the playback head). The pattern of magnetization on the tape produces a varying magnetic field in the core of the head, which then induces a varying electric current in the coil. This current is amplified, so the signal can be output through a loudspeaker. Cassette tapes are generally analogue in nature, but digital cassette systems have also been developed.

13 Find out about the use of tape drives to back up computer systems or large data files.

Floppy disks

Figure 14.16 A 3½ inch floppy disk

The floppy disk (Figure 14.16), like the cassette, is a magnetic form of data storage. The first floppy disks, developed in the 1960s, were 8 inches in diameter. By the end of the 20th century they were 3½ inches in diameter and could store 1.44 megabytes of digital data. Floppy disks are now virtually obsolete because of their relatively low capacity to store data, but have some use with old computer systems.

The data on a floppy disk is stored in patterns of magnetic particles arranged in concentric rings (called tracks). The floppy disk reader is able to access data on any track without having to search sequentially through the other tracks as on an analogue cassette tape. The floppy disk (and hard disk, see below) are examples of direct access storage devices. In a direct access storage device bits of data are stored at precise locations, enabling the computer to retrieve information directly without having to scan a series of records.

14 Find out how 'flash memory' works. Describe the forms it takes and the types of devices that use it.

Hard disks

Large amounts of data can be stored in magnetic form on a **hard disk**, for example in computers and in digital video cameras. The hard disk of a personal computer contains a stack of platters (Figure 14.17) which spin at high speed. The platters are rigid (stiff) and are coated with iron oxide particles. A head, which is an electromagnet that has reversible polarity, is used to put digital information on the disk surface. In one polarity the head aligns the magnetism in a tiny area of the disk in one direction so that information is stored as a binary one ('on'), and in the opposite polarity the information acts as a binary zero ('off').

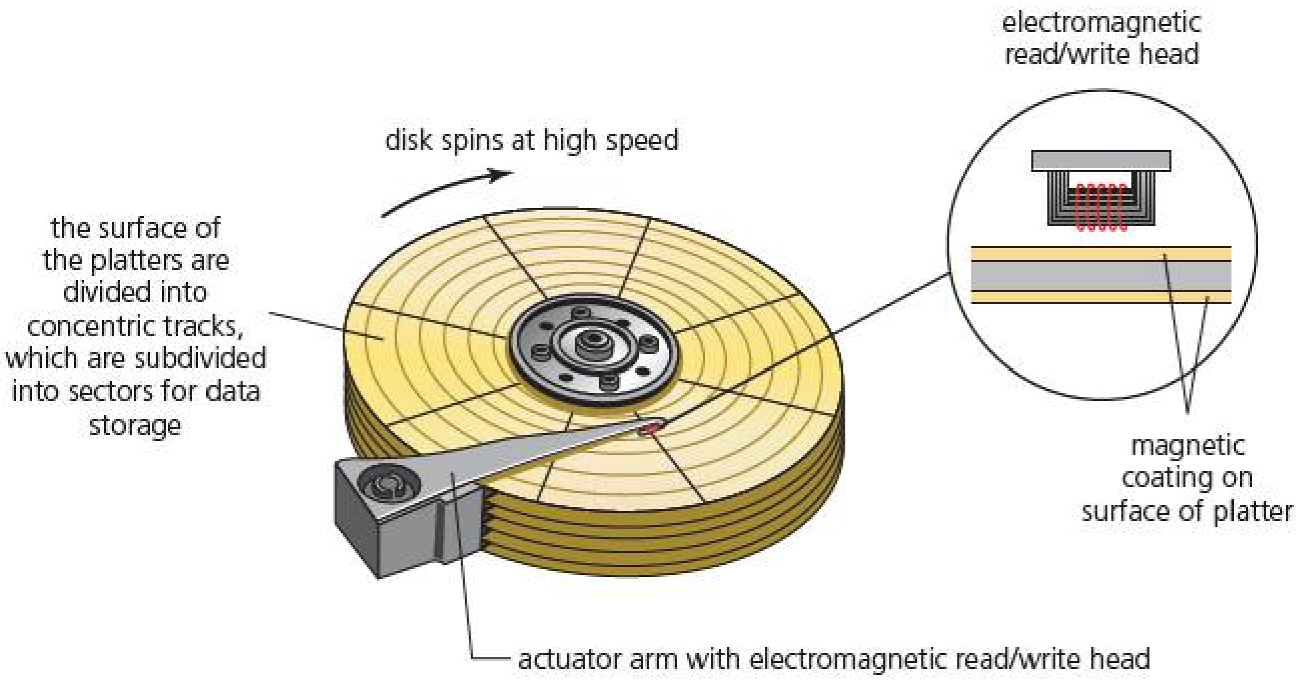


Figure 14.17 The spinning platters in the hard disk of a personal computer

To read the data, the platter or disk spins past the head, which induces a current in the electromagnet. If dust or smoke particles were to get into the drive, the disk surface would be irreversibly damaged. For this reason, the platter and head assembly is sealed during manufacture.

Compact discs

An optical storage medium uses light, usually in the form of a laser, to read and/or write data. The compact disc (CD) is an optical storage medium. CDs have been one of the most popular formats for storing digital sound recordings (music), large computer files and movies. CDs can store about 0.75 GB (gigabytes) of digital data and so are very convenient for transferring and storing digital information.

The data on a CD is stored in a track of microscopic 'pits' and 'lands', which are moulded into a thin layer of transparent plastic (Figure 14.18). These indentations are then covered with a thin layer of reflective aluminium. The pits and lands are arranged one after another in a spiral track, starting from the centre of the disc.

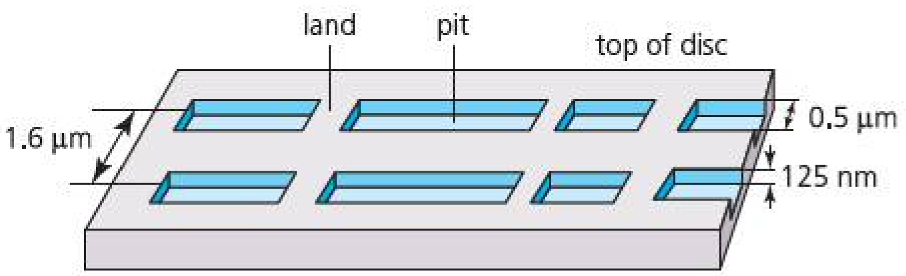


Figure 14.18 Diagram showing the three dimensional arrangement of 'pits' and 'lands' in a track on a CD (not to scale). The length of the pits ranges from 830 to 3560 nm

The disc rotates at about 500 revolutions per second, and a low power laser beam 'reads' the data off the track with its series of pits and lands. (The laser reads from the bottom of the disc, so the pits are in fact raised compared to the lands.) The pattern of pits forms a coded signal of 16 binary numbers, each representing one feature of the sound wave (if the CD stores music).

The laser starts reading the data from the centre and moves outward along a radial line as the disc rotates rapidly next to it. There is no mechanical contact between the laser and the CD disc, so it can be replayed repeatedly without introducing noise.

It is important that the laser reads the pits at a constant rate, otherwise it might misread the digital information in the pits and lands. If the CD is rotated at a constant rate, the speed at which the pits and lands pass the head would increase as the laser moved to the outer edge, so the CD rotation is slowed down (by a motor) as the laser reader moves outward.

15 Describe the similarities and differences between LPs, cassette tapes and CDs.

16 Find out how music from an LP can be transferred to a computer.

Additional Perspectives

Error correction in CD players

The music on an audio CD does not play in sequence along the track. The binary digits corresponding to a small piece of a music track are scattered around the disc. This means that if a portion of the CD is damaged, only a very small part of any section of music is lost. This is known as interleaving and it means that the laser reading head has to be accurately guided or steered around the disc; it is moving laterally as the disc is playing. This means that the disc speed is constantly changing, which requires high-quality motors and excellent control electronics. The CD contains digital information that allows the laser reading head to keep track of its location.

Question

1 Find out about the role of error correcting code (ECC) on an audio CD.

DVDs

DVD (digital video disc or digital versatile disc) is an optical disc storage format similar to the CD. Its main uses are video and data storage. DVDs are of the same dimensions (just under 12 cm) as compact discs, but are capable of storing almost seven times as much data.

DVD uses a 650 nm wavelength laser diode light as opposed to 780 nm for CD. This allows a smaller pit to be etched on the media surface compared with CDs (0.74 μm for DVD versus 1.6 μm for CD), which gives the DVD a greater storage capacity. The track length of a DVD is about twice as long as the track length of a CD.

DVD+R DL is a DVD format that contains double the information of a DVD by having two layers of pits. The top layer is coated with a semi-reflective coating, enabling light to also pass through to read the bottom layer. This DVD format is able to store 17 GB of data.

TOK Link: Digital representation

René François Ghislain Magritte (1898–1967) was a Belgian surrealist artist. He became well known for a number of thought-provoking images including *La trahison des images* (Figure 14.19), which shows a pipe and the words *Ceci n'est pas une pipe*, which translates into 'This is not a pipe'. This seems a contradiction, but is actually true: the painting is not a pipe, it is an image of a pipe. This reminds us that the digital information stored in digital storage devices is a *representation* of words, music and images. The image is reduced to binary 0s and 1s and the 'essence' of the image is lost. It is not the 'real' pipe.

Question

1 What important reminders does this painting give us about how we represent the world in physics?

Figure 14.19 *La trahison des images* (The Treachery of Images) (1928/29)

Recovery of information on a CD

14.1.3 Explain how interference of light is used to recover information stored on a CD.
14.1.4 Calculate an appropriate depth for a pit from the wavelength of the laser light.

The laser beam is focused on the CD surface, and is reflected back onto a detector (photodiode). How it is reflected depends on whether it falls on a pit or a land. If the laser beam is entirely incident on a pit (or on a land) then all the waves in the reflected beam are in phase with each other. Constructive interference takes place, and a strong signal (a binary one) is detected by the photodiode.

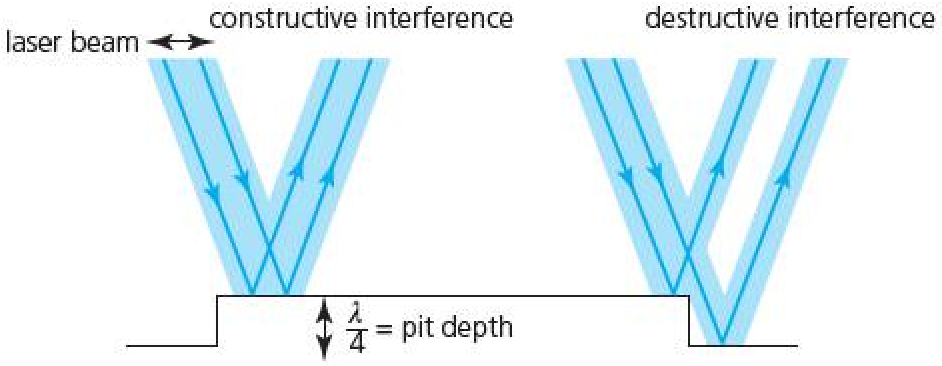


Figure 14.20 The reflection of the laser beam from a pit and from a land. Here the light is shown falling at an angle, but in a CD player it is *almost* normal to the disc surface

If part of the light beam is incident on a pit and part on a land, then there is a path difference between the two parts of the laser beam. The pit depth is such that, for a laser beam of wavelength λ, the path difference is $\lambda/2$, which means that destructive interference occurs and there is no signal produced (a binary zero) at the photodiode. Since the path difference is twice the pit depth, to obtain a path difference of $\lambda/2$:

$$\text{pit depth} = \frac{\lambda}{4}$$

As the laser beam reflects off the rotating spiral track, the signal received by the photodiode changes as the beam travels from pit to land to pit. This produces digital signals of 1s and 0s varying according the whether the interference is constructive or destructive, and also according to the lengths of the pits and lands. If the CD is an audio CD storing music, then a digital-to-analogue converter (DAC) converts this digital signal back into the corresponding analogue signal.

Calculating the depth for a pit from the wavelength of the laser light

We have seen that the depth of the pit on a CD track must be a quarter wavelength of the laser light used to read the CD so that destructive interference can occur. The lasers used for CDs, DVDs and Blu-ray all have different wavelengths, so the discs themselves must be made differently.

Worked example

6 Laser light of frequency 3.80×10^{14} Hz is used in the laser of a CD-ROM reader. Calculate an appropriate depth of a pit on a CD.

$$\lambda = \frac{c}{f} = \frac{3.00 \times 10^8}{3.80 \times 10^{14}} = 7.89 \times 10^{-7}\,\text{m}$$

$$\text{Depth of pit} = \frac{\lambda}{4} = \frac{7.89 \times 10^{-7}}{4} = 1.97 \times 10^{-7}\,\text{m}$$

17 If the laser light had a wavelength of 620 nm in plastic, what depth of pit would the CD have? Explain your answer.

18 If the CD rotates at a rate of 600 revolutions per minute (rpm) when the laser pickup is 2 cm from the centre, how fast should it rotate when the pickup is 6 cm from the centre?
(Recall from the discussion about circular motion in Chapter 4 that the speed of a rotating body is given by: $v = \omega r$ where $\omega = 2\pi f$.)

19 Find about holographic discs and describe how digital data is stored in them.

Additional Perspectives

Conversion of analogue to digital signals

Analogue signals can be easily and quickly converted into digital signals. In an analogue-to-digital converter, the analogue voltage is *sampled* at regular intervals of time. The sampling frequency (rate) is the number of samples in unit time. The value of the sample voltage measured at each sampling time is converted into a binary (digital) number that represents the voltage value.

For example, if a four-bit number is being used, then the number representing a signal that is sampled as 5.0 V would be 0101. When sampling, the number representing the sample would be the whole number (integer) below the actual value of the sampled voltage. If the signal were to be sampled as 11.4 V, then the four-bit number would be 1011. A sampled signal of 11.6 V would also be 1011.

Figure 14.21a shows an analogue signal that is to be sampled at a sampling rate of 10 kHz. A four-bit system is used for the binary (digital) numbers generated. The sample voltages are shown in Figure 14.21b. These sample voltages are converted into a digital signal, shown in Figure 14.21c, by the analogue-to-digital converter (ADC). After this digital signal has been transmitted, it is converted back into an analogue signal using a digital-to-analogue converter (DAC). Complex programs (algorithms) are required for converting analogue signals into digital signals and vice versa.

Figure 14.21d shows the analogue signal that has been recovered. The recovered signal has large 'steps'. It is known as a **pulse amplitude modulated** (PAM) signal. The size of these steps can be reduced and hence the accuracy of the reproduction of the initial analogue signal can be improved by using more voltage levels (greater quantization) and sampling at a higher frequency.

The number of bits in each binary number limits the number of voltage levels (**quantization levels**). In this conversion there are four bits and 16 (2^4) levels. An eight-bit number would give 256 (2^8) levels and a 16-bit number would give 65 536 (2^{16}) levels.

For good quality reproduction of music, the higher audible frequencies must be present; that is, frequencies up to about 20 kHz. For compact discs (CDs) the sampling frequency is 44.1 kHz. This quality of reproduction is not required for speech, and it would be expensive. For walkie-talkies, intercoms and the telephone system the sampling frequency is 8 kHz and the highest frequency to be transmitted is limited to 3.4 kHz.

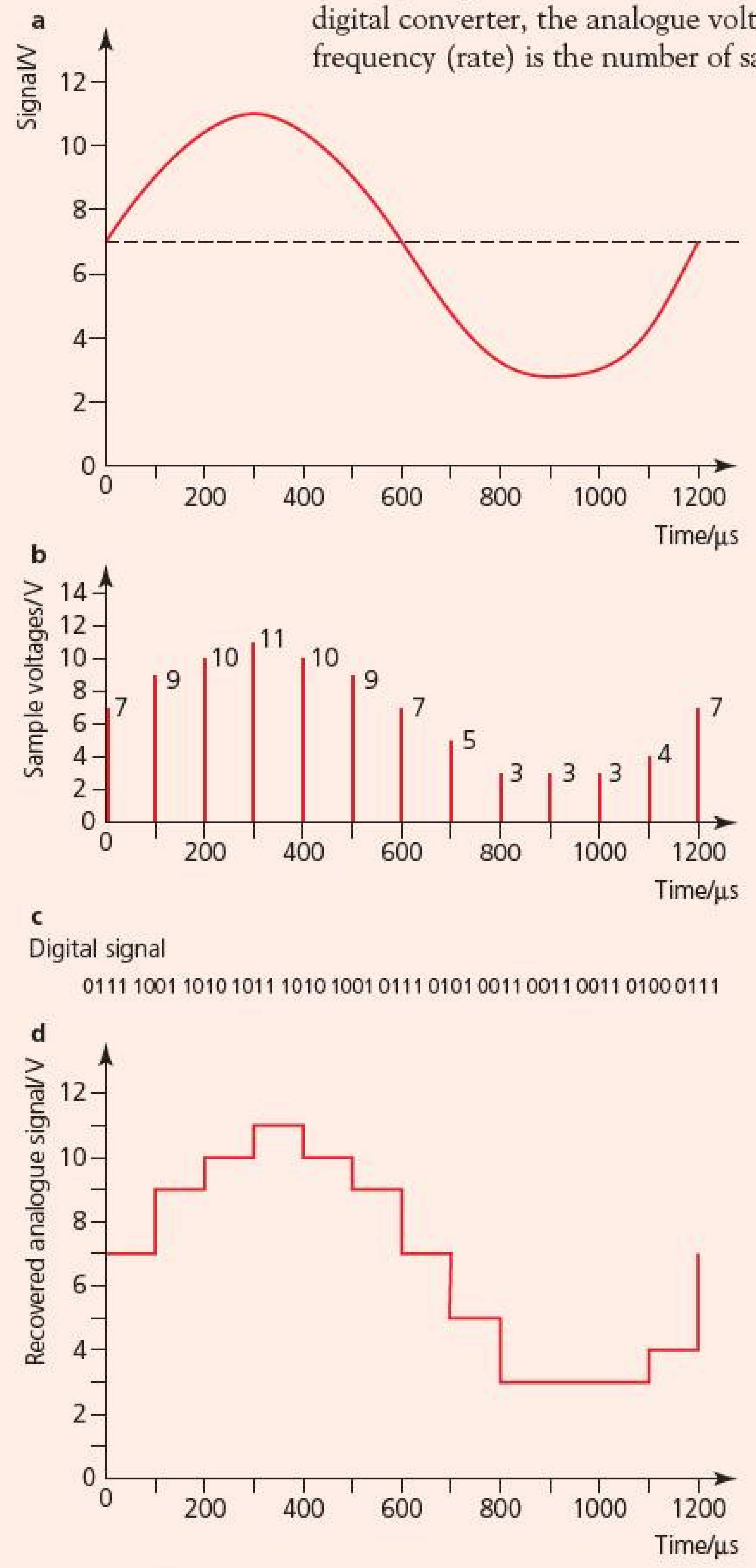


Figure 14.21 Analogue-to-digital and digital-to-analogue conversion

Questions

1 a Figure 14.22 shows a graph of the varying potential difference from a microphone. Convert this signal to a PAM signal by measuring the potential difference at a sampling rate of 200 Hz. Round off all of the potential difference values to the nearest tenth of a volt. Draw a graph showing the digital signal.

  b Repeat with a sampling rate of 400 Hz.

2 To measure the variation of a 100 Hz alternating current (ac) signal using a digital device, what is the minimum sampling rate you should use?

3 The following string of binary 1s and 0s is three-bit binary data sampled at 2 Hz:

  011 100 101 001 011 111

  a Present this data in the form of a table displaying time and a number to represent the potential difference.

  b Use this data to draw the digital signal.

4 Find out how digital data can be compressed and encrypted.

5 Find out about digital television (DTV). Outline the advantages of DTV over traditional analogue TV.

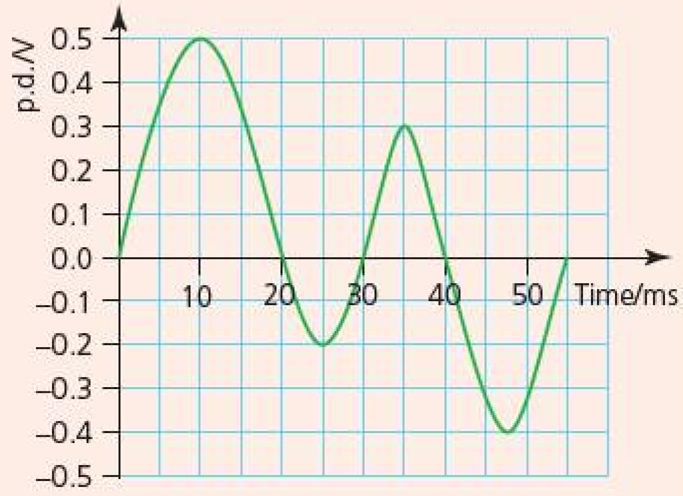


Figure 14.22 A graph of the varying potential difference from a microphone

Optical data storage capacity

14.1.5 Solve problems on CDs and DVDs related to data storage capacity.

Problems about optical data storage capacity can be about how much data is stored, how quickly it can be retrieved, the play time of a disc, or how long the tracks are that store the data. You will need to use what you have learned about the relationship between bits and bytes. You also need to know that the sampling rate is the rate at which an analogue signal is sampled (its value recorded). Thus 16-bit sampling at a rate of 44 kHz means that each second $44\,000 \times 16$ bits of data are recorded.

Worked examples

7 Information is imprinted (during the manufacturing process) on a music CD at a rate of 44 100 words per second. The digitized information consists of 32-bit words (two stereo channels of 16-bit samples each).

A mini-CD single contains 24 minutes of music. Calculate the capacity of a mini-CD single to the nearest megabyte.

Number of bits imprinted on the CD $= 44\,100 \times 32 \times 24 \times 60 = 2.03 \times 10^9$ bits

Since one byte equals eight bits this corresponds to:

$$\frac{2.03 \times 10^9}{8} = 254 \times 10^6 \text{ bytes} \approx 242\,\text{MB (since 1 MB} = 1\,048\,576 \text{ bytes)}$$

8 A track on a music CD moved from a radius of 21 mm to 54 mm with an average radius of 40 mm. The distance between spirals is 1.6 μm.

a Estimate the length of the track (in centimetres).

b The average scanning speed by the laser is $1.4\,\text{m s}^{-1}$. Estimate the play length of the audio CD in minutes.

c The CD can store 737 MB of information. What is the average length of track in micrometres per bit of information?

a Number of turns $= \dfrac{(54-21)\times 10^{-3}}{1.6\times 10^{-6}} = 2.06\times 10^4$

Track length $= 2.06\times 10^4 \times 2\times \pi \times 4.0 = 5.2\times 10^5\,\text{cm}$

b CD playing time $= \dfrac{5.2\times 10^3}{1.4} = 3.7\times 10^3\,\text{s} = 62$ minutes

c Average length per bit $= \dfrac{5.2\times 10^3}{7.37\times 10^8\times 8} = 8.8\times 10^{-7}\,\text{m} = 88\,\mu\text{m}$

9 Estimate the playing time of a 737 MB CD storing stereo music using 16-bit sampling. The maximum human audio frequency is 20 kHz, hence the sampling rate is 40 kHz.

Number of bits every second for each channel (of the stereo) $= 40\,000\times 16 = 6.4\times 10^5$ bits
Total number of bits per second for stereo $= 2\times 6.4\times 10^5$ bits $= 1.28\times 10^6$ bits
Total storage capacity of CD $= 7.37\times 10^8$ bytes $= 5.90\times 10^9$ bits

Maximum play time for CD $= \dfrac{5.90\times 10^9}{1.28\times 10^6} = 4610\,\text{s} = 77$ minutes

20 The track on a 12 cm audio CD is 5.7 km long and is made of a series of pits and lands that individually are a minimum of 0.8 μm long.
- **a** How many pits are there on a CD track (remembering that each pit is followed by a land)?
- **b** Each short pit has two edges so represents two bits of data. Calculate the number bits present on a standard audio CD.
- **c** How many megabytes (MB) of data are there on a standard 12 cm audio CD?

21 A standard 12 cm audio CD can store 74 minutes of stereo music.
- **a** If 16 bits are recorded at 44 100 samples per second, how many bits are recorded in 74 minutes?
- **b** The music to be stored is recorded as 'surround sound' so there are six channels. Calculate the number of bits.
- **c** 'Surround sound' music is supplied on a Super Audio CD (SACD), which has a much greater storage capacity than a standard audio CD. Calculate the number of megabytes stored in the 'surround sound' on the SACD.

Storage of information

14.1.6 **Discuss** the advantage of the storage of information in digital rather than analogue form.

■ Additional Perspectives

Transmission of information

When any electrical signal is transmitted over a long distance it will pick up noise. **Noise** is any unwanted random signal disturbance that adds to the signal being transmitted. It often occurs when two wires are next to each other. In addition, the power of the signal becomes progressively reduced with distance. We say that the signal has become **attenuated**. For long distance transmission, the signal has to be amplified at regular intervals. But the problem is that when an analogue signal is amplified, the noise is also amplified. The signal becomes distorted or 'noisy'. This effect is shown in Figure 14.23.

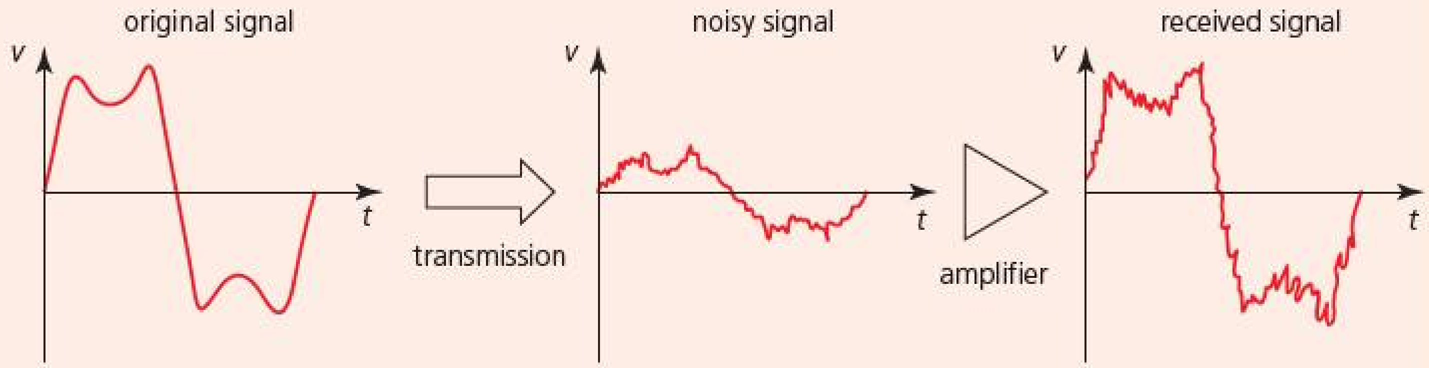


Figure 14.23 Amplification of a 'noisy' analogue signal

A digital signal still has noise and will be attenuated. However, when amplified the noisy 1s and 0s can be reshaped or regenerated to return the signal to its original form. Such amplifiers are known as regenerator amplifiers. They 'filter out' any noise and restore the digital signal (Figure 14.24). In contrast to an analogue signal, a digital signal can be transmitted over a long distance with regular regenerations without the signal becoming degraded.

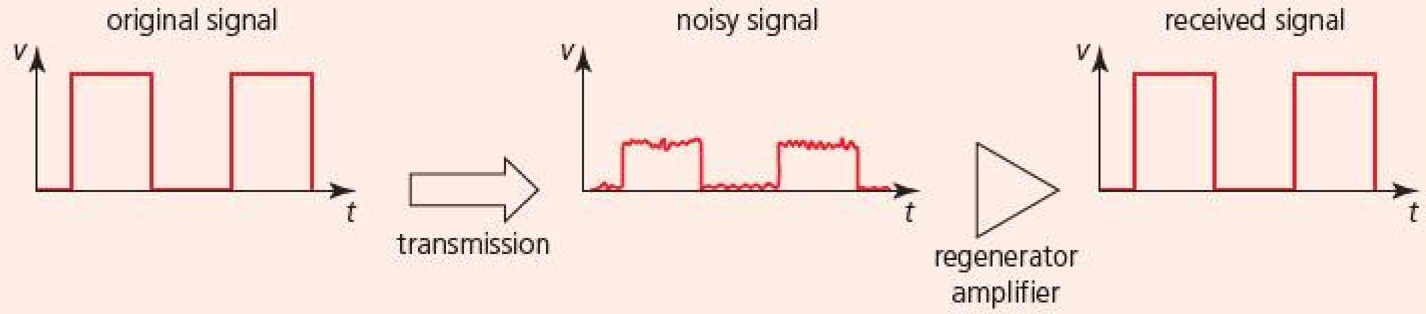


Figure 14.24 Amplification of a 'noisy' digital signal

Modern digital electronic circuits are more reliable, smaller and cheaper to produce than analogue circuits. An added advantage of digital systems is that extra information, or data known as a checksum, can be added to the transmission. These extra data are a simple code for the receiving system so that the transmitted signal may be checked and corrected before the signal is finally reproduced.

Digital data can be quickly and cheaply manipulated, compressed (to reduce storage space), processed, edited or made secure via a process of encryption. Analogue data is more difficult to process and encrypt.

Modern miniaturization techniques based on microprocessors mean that large amounts of digital data can be stored on a physically small device. Although stored analogue data can be compact, for example analogue tapes, many analogue storage systems are relatively large. Compare, for example, an MP3-player with an LP. A 120GB MP3 player can store the music of about 1500 LPs.

The retrieval speed for digital data is usually much more rapid than the retrieval speed for comparable analogue data. The process of retrieving analogue data often affects the quality of future data retrieval. For example, the quality of music on magnetic tape and LPs decreases with the number of times the data is retrieved and the music played. However, optical techniques, as used in CD-ROM readers, can ensure that the data is identical each time it is accessed and there is no degradation in its quality.

Data storage

14.1.7 Discuss the implications for society of ever-increasing capability of data storage.

Table 14.5 on page 489 outlines *some* of the implications for the ever-increasing capability of data storage. It is claimed computing power (the number of operations a computer can perform per second) doubles about every 18 months and data storage costs are rapidly declining. Networking advances and the Internet make copying data from one location to another and accessing personal data from remote locations much easier. Some of the moral, ethical, social and environmental implications of this global issue are considered.

Table 14.5 Implications for increasing data storage

Moral/ethical	• Issues concerning the privacy and anonymity of personal digital data. • Issues related to access and ownership of personal digital data.
Cultural	• Some people are concerned about cultural integrity of their regional or lingual groups due to the widespread use of English on the Internet.
Economic	• Many of the recent financial scandals involved the secret modification of accounting and financial data. • The integrity of currencies may be undermined by digital or e-cash. • Increased access to training and education via computer-assisted teaching and learning. • Economic decision-making by consumers will be faster and more accurate. This may lead to more stable markets for goods and services and perhaps a tendency towards low inflation. There is also a counter argument, that the 'herd instinct' seen in stock markets, based on instant digital information, results in a high degree of instability in the system.
Social	• The emergence of new forms of communities due to social networking sites (such as Facebook and Twitter), which have both positive and negative features.
Environmental	• Electronic data storage could replace traditional techniques, thus saving in resources, e.g. e-books do not use wood pulp. However, data centres consume electricity (and hence in most cases contribute to global warming and climate change). • The resources needed for the maintenance of electronic data will be in higher demand, e.g. fossil fuels and silicon.

22 Find about the new emerging science of bioinformatics, a combination of biology and information technology.

23 Find about any laws in your country, such as the United Kingdom's Data Protection Act 1998, which govern the protection of personal data, including digital data.

24 Research the use of 'cookies' and 'web bugs' by websites on the Internet.

25 Give one additional example of your own for each of the five headings in Table 14.5.

14.2 Data capture; digital imaging using charge-coupled devices (CCDs)

Capacitance

14.2.1 Define *capacitance.*

Capacitors (Figure 14.25) are devices that can store charge (electrons). They store energy in an electric field. Capacitors are common components of electrical circuits and perform a variety of functions. Camera flash units use capacitors to store energy and capacitors are an essential part of radio tuners. Digital cameras contain a sensor known as a charge-coupled device (CCD) which is made up of millions of tiny pixels, which behave like tiny capacitors.

Two parallel metal plates separated by a small gap filled with air form a simple capacitor, as illustrated in Figure 14.26. The metal plates are connected to a battery (a source of a potential difference). When the switch is closed, a current will flow for a very short time. The capacitor becomes charged. Negative charge (electrons) will accumulate on the bottom plate, leaving behind an equal amount of positive charge on the top plate.

The amount of charge stored on the surface of the plates depends on the design of the capacitor and the p.d. across them. In fact, the charge, q, is directly proportional to the

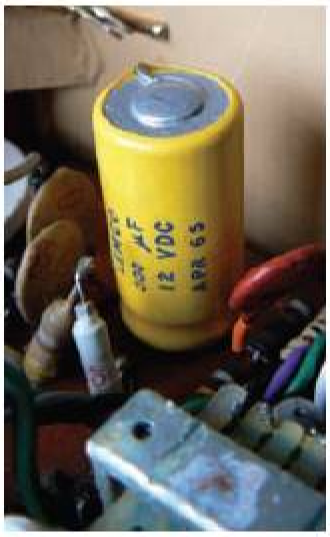


Figure 14.25 A capacitor (with a capacitance of 500 μF)

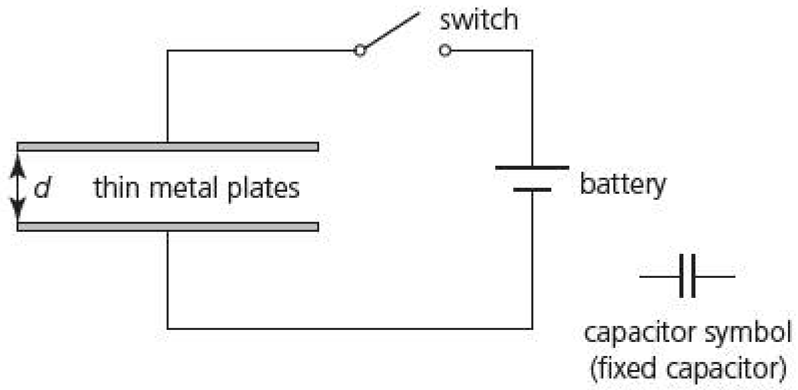


Figure 14.26 A simple circuit with a capacitor, and the symbol for a capacitor (with fixed capacitance)

potential difference between the metal plates (i.e. $q \propto V$). The constant in this relationship is called the **capacitance** C of the plates.

$$q = CV$$

Capacitance is the charge per unit potential difference that can accumulate on a conductor. It is defined as the ratio of charge stored to the applied potential difference:

$$C = \frac{q}{V}$$

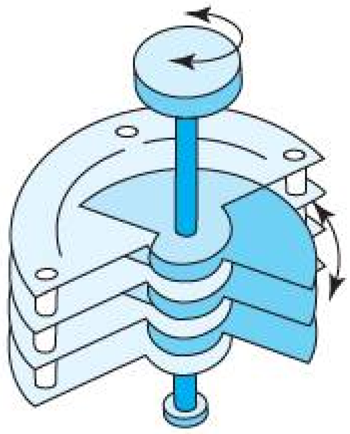


Figure 14.27 A variable capacitor

The SI unit of capacitance is the farad (F), with one **farad** (1 F) being a capacitance of one coulomb per volt ($1\,CV^{-1}$); one farad (1 F) is a relatively large capacitance. The unit is named after Michael Faraday.

Smaller multiple units of the farad are often used: $1\,\mu F = 10^{-6}\,F$ (a microfarad), $1\,nF = 10^{-9}\,F$ (a nanofarad) and $1\,pF = 10^{-12}\,F$ (a picofarad). Variable capacitors (Figure 14.27) can be constructed whose capacitance can be varied by changing the overlap of the plates. A variable capacitor is a critical part of a radio tuning system that selects the frequency of a radio station, thus it is often referred to as a tuning capacitor. Resonance occurs when the discharge frequency of the capacitor equals the radio frequency.

Worked example

10 The capacitance of two parallel metal plates is 9.50 pF. Calculate the charge on one of the plates when a potential difference of 18.0 V is established between the plates.

$$q = CV = 9.50 \times 10^{-12} \times 18.0 = 1.71 \times 10^{-10}\,C = 171\,pC$$

Additional Perspectives

Capacitors

The voltage rises as we charge up a capacitor, and falls as the capacitor discharges. The current falls from a high value as the capacitor charges up, and falls as it discharges. During the discharge of a capacitor the current can cause the heating of a resistor or even drive a small electric motor for a short period of time. So the charged capacitor is acting as an energy store. A graph of potential difference (V) across the capacitor against charge (q) is a straight line, with the gradient being a constant for the capacitor, and equal to 1/C. For a capacitor the energy stored is equal to the area under the graph of potential difference versus charge (Figure 14.28).

Thus:

$$\text{energy stored} = \text{area under } V\text{–}q \text{ graph}$$
$$= \frac{1}{2} \times \text{final voltage} \times \text{final charge}$$
$$= \frac{1}{2}Vq$$

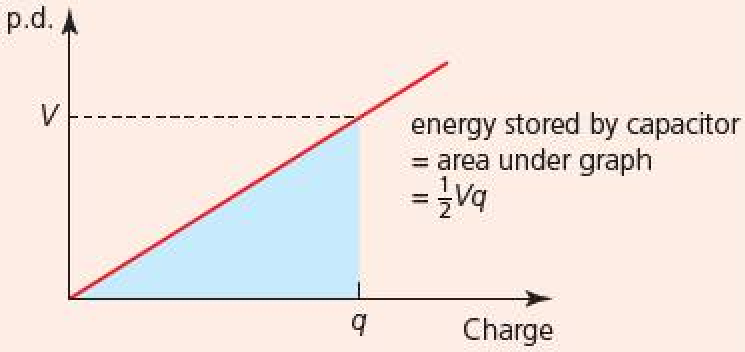


Figure 14.28 Energy stored in a capacitor

Since $C = \frac{q}{V}$,

$$\text{energy stored} = \frac{1}{2}\frac{q^2}{C} \quad \text{or} \quad \frac{1}{2}CV^2$$

When a capacitor is discharged the charge decays exponentially (Figure 14.29). A graph of charge q against time t is an exponential decay curve (compare with radioactive decay – Chapters 7 and 13). The rate of decay, or steepness of the curve, depends on the values of R (resistance) and C (capacitance) and on the charge on the capacitor at that moment. Such a proportional relationship between the rate of change of value and the value itself always leads to an exponential relationship.

The exponential decay curve for a capacitor is described by the following equation:

$q = q_0 e^{-t/RC}$

where RC is known as the *time constant*.

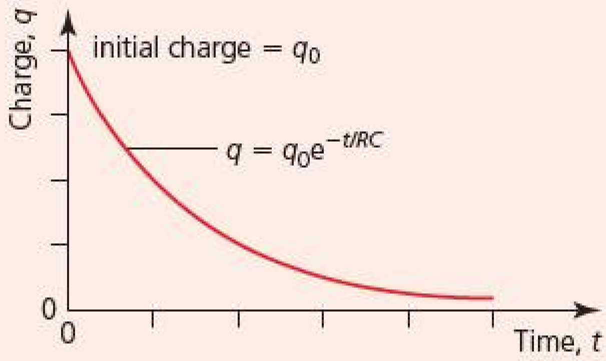


Figure 14.29 A decay curve for a discharging capacitor

Question 1

1 Use a spreadsheet to model capacitor discharge: use R = 47 kΩ and C = 22 μF. In Excel the EXP function returns e raised to the nth power, where e = 2.718. The syntax for the EXP function is: EXP (number). Use the generated numbers and plot a graph of charge versus time. Plot a graph of ln$q$ against time and show that the slope of the graph will be $-1/RC$.

26 What will be the potential difference between the plates of a 20 μF capacitor if there is a charge of 5 μC on the plates?

27 The capacitance of a device known as a charge-coupled device (CCD) is 100 μF. What charge (in μC) will cause a potential difference of 1.2 V across it?

28 If there are 500 electrons stored on a 100 nF photodiode, what will be the potential difference across it?

29 Find out about the use of capacitors in the tuning of a simple radio with a tuning circuit.

30 A capacitor is charged with a 20 V battery. Calculate the capacitance of the capacitor if the maximum charge on the plates is ± 500 μC.

31 A graph of the variation of charge with potential difference (p.d. or voltage) across a capacitor is shown in Figure 14.30. What is the capacitance of the capacitor?

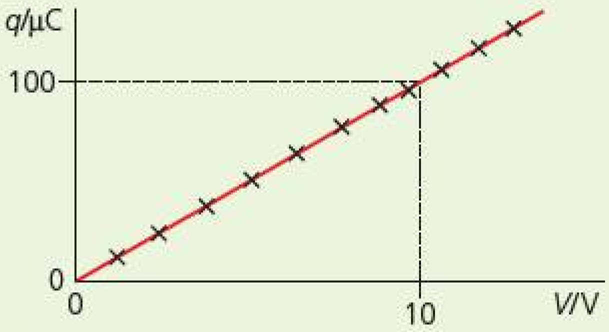


Figure 14.30

Charge-coupled devices

A charge-coupled device (CCD) is a highly sensitive light detector originally developed in astronomy for digital imaging. These devices doubled the ability of telescopes to detect light and allowed astronomers to obtain colour images of very faint stars. They are now to be found in digital cameras, web cams, digital video recorders and scanners.

Image capture in a charge-coupled device

14.2.2 **Describe** the structure of a charge-coupled device (CCD).
14.2.3 **Explain** how incident light causes charge to build up within a pixel.

A charge-coupled device (CCD) is a small silicon chip with a surface covered with a very large number of light sensitive elements called **pixels** (picture elements). The pixels are defined on the surface of the silicon wafer by insulating channels in one direction and by rows of transparent electrodes deposited on the surface in the other direction (Figure 14.31).

When exposed to light, for example by opening the shutter on a digital camera, each pixel within the CCD releases electrons as result of a process similar to the photoelectric effect (Chapter 13). The greater the intensity of the light on a pixel, the higher the number of photons that are incident on the pixel per second, and so the higher the number of electrons released.

The voltages that develop on the electrodes across the array are set so that each individual pixel in the array behaves as a small capacitor, trapping the electrons released in a potential 'well' (Figure 14.32). This results in a build up of charge within the pixel, and a potential difference V develops across the pixel given by $V = q/C$, where q is the charge of the electrons and C is the capacitance of the pixel. Because this potential difference is proportional to the number of electrons released, it is also proportional to the intensity of light falling on that pixel.

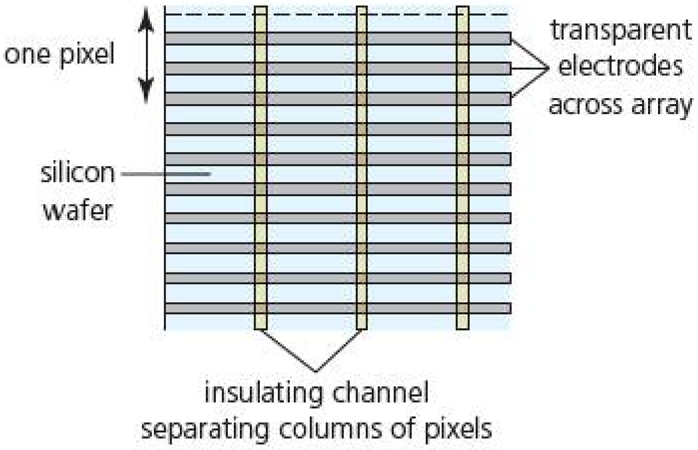


Figure 14.31 Part of the pixel array on the surface of a CCD sensor

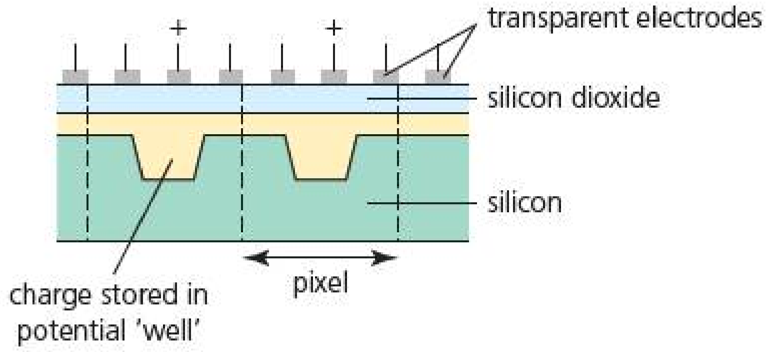


Figure 14.32 Storage of charge in a pixel array. Electrons released by the incident photons are stored in the potential 'well' beneath the positively charged electrode

Once the exposure to light has stopped, for example by closing the shutter on a camera, the charge that has built up in each pixel is stored.

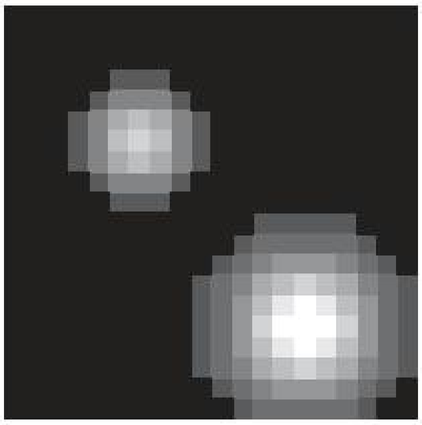
32 Find out how charge passes through a semiconductor, like silicon. (Investigate both the flow of electrons and of positively charged 'holes'.)

Processing a CCD image

14.2.4 **Outline** how the image on a CCD is digitized.
14.2.10 **Outline** how the image stored in a CCD is retrieved.

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	3	3	3	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	3	5	6	5	3	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	3	6	7	6	3	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	3	5	6	5	3	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	3	3	3	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	1	0	0
0	0	0	0	0	0	0	0	0	0	1	2	3	4	4	4	3	2	1	0
0	0	0	0	0	0	0	0	0	1	2	3	5	6	7	6	5	3	2	1
0	0	0	0	0	0	0	0	0	1	2	4	6	8	9	8	6	4	2	1
0	0	0	0	0	0	0	0	0	1	2	4	7	9	9	9	7	4	2	1
0	0	0	0	0	0	0	0	0	1	2	4	6	8	9	8	6	4	2	1
0	0	0	0	0	0	0	0	0	1	2	3	5	6	7	6	5	3	2	1
0	0	0	0	0	0	0	0	0	0	1	2	3	4	4	4	3	2	1	0
0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	1	0	0

a



b

Figure 14.33 **a** Data from a CCD (as decimal numbers) **b** Interpretation of numbers as light intensity levels

Image formation

The stored charges across the pixel array represent the variation in intensity of the light that fell upon the detector surface. It is analogue information. By measuring and digitizing the voltage developed across each pixel, at the same time recording the position of each pixel, a digital intensity 'map' of the CCD surface is built up. This intensity 'map' is in effect an image of the light focused onto the CCD (Figure 14.33).

Figure 14.33a shows the typical data from a CCD consisting of an array of numbers, running from 0 to 9 in this simplified example (rather than using binary numbers). Each number represents the intensity of the light incident on that specific pixel. The larger the number, the greater the intensity of the incident light. These numbers correspond to the intensity levels shown in Figure 14.33b.

Retrieving the information from a CCD

As shown in Figure 14.31, the pixels are arranged in columns that are separated from one another by an insulator. By applying a potential difference to corresponding electrodes on each row of pixels, the charges in each row are pushed down to the row below. The charge on the bottom row moves off the array onto the **serial register** (Figure 14.34).

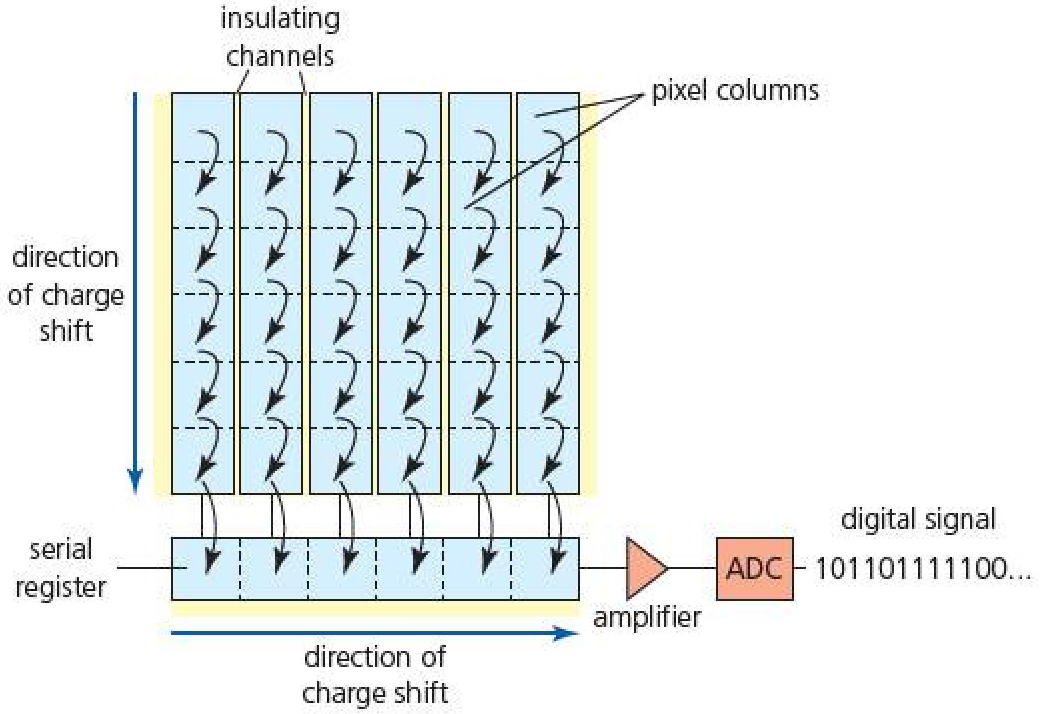


Figure 14.34 A clocking potential shifts the charges down the array

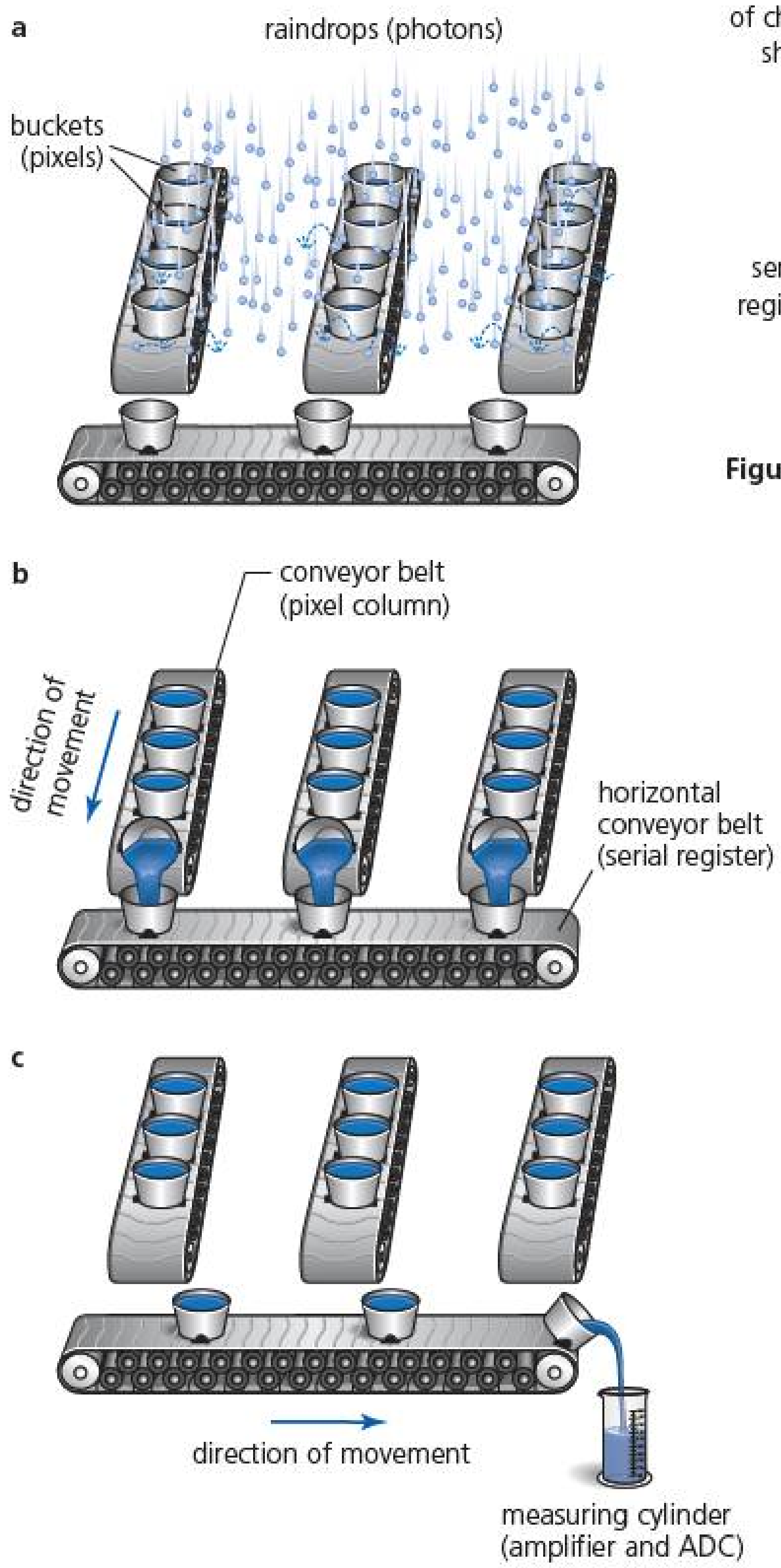


Figure 14.35 A simple mechanical analogy for the clocking of a CCD

The packets of charge on the serial register are then moved sideways. One by one they pass through an amplifier and then to an analogue-to-digital converter (ADC), which outputs the value in digital form. This process continues until the charge in the entire row of pixels has been read and converted to a digital signal. This process is known as 'clocking' the CCD.

The process is repeated for the next row, until the entire array has been processed. The name 'charge-coupled' device comes from this process, where the stored charge passes down from one row to the next to be read. Once the data from whole array has been read, digitized and stored, which can happen very quickly, the pixels can be reused to record another image.

Figure 14.35 shows a simple mechanical analogy (using rain and buckets) describing how a CCD is clocked and its charge collected. Once the exposure to rain is finished the buckets will contain samples of rain water (Figure 14.35a). The conveyor belt starts turning and transfers the water in the first buckets (only) to other buckets on a stationary 'horizontal' belt at the end (Figure 14.35b). The conveyer belts stop and the horizontal conveyor starts up and tips each bucket in turn into the measuring cylinder (Figure 14.35c). After each bucket has been measured, the measuring cylinder is emptied, ready for the next bucket load.

To produce a black and white image, the only information required is the overall light intensity on each pixel and the pixel's position (its x- and y-coordinates). Computer software can use the digital data retrieved from the CCD to generate a black and white image of the scene recorded. This can be displayed on a computer screen and be further processed or manipulated.

A colour image requires more information. One type of colour CCD has the pixels arranged into groups of four, with green, red and blue filters, known as the Bayer array filter (Figure 14.36). There are two green filters because the human eye is more sensitive to green light than to red or blue light. The intensity of the light in each of the four pixels is measured. The group of four pixels acts as a sensor unit giving information for all three primary colours for that area. Computer software combines the information to produce digital data that can recreate a full colour image.

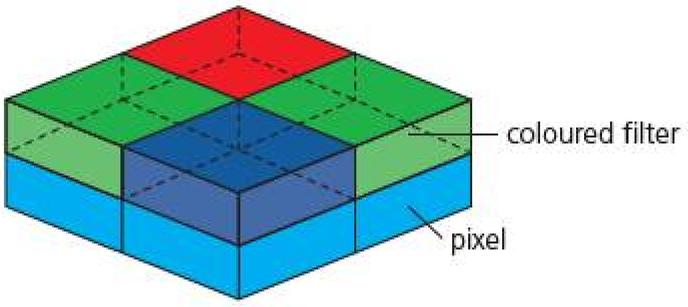


Figure 14.36 A group of four pixels with green, red and blue filters (Bayer array filter)

Characteristics of a CCD image

Quantum efficiency

14.2.5 Define *quantum efficiency* of a pixel.

Not every photon that is incident on a pixel of a CCD will cause an electron to be released. Some photons are reflected and other photons may pass through the pixel.

The **quantum efficiency** of a pixel is definded as the ratio of the number of emitted electrons to the number of incident photons. It is an accurate measurement of a CCD's electrical sensitivity to light.

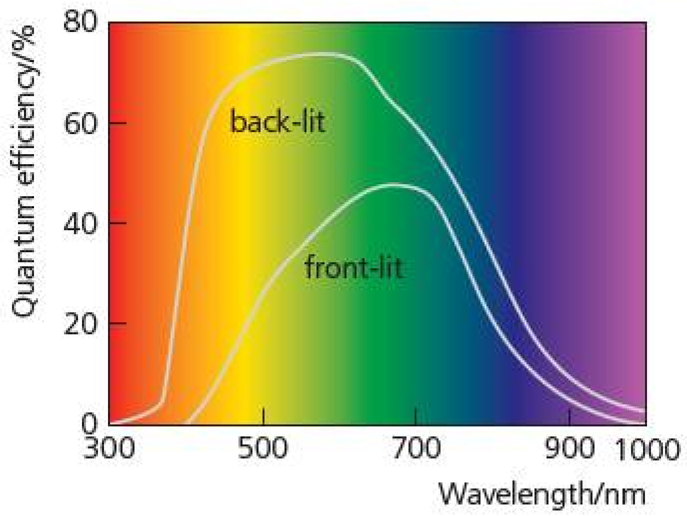


Figure 14.37 Comparison of the quantum efficiency of front-lit and back-lit CCDs over the range 300–1000 nm

A higher quantum efficiency means that a clear image is formed even at low light intensity. This is very important in astronomical imaging where very faint light signals are being studied. Since the energy of a photon varies with wavelength, quantum efficiency is often measured over a range of different wavelengths to quantify a CCD's efficiency at each photon energy.

Charge-coupled devices have very high values of quantum efficiency ranging between 70% and 80%. The human eye has a quantum efficiency of only about 20%; photographic film (an analogue medium) has a quantum efficiency of around 10%. However, the quantum efficiency is not constant for all wavelengths of light and is generally higher for back-lit CCDs than for those that are illuminated from the front (Figure 14.37). The use of back-lit CCDs in digital cameras allows the user to take high-quality night shots.

33 One-fifth of the photons incident on a CCD do not result in electrons being emitted. Calculate the quantum efficiency of the CCD.

34 There are 50 photons incident on a CCD pixel (photodiode) and 30 electrons are released. Calculate the quantum efficiency.

Magnification

14.2.6 Define *magnification*.

The magnification of the image on a CCD is defined as the ratio of the length of the image as it formed on the CCD to the actual length of the object. Magnification has no units.

$$\text{magnification} = \frac{\text{length of the image on CCD}}{\text{actual length of the object}}$$

The magnification of a CCD system is determined by the overall properties of the lenses that are used to focus the light from the CCD. A greater magnification means that more pixels are used for a given section of the image. Hence, the image will be more detailed.

Worked example

11 A digital camera is used to take a photograph of a small insect embryo. The area of the embryo is $0.012\,cm^2$ and the area of the image is $9.8 \times 10^{-6}\,m^2$. Calculate the magnification of the CCD.

The ratio of the image to the object areas is:

$$\frac{9.8 \times 10^{-6}}{1.2 \times 10^{-6}} = 8.2$$

So the ratio of corresponding linear sizes is $\sqrt{8.2} = 2.9$
Therefore the magnification is 2.9.

Resolution

14.2.7 State that two points on an object may be just resolved on a CCD if the images of the points are at least two pixels apart.

An important characteristic of a CCD is its ability to *resolve* two points very close to each other on an object whose image is required (resolution was discussed in Chapter 11).

Two points are resolved (distinct) if their images are at least two pixels apart (Figure 14.38), so that there is a noticeable decrease in intensity between them.

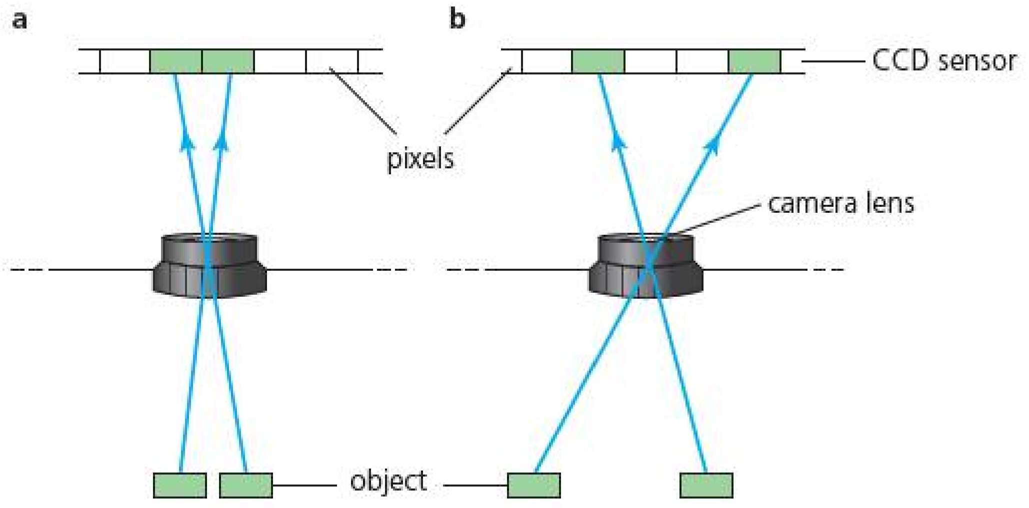


Figure 14.38 A pair of objects which are **a** not resolved and **b** resolved by a CCD

Worked example

12 The magnification produced by a 8.0 megapixel digital camera with a light-collecting area of $14\,mm^2$ is 1.8. Determine if this digital camera can resolve two points that are $1.2 \times 10^{-3}\,mm$ apart.

The area of a pixel $= \dfrac{14}{8.0 \times 10^{-6}} = 1.8 \times 10^{-6}\,mm^2$
and so the length of a pixel is $\sqrt{1.8 \times 10^{-6}} = 0.0013\,mm = 1.3 \times 10^{-3}\,mm$
The distance between the two points is:
$1.8 \times 1.2 \times 10^{-3}\,mm = 2.2 \times 10^{-3}\,mm$
This is less than two pixel lengths and so the points are not resolved.

Factors affecting the quality of a CCD processed image

14.2.8 Discuss the effects of quantum efficiency, magnification and resolution on the quality of the processed image.

A higher quantum efficiency means that the image produced by a CCD will require less time to form if the incident light intensity is very low. This is important in astronomy because the Earth is rotating relative to the stars in the night sky.

The greater the magnification of an object, the greater the length of the image on the CCD surface. This results in a larger number of pixels that will accumulate charge due to the incident light. This means that the image will be more detailed and have a higher resolution.

The resolution is greatest with a high pixel density (that is, number of pixels per unit area). An image of high resolution is of high quality because it includes more detail than an image of low resolution.

Practical uses of CCDs

14.2.9 Describe a range of practical uses of a CCD, and list some advantages compared with the use of film.

Charge-coupled devices have a range of practical uses, as described below.

Endoscopes

An endoscope is a thin and flexible tube with a light and lenses attached at the end that is used to look inside the body. The endoscope can make use of the body's natural openings, so it is often inserted through the mouth, nose, ear or anus. Modern electronic endoscopes are equipped with CCDs that produce high-quality colour images in real time.

Medical X-ray imaging

Special CCDs have been developed which can detect X-rays. This means that exposure times to X-rays are shorter, benefiting patients and operators. At present these devices are relatively expensive and many X-ray machines still use photographic film.

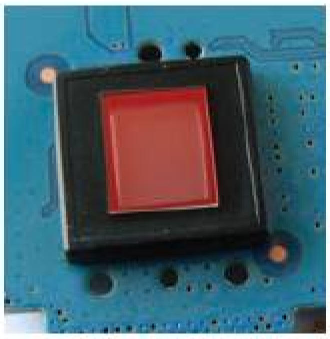


Figure 14.39 The CCD of a disposable polaroid camera

Digital cameras and video cameras

Digital cameras (Figure 14.39) are widely used and have many advantages over photographic film. The image produced by a CCD can be enhanced and edited using electronic processing techniques. The storage, archiving and sorting of a large number of photographs from a digital camera is relatively easy and cheap. The images taken by a digital camera can be viewed immediately – the processing time is very fast. A high definition digital video camera generally takes about 30 pictures per second. This means that the information from the CCD must be read very quickly, which limits the number of pixels.

Telescopes

Figure 14.40 An HST image of the Butterfly Nebula, 3800 light years away from Earth, taken using the Wide Field Camera 3

Charge-coupled devices are very useful in data collection in astronomy because they can respond to a wide range of electromagnetic radiation, and their response is in electrical form. They are also very sensitive to low intensities of light.

Electromagnetic radiation from distant sources is more intense in space than it is at the bottom of the Earth's atmosphere. The Hubble Space Telescope was launched into orbit in 1990 and does not suffer from the effects of atmospheric refraction and scattering that affect telescopes on Earth. The two cameras currently in use have specialized CCD arrays for recording different parts of the electromagnetic spectrum. Figure 14.40 shows an HST image of hot gases ejected from a dying star, a so-called planetary nebula.

35 Find about Super CCD developed by Fujifilm.

Solving problems involving CCDs

14.2.11 Solve problems involving the use of CCDs.

A variety of problems can be solved relating to CCDs. The calculations may involve energy, power, surface area and the energy carried by a photon, or the charge carried by a specific number of electrons.

Worked examples

13 Calculate the number of megapixels on a $20\,\text{mm} \times 20\,\text{mm}$ CCD where the pixel size is $25 \times 10^{-6}\,\text{m}$.

The collecting area of the CCD is
$20\,\text{mm} \times 20\,\text{mm} = 4.0 \times 10^2\,\text{mm}^2 = 4.0 \times 10^2 \times 10^{-6}\,\text{m}^2 = 4.0 \times 10^{-4}\,\text{m}^2$

and the area of each pixel is
$25 \times 10^{-6}\,\text{m} \times 25 \times 10^{-6}\,\text{m} = 6.25 \times 10^{-10}\,\text{m}^2$

The number of pixels is therefore

$$\frac{4.0 \times 10^{-4}}{6.25 \times 10^{-10}} = 6.4 \times 10^5 = 0.64 \text{ megapixels}$$

14 Light of intensity $6.9 \times 10^{-6}\,\text{W}\,\text{m}^{-2}$ and wavelength $6.0 \times 10^{-7}\,\text{m}$ is incident on a CCD with a collecting area of $6.25 \times 10^{-10}\,\text{m}^2$. Calculate the number of photons incident on each pixel in a period of 35 ms.

Power incident on CCD area $= 6.9 \times 10^{-6}\,\text{W}\,\text{m}^{-2} \times 6.25 \times 10^{-10}\,\text{m}^2 = 4.3 \times 10^{-15}\,\text{W}$
The total energy incident on the CCD in a time of $35\,\text{ms} = 4.3 \times 10^{-15}\,\text{W} \times 35 \times 10^{-3}\,\text{s}$
$= 1.5 \times 10^{-16}\,\text{J}$
The energy of one photon is

$$E = \frac{hc}{\lambda} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{6.0 \times 10^{-7}} = 3.3 \times 10^{-19}\,\text{J}$$

and hence the number of photons per pixel is equal to the total energy incident on the CCD/energy of one photon:

$$\frac{1.5 \times 10^{-16}}{3.3 \times 10^{-19}} \approx 460$$

15 The area of a pixel in a CCD is $8.5 \times 10^{-10}\,\text{m}^2$ and its capacitance is 54 pF. Light of intensity $3.2 \times 10^{-3}\,\text{W}\,\text{m}^{-2}$ and wavelength $4.6 \times 10^{-7}\,\text{m}^2$ is incident on the collecting area of the CCD for 150 ms. Calculate the potential difference established at the ends of the pixel, assuming that 80% of the incident photons cause the emission of electrons.

The energy incident on a pixel is
$3.2 \times 10^{-3} \times 8.5 \times 10^{-10} \times 150 \times 10^{-3} = 4.08 \times 10^{-13}\,\text{J}$

The energy of one photon is

$$E = \frac{hc}{\lambda} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{4.6 \times 10^{-7}} = 4.32 \times 10^{-19}\,\text{J}$$

The number of incident photons is then equal to

$$\frac{4.08 \times 10^{-13}}{4.32 \times 10^{-19}} = 9.44 \times 10^5$$

Since the quantum efficiency of the CCD is 80%, the number of ejected electrons is
$0.80 \times 9.44 \times 10^5 = 7.55 \times 10^5$

The charge corresponding to this number of electrons is
$7.55 \times 10^5 \times 1.60 \times 10^{-19} = 1.21 \times 10^{-13}\,\text{C}$

The potential difference is then

$$V = \frac{q}{C} = \frac{1.21 \times 10^{-13}}{54 \times 10^{-12}} = 2.2 \times 10^{-3}\,\text{V} = 2.2\,\text{mV}$$

36 **a** A square CCD has a diagonal measurement of 6.0 mm. Calculate the width of the CCD.
b If each pixel is a square of side 12 μm, how many pixels will this CCD have?
c A picture of a 5.00 m high tree is taken with a digital camera containing a 6.0 mm CCD. What is the magnification of the camera if the tree image just fits on the camera's CCD?

37 **a** If 10^{11} photons enter a digital camera and are incident on the CCD of a 6 megapixel camera, how many photons land on each pixel?
b If the quantum efficiency is 75%, how many electrons are liberated in each pixel of a back-lit CCD by the photoelectric effect?

38 The rate of reading each pixel in a CCD is 8 MHz. How long will it take to read (clock) all of the pixels in a 1 megapixel CCD?

39 A video camera takes 25 frames per second (smooth action video must be shot at a minimum of 25 frames per second). If the rate of reading pixels is 6 MHz, how many pixels does the digital video camera have?

40 Find about complementary metal oxide semiconductors (CMOS) and their use as image sensors. Summarize their uses and relative advantages and disadvantages compared to CCDs.

41 Outline current research into 'vision chips' using CCDs and CMOS. Could such devices be used to help blind people (with defective retinas) see in the future? How are they currently being used in robotics?

42 CCDs used in astronomy are often cooled down to low temperatures. Find out why this is done. Refer to dark currents.

SUMMARY OF KNOWLEDGE

14.1 Analogue and digital signals

- Information can be stored and transmitted in digital and analogue forms.
- An analogue signal has a large number of different values (between given limits), and hence varies continuously with time.
- A digital signal repeatedly changes between two possible values of 0 (off) and 1 (on).
- Analogue storage devices include the LP and cassette tape.
- In an LP the sound variations are stored as physical variations in a track (groove) on the LP surface.
- In a cassette tape the data is stored as variations in the orientations of magnetic particles on the tape surface.
- Digital storage devices include floppy and hard disks, and optical devices such as CDs and DVDs.
- Floppy and hard disks store data in a series of magnetic variations on the disk surface.
- Digital techniques involve signals consisting of a large number of binary digits, or bits.
- Each binary digit or bit can only take one of two possible values: binary 1 (a specific voltage) or binary 0 (zero voltage).
- Decimal (base 10) numbers (0–9) are represented by counting in powers of 10, so each digit (from left to right) represents 1, 10, 100, 1000, etc.
- Binary (base 2) numbers (0 and 1) are represented by counting in powers of 2, so each digit (from left to right) represents 1, 2, 4, 8, 16, 32, etc.
- In binary notation, the largest power of a series of binary digits is known as the most-significant bit (MSB), and the smallest power is known as the least-significant bit (LSB).
- When the number of bits used is n, the number of different possible values is 2^n.
- One byte is eight bits. 1 KB (one kilobyte) is 2^{10}, or 1024 bytes; 1 MB (one megabyte) is 2^{20}, or 1 048 576 bytes. 1 GB (one gigabyte) is 2^{30} or 1 073 741 824 bytes.
- American Standard Code for Information Interchange (ASCII): eight-bit code representing 256 or 2^8 different possible characters and formatting codes.
- To convert analogue data to digital data the analogue data is sampled at intervals. Each sampled signal is converted into one binary value among a fixed range of possible values/quantum levels.
- To improve the accuracy of the digital data the sampling frequency and the number of available quantum levels can be increased.
- A CD contains a single very long spiral-shaped track that starts in the centre. It is composed of a large number of 'pits' and spaces known as 'lands'.
- The digital information in the CD is read by sensing the amplitude of the reflection of a laser beam (of visible light) reflecting off the 'lands' and 'pits'.
- The speed of rotation of the CD is controlled by an electric motor so that a constant length of the track is scanned by the laser beam in a given period of time.
- The CD has a higher speed of revolution when the laser is reading the track near the centre compared with the outer edge.

- When the laser beam reflects from a 'pit', a strong signal (binary 1) is received and detected.
- When the laser beam reflects from the edge between a 'pit' and a 'land' destructive interference occurs and a suppressed (weak) signal (binary 0) is received and detected.
- An appropriate depth of a 'pit' for the wavelength, λ, of laser light is $\frac{\lambda}{4}$.
- Number of turns on CD = change in radius/distance between spirals on track; length of track = number of turns $\times 2\pi \times$ average radius.
- CD play time = length of track/scanning velocity, and the average length of track per bit of information = length of track/number of bits.
- Pits and lands on a DVD are closer together than on a CD, hence it can store more data. Also, DVDs allows multilayered data storage.
- Characteristics of information stored in digital format: high quality (if sampling rate is high), exact reproducibility, large quantities can be stored in a small device and data can be readily manipulated (changed and copied) without corruption.
- Characteristics of information stored in analogue format: variable quality, reproducing the data introduces noise, analogue storage devices may be large and data may be corrupted if manipulated (changed).

14.2 Data capture

- Capacitors are electrical components that can store charge (and energy). The charge stored is proportional to the potential difference applied across the capacitor.
- The simplest capacitor consists of a pair of flat metal plates (separated by an insulating material).
- Capacitance is the charge stored per unit of potential difference, $C = q/V$
- The unit of capacitance is the farad, $1\,F = 1\,CV^{-1}$.
- The charge-coupled device (CCD) is a silicon-based device that is used to record an image focused on to its surface. The incident light releases electrons because of the photoelectric effect.
- The surface of the CCD is divided into a large number of small pixels (photodiodes) which behave as tiny capacitors.
- When light is incident on the CCD the photons cause electrons to be stored in each pixel (photodiode); the number of electrons stored depends on the intensity of light (number of photons incident per second).
- The image data is stored as charge on the two-dimensional grid of pixels on the surface of the CCD.
- A potential difference is applied across the CCD to move all the charge down one row. The end row forms a serial register, which is a row of pixels whose potential differences can be measured and recorded. The measured potential difference is proportional to the charge stored.
- An analogue-to-digital converter converts the potential differences into binary data for processing. After measurement, the charge is removed from the rows of pixels and the CCD array is ready to record another image.
- Quantum efficiency is defined as the ratio of the number of photoelectrons emitted to the number of photons incident on the pixel of the CCD.
- The magnification is defined as the ratio of the length of the image on the surface of the CCD to the actual length of the object.
- Two points on an object may be just resolved on a CCD if the images of the points are two pixels apart.
- If the quantum efficiency is high then a high quality image is generated; if the quantum efficiency is low then a poor quality image is generated and some less bright parts of the image will be lost.
- A large CCD means that the magnification can be greater, resulting in a higher resolution and better quality image.
- A higher resolution means a better quality image with a greater amount of detail.

- CCDs are used for image capturing in digital cameras (including mobile phone cameras), digital video cameras, telescopes (including the orbiting Hubble Telescope), scanners, laser printers and imaging of X-rays.
- Advantages of a CCD (in a digital camera) compared to photographic film: lower cost, much higher quantum efficiency, the image can be readily copied, deleted or processed; and storage, viewing and archiving of a large number of images is easy.
- CCDs have been developed that detect photons from the infrared, X-ray and ultraviolet regions of the electromagnetic spectrum.

Examination questions – a selection

Paper 1 IB questions and IB style questions

Q1 A compact disc player uses laser light to read a disc. The height of one pit on the CD is about $\frac{1}{4}$ of the wavelength of the laser light. The light illuminates the edge of a pit. Which of the following is correct with reference to the interference of the light and the binary information registered?

	Interference	Binary information
A	constructive	0
B	destructive	1
C	constructive	1
D	destructive	0

Higher Level Paper 1, May 09 TZ2, Q38

Q2 Laser light of wavelength 560 nm is used to read the information on a CD. The approximate depth for a pit on the CD is

A 1120 nm. **B** 560 nm.
C 280 nm. **D** 140 nm.

Q3 Two CCDs from digital cameras are identical in all respects, except that the pixels (photodiodes) of one device have a greater quantum efficiency. This digital camera will

A have a greater resolution.
B have a greater magnification.
C use less solar energy.
D be able to record images at lower intensities of light.

Q4 The maximum number of voltage levels that can be converted into a three-bit binary number (string) are

A 18 **B** 5
C 8 **D** 3

Q5 A CCD camera is used to record a digital picture of an object of length 60 m. The image of the object on the chip measures 0.06 mm. What is the magnification?

A 10^7 **B** 10^{-3}
C 10^3 **D** 10^{-6}

Q6 A CCD camera is used to capture the image of a painting. The area of the painting is 2.0 m^2 and the area of the image is 500 mm^2. Which of the following is the linear magnification of the image?

A 2.5×10^{-4}
B 5.0×10^{-3}
C 2.0×10^2
D 4.0×10^4

Higher Level Paper 1, May 11 TZ2, Q39

Q7 An analogue signal is sampled at time intervals t. Each sample is converted to a digital number having n bits. What is the total number of bits produced in time T?

A $\frac{nT}{t}$ **B** $\frac{nt}{T}$
C nT **D** $\frac{n}{Tt}$

Higher Level Paper 1, May 11 TZ2, Q40

Q8 To retrieve information stored on a CD, light of wavelength 780 nm is used. To retrieve information stored on a DVD, light of wavelength 650 nm is used.
Which of the following gives the ratio pit height of the CD/pit height of the DVD?

A 4 **B** 1.2
C 1 **D** 0.83

Q9 N photons are incident on a pixel of a CCD. Each photon causes one electron of charge e to be emitted. The capacitance of the pixel is C. What is the resulting potential difference across the pixel?

A NeC

B $\frac{C}{Ne}$

C $\frac{Ne}{C}$

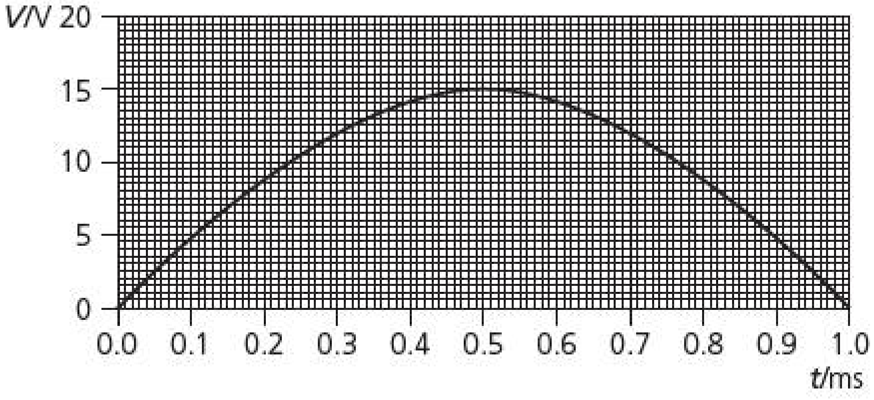
D $\frac{Ce}{N}$

Higher Level Paper 1, May 10 TZ2, Q39

Paper 2 IB questions and IB style questions

Q1 This question is about storing information on a CD.

The graph shows how the voltage V of part of an analogue signal varies with a time t.



In order to convert this signal to a digital signal that can be stored on a CD, the signal voltage is measured at regular time intervals. The measured value is then converted into four-bit binary number by dividing the signal into 1 V levels.

a State:

i the value of the voltage at 0.30 ms. [1]

ii the four-bit binary number corresponding to the value of the voltage at 0.30 ms. [1]

iii and explain the value of the least significant bit of the four-bit binary number in **a ii**. [1]

b The binary number in **a** is encoded onto the surface of a CD as a series of pits. Outline, with the use of an appropriate diagram, how light from a laser is used to distinguish between a binary 0 and a binary 1. [1]

Standard Level Paper 3, Nov 09, QC1

Q2 a A digital camera is used to take a photograph of a plant. The CCD in the camera has 1.6×10^7 square pixels. Each pixel has an area of $2.3 \times 10^{-10}\,m^2$. A particular leaf of the plant has an area of $2.5 \times 10^{-2}\,m^2$. The image of the leaf formed on the CCD is $1.0 \times 10^{-3}\,m^2$. Two indentations on the leaf are separated by 0.50 mm. Deduce whether the images of the two indentations will be resolved. [4]

b Light is incident on the image collection area for a time of 100 ms. The number of photons incident on one pixel is 5.5×10^4. Each pixel has a quantum efficiency of 80% and a capacitance 40 pF.

i State what is meant by *quantum efficiency*. [1]

ii Estimate the change in potential difference across each pixel. [4]

c Outline how the variation in potential difference across individual pixels enables a black and white image to be produced by a digital camera. [2]

Higher Level Paper 2, Specimen Paper 09, QB1 (Part 2)

Answers to self-assessment questions in Chapters 1 to 14

1 Physics and physical measurement

1 a 10^{-3} kg
b 10^{-3} kg
c 10 kg
d 10^{21} kg

2 a 10 m
b 10^{8}
c 10^{6}
d 10^{-4} m

3 a 2×10^{9} s
b 10^{8} s
c 1×10^{-8} s

4 a About 10 to 1; 1 order of magnitude
b About 10^{25} to 1; 25 orders of magnitude
c About 10^{-9} to 1; 9 orders of magnitude

5 a ± 3 g
b 0.63%

6 a 4.2 (m)
b a: 4.7%; t: 7.1%
c 19%
d ± 0.8 (m)

7 ± 0.03

2 Mechanics

1 a 1.9 m s^{-1}
b 2.7 m
c 1.0 m s^{-2}

2 a 1540 m
b 35.8 s

3 a 4100 m
b 1600 s

4 a 3.0 m s^{-2}
b 24 m s^{-1}

5 a 5.3 s after the police car started to move
b No; private car has travelled 210 m; police car has travelled 192 m
c After 12 s

6 a –13 m s^{-2}
b 32 m s^{-1}

7 a 8.72 km s^{-1}
b 3.01×10^{6} km

8 a 0.65 s
b 7.8 m s^{-1}
c 7.8 m s^{-1}

9 After 1.2 s or 3.3 s

10 0.34 s later

11 a 3.77 s later
b 24.2 m s^{-1}
c 29.9 m

12 a 10.1 m
b 24.8 m

13 5.5 m s^{-1}

14 The acceleration of the driver was twice the acceleration due to gravity, so the force acting on him was twice the driver's weight

15 a The distance increases as they fall
b Consider $s = ut + \frac{1}{2}at^2$ for both stones from the moment the second stone is dropped; the first stone will always have travelled an extra distance of ut, which increases with time

16 a 17 cm, ignoring air resistance
b For example, a 60 cm height results from a take-off speed of 3.4 m s^{-1}
c For example, a 30 cm height reduction suggests an acceleration of about 20 m s^{-2}
d About 0.2 s

18 a 13.0 s
b 127 m s^{-1}

19 a Acceleration for the first 100 s; then a constant speed until 200 s; followed by a deceleration to rest at 300 s
b 3600 m
c 20 m s^{-1}
d 12 m s^{-1}

20 b 2.13 m s^{-1} (men) and 1.92 m s^{-1} (women) in a 50 m pool in July 2009
c Swimmers can accelerate by pushing on the end walls of the pool

22 The runner starts from rest and accelerates towards a reference point (zero displacement); after that the runner has a constant speed in the opposite direction

23 a, c The object is moving (oscillating) backwards and forwards about a reference point; its greatest speed is in the middle of its motion and it decelerates as it moves away until its speed is zero at the maximum displacement; then it accelerates back
b Equal magnitude; opposite directions
d 8 cm s^{-1} and 0 cm s^{-1}
e A pendulum or a mass on the end of a spring

24 a Starts from rest, constant acceleration for 2 s; followed by constant velocity for 4 s; and finally a deceleration for 1 s to rest
b 1.5 m s^{-2}; 0 m s^{-2}; –3.0 m s^{-2}
c 16.5 m
d 2.4 m s^{-1}

25 2.8 m s^{-2} and 190 m

26 a The object experiences a constant acceleration for 8 s. In the first 4 s this decelerates the object to rest; then it accelerates in the opposite direction. Between 8 s and 12 s it has a constant velocity
b –3.0 m s^{-2}
c 96 m
d 48 m
e 8 m s^{-1}

29 10 m s^{-1}

30 Graph should show a constant velocity changing very quickly to another constant velocity of the opposite sign. The acceleration should be zero except during the impact

32 $50\,m\,s^{-1}$ to the east

33 $21.4\,m\,s^{-1}$ at an angle of 11° to the vertical

34 **a** $250\,km\,h^{-1}$ at 45° to the wind
b Air speed

35 **a** Observer on bus: the apple core will move backwards and downwards (it only moves backwards because of air resistance). Observer on the opposite side of the road: the apple core will move forwards and downwards
b They are both equally correct

36 **a** $1.23 \times 10^4\,N$
b 318 N
c $2.43 \times 10^{-3}\,N$

37 1.33 N

38 **a** $5.60 \times 10^6\,N$
b Approximately 50 000 kg; about 9%
c The forces stopping the motion of the plane when it lands are not as large as the forces accelerating the plane when it is taking off
d The mass has reduced by 180 tonnes; most of the fuel it was carrying has been burned

39 There is less of the mass of the Earth beneath it, and some of the mass of the Earth is above it

40 $8.9\,N\,kg^{-1}$; this is about 10% less than on Earth

41 A has twice the circumference, four times the surface area and eight times the volume, mass and weight

44 10.6 N at an angle of 41° to the 12 N force

45 9.1 N at an angle of 32° to the 7.7 N force

46 76 N at an angle of 14° to the 74 N force

47 **a** 3.33 N
b 1.72 N

48 **b** 14 N parallel to the slope; 30 N perpendicular to the slope

49 **a** $6.68 \times 10^4\,N$
b $1.74 \times 10^4\,N$
c $0.0045\,m\,s^{-2}$
d The component of weight (acting down the slope) of a heavy train is so large that it may be larger than the it resultant forward force provided by the engine

51 All arrows identical; pointing downwards

52 **a** A force vector of 300 N acting downwards from the middle of the suitcase, labelled weight; an equal and opposite arrow pushing up on the suitcase, labelled normal reaction force
b Add an arrow upwards from the handle of half the length of the previous vectors, labelled pull of hand, 150 N; the normal reaction force reduces to 150 N

53 The reading rises when a force is needed to accelerate the book upwards; and falls when the book is accelerated downwards

54 **a** Not if the elevator is moving with constant velocity; this is because the forces acting on you would be the same in all three cases
b In all diagrams the weight vector will be the same; if the person is accelerating, the force from the floor will be greater or less than the weight

55 **a** **i** Weight and air resistance on the sky diver's body will be equal and opposite; then there will be an upwards force from the parachute on the skydiver
ii The upwards forces from the parachute and air resistance on the skydiver will together be equal and opposite to the weight

56 14 N

57 Weight acting downwards from climber's centre of mass; tension acting along the rope, away from the climber; push of rock acting from climber's feet to the point where the other two forces cross. (This could be resolved into a normal force and friction.)

58 **a** The weight of B will be eight times greater than A; the air resistance acting on B will be four times greater than A
b A moves with constant velocity because the forces are balanced; B accelerates because there is a resultant force, since its weight is greater than the air resistance acting on it

59 $2.56 \times 10^5\,N$

60 2.4 N

61 **a** $2.39\,m\,s^{-2}$
b 1120 m

62 **a** $-3.5\,m\,s^{-2}$
b 4200 N

63 **b** $0.27\,m\,s^{-2}$
c A much greater force would be needed for the larger acceleration; the thin rope may not be strong enough, and may break

64 **a** 124 N
b 933 N

65 **a** $1.6\,m\,s^{-2}$
b 3.3 N
c (Tension = 16.4 N)

66 **a** $4.58\,m\,s^{-1}$
b 1570 N
c 2310 N
d Pushed down hard on the ground
e The impact takes longer; reducing the size of the deceleration and the force needed to stop him

67 **a** The impact takes longer reducing the size of the deceleration and the force needed to stop
b As **a**. The force is also spread over a large area, reducing the pressure

68 **a** $230\,m\,s^{-2}$
b 0.14 s

69 **a** 6800 N s
b $12\,m\,s^{-1}$

70 **a** $5\,m\,s^{-1}$ to the left
b $2.5\,kg\,m\,s^{-1}$ to the left
c 7.4 N to the left
d Force changing (linearly) from 0 to 14.8 N in 0.17 s; then back to zero in another 0.17 s
e Shorter times; greater forces
f Identical, except the forces are in opposite directions

71 $0.45\,m\,s^{-1}$ to the left

72 **a** $5.35\,m\,s^{-1}$
b Air resistance under these circumstances will be almost insignificant; so the actual speed will be close to the predicted value
c About $5 \times 10^{-23}\,m\,s^{-1}$; cannot be measured

73 $260\,m\,s^{-1}$

74 The ball is not an isolated system; it is acted on by the external force of gravity; the Earth loses an equal amount of momentum

75 She moves in the opposite direction with a speed of $2.0\,cm\,s^{-1}$; an external force is needed to stop her motion

76 $22.8\,cm\,s^{-1}$; in the original direction of A

77 **a** $1.8\,m\,s^{-1}$
b The total (kinetic) energy of the balls would have increased

78 At the 20 cm mark on the ruler

79 The moving air from the fan will collide with the sail and exert a forward force. But, in order for the air to be pushed forward by the fan, the air pushes backwards with an equal force on the fan. The resultant force on the boat will also depend on how the air flows past the sail

80 **a** 340 N
b $0.034\,m\,s^{-2}$
c 10 000 N s
d $1.0\,m\,s^{-1}$

81 The two forces are acting on the same object

82 **a** 59 N
b 130 N

83 29 J

84 **a** 54 N
b 140 J

85 0.11 J

86 **a** Your weight (say 600 N); 20 cm
b 60 J

87 **a** $320\,N\,m^{-1}$
b 1.0 J
c 33 cm
d A mass of 10 kg would exert a force of approximately 100 N, which is well beyond the range shown on the graph; the spring may not continue to stretch proportionally to its load for such a large force

89 **a** Loudspeaker
b Battery
c Nuclear power station
d Microphone
e Flame
f Car
g Bow and arrow
h Electrical generator
i Gas cooker
j Photovoltaic cell

90 **a** 2.0×10^6 J
b The field strength is slightly less at the top of the mountain; the difference is too small to affect the answer

91 **a** 1.5×10^7 J
b Energy is dissipated in the electric motor; because of friction and air resistance

92 **a** 5.1×10^6 J
b As the elevator comes down a significant amount of its decrease in gravitational potential energy is transferred to the rising counterweight, rather than being dissipated as internal energy

93 **a** **i** 130 J
ii 2100 J
b 3000 J

94 $15\,m\,s^{-1}$

95 4.1×10^{-16} J

96 900 kJ

97 **a** $3.39 \times 10^4\,kg\,m\,s^{-1}$
b 100 J
c 9.11×10^{-31} kg

98 1.27 m

99 10 m

100 **a** 22 m
b Measuring the time accurately; determining when the ball has reached the right height; the ball is not released from ground level

101 **a** $2.8\,m\,s^{-1}$
b $2.8\,m\,s^{-1}$
c It moves a smaller distance at a greater average speed because it has a greater acceleration (the average component of its weight down the slope is greater)

102 **a** Elastic strain energy.
b Gravitational potential energy to kinetic energy; to strain energy (and internal energy); to kinetic energy;

to gravitational potential energy. This assumes that air resistance was negligible
c 4.4 m and 3.4 m
d 22 cm; assuming that it loses the same fraction of its kinetic energy each time it bounces

103 1.1 $m s^{-1}$

104 3.0×10^5 N

105 The average force exerted in the accident can be calculated from the kinetic energy before the collision divided by the distance the vehicles 'crumple'; the greater the deformation, the less the force

106 a 16 000 J
b The work done in stretching the cord
c Equal increases in force produce smaller and smaller extensions (except for the first 250 N); the cord becomes stiffer
d Approximately 25 m

107 a 2790 J
b Probably not; because this is quite a large mass falling a short distance
c 3400 N

108 To increase the length (and time) of the impact; and so reduce the force and prevent injury to the knees and legs

109 a 6700 N
b The hammer comes to rest

110 a 240 J
b 13 W

111 About 25 kW

112 1400 N

113 a 1.7×10^8 W
b 170 000 homes

114 900 MW

115 a 338 N
b 54 N; friction helps to stop the box sliding down the slope when the force is reduced
c 830 J
d 0.86 (86%)

116 a 0.093 $m s^{-2}$
b 0.040 N towards the centre of the circle.
c 17 s

117 a 372 $m s^{-2}$
b 1500 N

118 a 1010 $m s^{-1}$
b 2.7×10^{-3} $m s^{-2}$

119 a 4500 N
b Friction between the tyres and the road
c The force needed will be four times greater; there may not be enough friction
d Water and/or ice will reduce friction between tyre and road
e The centripetal force needed for this curve in the road will increase to 6400 N; but the extra weight will probably increase the friction as well

120 a 3.0 $m s^{-1}$
b 0.53 revolutions per second

121 a 2.1 N
b 637.7 N
c (Normal reaction force acting upwards on boy = 635.6 N)

122 b 0.68 N
c 1.3 $m s^{-1}$ and 1.4 s

123 There is no air resistance to slow it down and the only force acting on it is the force of gravity, which does not change its speed because it is always acting perpendicular to motion

3 Thermal physics

1 a 331 K; 184 K
b 38 K

2 a 127 °C
b 32 cm^3

3 a Probably not; they may have the same temperature, but they are different substances; there will probably be different numbers of molecules with different masses and different speeds
b Probably not; different substances require different amounts of energy for the same temperature rise

4 a i A relatively large attractive force pulls the molecules back together
ii A large repulsive force pushes the molecules apart
b Forces between gas molecules are much, much smaller; almost zero
c If the separation is ten times greater in each dimension, the average volume occupied by one molecule must be 10^3 times greater

5 Each 'sparkle' is so small that it contains very little internal energy, although it has a high temperature. If a sparkle comes in contact with the skin it transfers only a tiny amount of thermal energy as it cools down rapidly

6 a 108 g
b 5.7 g
c 167 mol
d 1.4×10^{25} molecules

7 a 0.062 mol
b 1.9×10^{11} atoms per second

8 a 200 kg
b 6700 mol
c 4.0×10^{27} molecules

9 a 10 cm^3
b 1.7×10^{-23} cm^3
c 2.6×10^{-8} cm

10 143 g

11 9.5×10^4 J

12 1370 $J kg^{-1} K^{-1}$

13 4.8 K

14 34 °C

15 4.8×10^4 J

16 40 °C

17 128 s

18 23 °C

19 41 °C

20 980 W

21 **a** 3670 J
b To reach thermal equilibrium with the oven
c To reduce the thermal energy transferred to the surroundings
d $313\,J\,kg^{-1}\,K^{-1}$
e To make sure that it was all at the same temperature
f Underestimate; the final temperature would have been higher if there had been no transfer of energy to the surroundings

22 $4.37 \times 10^5\,J\,kg^{-1}$

23 **a** 38.3 °C
b All of the thermal energy that flowed out of the hot water went into the cold water; none went into the surroundings

24 17.3 °C

25 320 s

26 $3.4 \times 10^5\,J\,K^{-1}$

27 840 s

28 **a** $1.80 \times 10^5\,J$
b 67 °C
c There would have been less time for energy to be transferred to the surroundings

29 **a** 1.4 °C
b All the kinetic energy of the bullet is transferred to internal energy in the block and bullet
c Half the speed has a quarter of the kinetic energy; so temperature rise would be 0.36 °C

30 **a** $4 \times 10^4\,J$
b 19.8 °C
c The temperature rises are too small to be easily measured with accuracy

31 **a** 1.1 °C
b This assumes that all of the gravitational potential energy of the lead is transferred to internal energy in the lead; no energy is transferred to the surroundings
c Twice the potential energy has to be spread to twice the mass

32 0.12 °C

33 1600 J

34 $2.56 \times 10^6\,J\,kg^{-1}$

35 More bonds between molecules have to be broken during boiling than during melting

36 **a** **i** $1.2 \times 10^3\,J$;
ii 144 J
b The energy transferred from the steam as it condenses is much greater than that from the water as it cools

37 **a** Some ice will melt because of thermal energy absorbed from the surroundings; by comparing the two masses of melted ice we can estimate the amount melted by the heater in A
b $4.0 \times 10^5\,J\,kg^{-1}$
c Some energy from the heater is transferred to the surroundings
d Carry out the experiment in a refrigerator; insulate the apparatus

38 $5.2 \times 10^4\,J$

39 $8.0 \times 10^{10}\,J$

40 10 °C

41 **a** 0.1 N
b $1.5 \times 10^5\,N$
c Because there is equal pressure inside the body

42 Because of the extra weight of the water above them

43 **a** Pressure under our feet is due to the force with which we push down – our weight, if we are standing still. The pressure in a gas is explained by random molecular collisions with surfaces, which can be in any direction
b The pressure in a liquid also acts in all directions

44 **a** The molecules are closer together
b Electric forces act across space, without contact, and in theory they will never go completely to zero; but the forces become very small (negligible) if the separation of molecules is more than a few molecular widths

45 **a** When the molecules of the warmer gas hit the molecules of the colder surface, on average, kinetic energy will be transferred to the molecules in the wall.
b Internal energy and the temperature of the wall will increase; the internal energy and temperature of the gas will decrease

46 **a** The molecules will have stopped moving, and stopped colliding with the walls
b A gas will condense to a liquid if its molecules do not have enough kinetic energy to overcome the electric forces between them when they get close together

47 The graphs have the same shape; but for all volumes at the the higher temperature, the pressures will be greater

4 Oscillations and waves

1 **a** 0.872 s
b 1.15 Hz
c 344
d The absolute uncertainty in measurement remains the same; but the percentage uncertainty decreases with larger measurements

2 3.1×10^{-4} s

3 **a** 2.0×10^{8} Hz
b 200 MHz

4 **a** 4 cm
b 0.5 s
c –1.4 cm
d +1.4 cm

5 0.045 s and 140 rad s^{-1}

6 450 rad s^{-1}

7 **a** 7.3×10^{-5} rad s^{-1}
b 7.2 rad

8 0.131 s

9 **a** 3.14 rad (π)
b 1.57 rad ($\pi/2$)
c 0.785 rad ($\pi/4$)
d 1.95 rad

10 14 cm

11 **a** 9.2 rad s^{-1}
b 0.68 s

12 **a** 3.9 m s^{-2}
b Simple harmonic motion

13 –3.5 cm and –11 cm s^{-1}

14 **a** 0.23 m s^{-1} and 1.2×10^{-3} J
b –2.9 mm

15 5.63 Hz

16 **a** 0.39 m s^{-1}
b 6.0 cm

17 1.61 m above the low tide level

18 The area under each half of an oscillation is equal to twice the displacement

19 **a** 4.32 s
b The accelerations are equal because $a = F/m$; the heavier pendulum has twice the mass, but also twice the weight
c Amplitude does not affect period; if the amplitude is doubled, the restoring force is also doubled (for small amplitudes)
d 1.45 rad s^{-1}
e 3.97 J
f 1.75 m s^{-1}

20 0.83 m and 4.44 m s^{-1}

21 **a** 0.58 s
b 2.2 m s^{-1}

22 Same time period; reducing amplitude

24 **a** The graph should have three resonant peaks
b Fit rubber padding between the mirror and the car body (to change the stiffness of the mounting and dissipate energy)
c The second graph should be significantly below the first

25 **a** 0.155 m
b 4510 m s^{-1}
c The particles are closer together; there are much larger forces between them

26 **a** 5.4×10^{5} Hz
b **i** 4.7×10^{-3} s
ii 0.11 s
iii Between 180 s and 1300 s

27 **a** 200 kW
b The efficiency of the power station is constant
c 2.1 m
d 35 kW

28 **a** 6.8×10^{13} MHz
b (Photons of) higher frequencies transfer more energy

29 **a** 5.5×10^{-7} m
b Yellow
c Ultraviolet

30 3.6 s

32 26°

33 Wavefronts in shallower water will be at 27° to the boundary

34 **a** 1.35
b 35°

35 51°

37 **a** 49°
b The light will be totally reflected back inside the water

38 Microwaves are part of the electromagnetic spectrum; with a wavelength suitable for diffraction by gaps and obstacles with a size of a few centimetres

39 55 cm and 610 Hz

40 Destructive interference

41 **a** When he is the same distance from both speakers, the waves he receives have both travelled the same distance and will interfere constructively. If he moves in any direction there will then be a path difference between the waves and they will no longer interfere perfectly constructively
b 71 cm

42 The waves from the two sources of light are not coherent

43 **a** To maximize diffraction of the waves emerging from the slits; so that they cross over each other and interfere
b 12, 6, 4, 3 cm … etc.
c Slowly move the receiver until it detects an adjacent maximum; the difference in the two path differences will equal the wavelength

5 Electric currents

1 **a** 36 J
b 860 J

2 **a** 220 V
b 40 C

3 P: 9.1 V; Q 1.1 V

4 **a** **i** 7.7×10^{-17} J
ii 480 eV
b No; its kinetic energy is not large enough to overcome the repulsive force

5 **a** 3.5×10^{-13} J
b 1.0×10^{7} m s^{-1}

6 8.0×10^{-16} J

7 a 0.25 C
b 1.6×10^{18}

8 a 432 C
b 648 J

9 37.5 V

10 27 Ω

11 1.2×10^{-4} A (0.12 mA)

12 136 V

13 a 1.17 m
b 120 Ω

14 They are inversely proportional

15 5.5×10^{-4} m

16 2.1×10^{-6} Ω m

17 a i 0.05 A
ii 0.60 W
iii 72 J
b 120 Ω

18 a 9.07 A
b 220 V

19 a 2×10^{5} W
b A power 'loss' of only $10\,W\,m^{-1}$ may seem quite low; but it can be considerable for long cables if high currents are used

20 a 150 W
b Heater, ammeter and power supply connected in series; voltmeter connected across the heater

21 a 0.86 A
b Energy is dissipated; because of friction and resistive heating

22 a 9.68 Ω
b 11.4 A
c The heater's resistance is lower when it is colder

23 a 5.5 Ω
b 0.37 A
c Internal resistance is constant

24 a 3.9 Ω
b 3.0 V
c i 2.0 W
ii 0.12 W

25 a 12.5 V
b 0.28 Ω
c 4.7 Ω

26 a To get a high power from a low voltage ($P = VI$)
b The voltage across the battery falls because of the 'lost volts' (Ir) due to the internal resistance of the battery

27 a 48 A
b No resistance in the wires causing the short circuit; internal resistance constant
c 580 W
d Rapid rise in temperature; the battery may be damaged

28 Usually a voltmeter would be connected across the resistor; but a voltmeter connected across a battery does not measure the p.d. across its internal resistance

29 a 4.5 V; 2.4 Ω
b 1.5 V; 0.27 Ω
c i 0.80 A and 2.6 V
ii 0.43 A and 1.4 V

30 a 3.0 Ω
b 4.8 V

32 In the 2 Ω resistor; 4 Ω and 6 Ω in parallel have a total resistance of 2.4 Ω, so that the voltage across them is (only) slightly higher than the voltage across the 2 Ω resistor; because $P = V^2/R$, the lower resistance of the 2 Ω is the dominant factor

33 a Low: two in series; medium: only one; high: two in parallel
b 605 W, 1210 W, 2420 W

34 a 4.4 Ω
b 0.29 A
c 0.70 V

35 a The ammeter will show a very low reading because the large resistance of the voltmeter prevents a greater current; the voltmeter will show (almost) 12 V because it has a resistance very much larger than 30 Ω; in a series circuit the voltages are in the same ratio as the resistances
b 4.0 V and 0.4 A
c Because of the resistance of the voltmeter; the total resistance between the terminals of the voltmeter is not just 10 Ω, it is calculated from the two resistances in parallel
d 1300 Ω
e Negligible resistance

36 a 4.3 V
b 5.2 Ω; 3.6 W
c The lamp becomes dimmer; the current reduces when the overall resistance of the circuit increases
d 3.4 V (assuming R is constant)

37 a 6.0 V
b 2.4 Ω
c The student does not understand that the value of 6.0 V will change when the bulb is connected because the total resistance between A and B is no longer 100 Ω
d 0.28 V; no
e Connect the lamp in series with an equal resistance (2.4 Ω).

38 a The circuit should include a variable resistance and a thermistor connected in series to a power source; the output to the cooler control circuit will be taken across the variable resistance, so that if the temperature rises the output voltage to the control circuit also rises and turns the cooler on
b Refrigerators, water heaters, room heaters, irons, air-conditioners, ovens

39 a The energy of the light releases more free electrons
b It would not be possible to plot or interpret the graph accurately for low values if a linear scale was used; (because of the very large

difference in the magnitude of different values)

c $\log R = 5.2 - (0.8 \times \log I)$

d 2300 Ω

e 5200 Ω

40 a 2.02%

b 0.942 V, assuming the voltmeter has infinite resistance

c 0.944 V

41 b 96 Ω

c 0–10 Ω

d 4 V

e The voltage would only be 6 V if the potentiometer was used on its own; the lamp is in parallel with half of the potentiometer, and their combined resistance is less than the other half of the potentiometer

6 Fields and forces

1 Around 4×10^{-11} N

2 6×10^{-4} N

3 3.6×10^{22} N

4 3.7×10^{-47} N

5 3.7 N kg^{-1}

6 a 8.9 N kg^{-1}

b 91%

7 10.3 N kg^{-1}

8 9.81 N kg^{-1}

10 b 1.42 N kg^{-1}

11 a B has eight times the mass of A

b 0.057 N kg^{-1}

c 0.086 N kg^{-1} at an angle of 48° to the line joining B and Z

12 c 379 000 km from the centre of the Earth

13 a All three on a straight line, with the Moon between the Earth and the Sun

b All three on a straight line, with the Earth between the Moon and the Sun

14 15 cm

15 a Three forces are acting: weight, tension and the repulsive force between the charges

b 4.3×10^{-4} N

c 54 nC

16 3.7×10^{-6} N to the left

17 5.4×10^{-3} N, diagonally away from the square

18 a 8.2×10^{-8} N

b 10^{39} : 1

21 There will be some charge separation on the uncharged sphere; the two spheres are then attracted to each other and hang at equal angles to the vertical

22 a 1.4×10^{3} N C^{-1}

b 4.0×10^{-4} N

23 a 1.6×10^{5} N C^{-1}

b 0.092 N

24 a 3.4×10^{-5} C

b 24 cm

25 a 5.1×10^{11} N C^{-1}

b 8.2×10^{-8} N

c The same magnitude but in the opposite direction

d The force may be considered to maintain the motion of the electron in a circular path around the proton

26 a Zero

b 2.0×10^{6} N C^{-1} and 2.2×10^{6} N C^{-1}

27 1.9×10^{6} N C^{-1} at an angle of 23° to a line joining P to the negative charge

28 a 5000 V m^{-1}

b When the separation halves, the field doubles

c 2.4×10^{-5} N

d 0.2 ms^{-2}

e Towards the positive plate

f The acceleration to the positive plate is combined with a greater downwards acceleration to give a curved trajectory

29 a 1.5×10^{4} V

b Water vapour helps conduct electricity

c Less charge can build up on objects if there is water vapout present

30 a Q downwards; P to the right

b The field strength at R is one-third of the field strength at Q (it is three times further away)

31 a The needle will become more aligned with the solenoid

b Towards the north-west

32 a 4.3×10^{-3} N

b 7.5×10^{-3} N

c 8.6×10^{-3} N

d 0

33 a 2.4×10^{-3} N m^{-1}

b From west to east

34 b 3.1 A

35 a 37°

b 5.7×10^{-3} N (the same)

37 a Only if it is moving parallel to the magnetic field

b Unlike magnetic forces, electric and gravitational forces exist regardless of the direction of motion (indeed, motion is not needed for the forces to exist)

38 a 6.8×10^{-13} N

b 1.0×10^{14} m s^{-2}

c Helical

39 a 1.3 T

b 1.1×10^{-25} kg

c 70 cm

40 a 7450 eV

b 1.19×10^{-15} J

c 5.12×10^{7} m s^{-1}

d 1.97×10^{-3} T

e 10.5 cm

41 a 1.0×10^{7} m s^{-1}

b 2.0×10^{5} V m^{-1}

c The fields need to be perpendicular to each other; and to the direction of movement of the particle

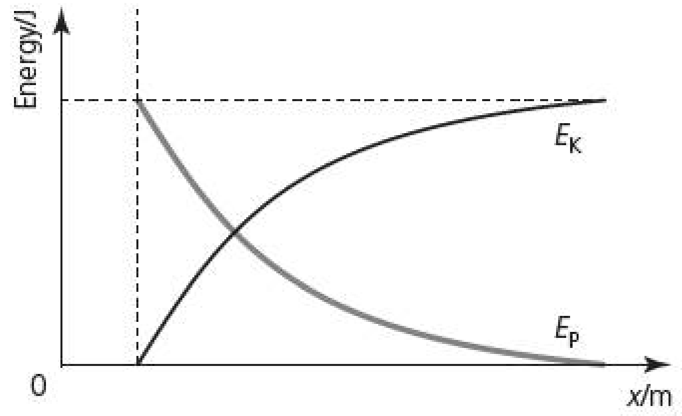
7 Atomic and nuclear physics

1 Mass = 9.3352×10^{-26} kg; charge = zero

2 3.9×10^{-29} m^3

3 At surface of nitrogen nucleus: 2.8×10^{20} N C^{-1}; over sphere radius 10^{-10} m: 1.0×10^{12} N C^{-1}

4



5 **a** Because neutrons are not charged they would not be deflected so it is unlikely that any would 'bounce back'; some could be absorbed by the nuclei
b The charge on the nucleus is much smaller so the deflection would be smaller
c To allow alpha particles to only interact with single gold nuclei and to avoid absorption of alpha particles

6 **a** Gaseous atoms of different elements are absorbing specific frequencies of light from the solar emission spectrum
b Each element has its own unique emission spectrum

9 **a** $f = 1.18 \times 10^{15}$ Hz; $E = 7.80 \times 10^{-19}$ J
b 470 kJ

10 13 nm

11 Iodine-129: 53 protons, 53 electrons and 76 neutrons
Caesium-137: 55 protons, 55 electrons and 82 neutrons
Strontium-90: 38 protons, 38 electrons and 52 neutrons

12 $+3.20 \times 10^{-19}$ C

13 $^{32}_{16}S$

14 13

15 **a** Strong nuclear force
b Electric (Coulomb) force

16 1 cm

17 Alpha particles have slower speeds than beta particles of the same energy; but their momentum is greater and they are less readily deflected during collision; at close approach there is electrostatic attraction with electrons of the atom or molecule it collided with.

18 At lower pressure the number of particles per m^3 in the air is reduced; the number of collisions is decreased; the ionizing radiation travels further before its energy is dissipated in collisions

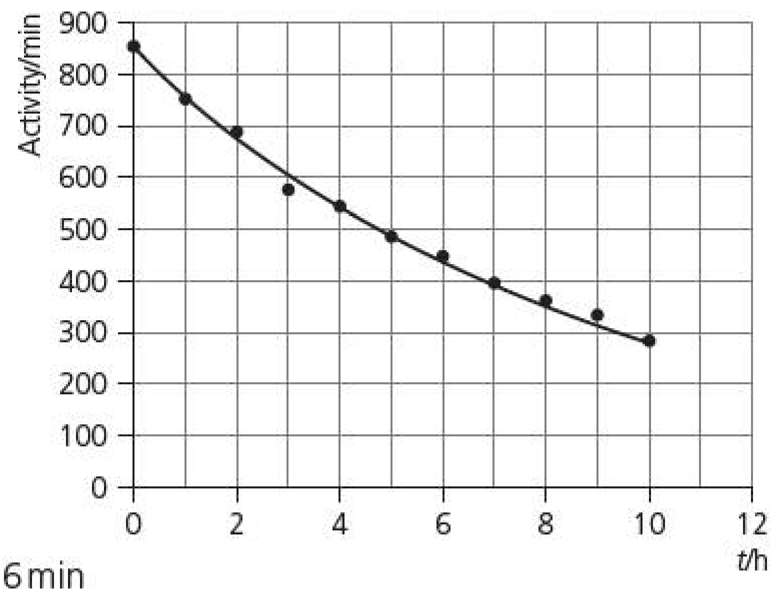
21 Beta-negative emitters are found above the line of stability (stable nuclei); a neutron is replaced in the nucleus by a proton with the emission of an electron; so that the neutron/proton ratio decreases

Beta-negative decay is above the line because the nuclei are 'neutron-rich'; neutrons turn into protons to make the nucleus closer to that line

Alpha emitters are also found above the line of stability; during alpha decay, two protons and two neutrons are expelled; the nucleus moves parallel to the line of neutron = proton number

22 Graph passes through (0, 8000), (5, 4000), (10, 2000), (15, 1000), (20, 500) and (25, 250)

23 **a**



b 6 min

24 80 min^{-1}

25 **a** 1/32 (or 0.31)
b 15/16 (or 0.94)
c 6.25×10^{13} Bq
d 160 min
e **i** 7/8
ii 1/16

26 $^{232}_{90}Th \xrightarrow{\text{alpha}} {}^{228}_{88}Ra \xrightarrow{\text{beta}} {}^{228}_{89}Ac \xrightarrow{\text{beta}} {}^{228}_{90}Th$

27 $^{184}_{78}Pt \rightarrow {}^{180}_{76}Os + {}^{4}_{2}\alpha$; $^{24}_{11}Na \rightarrow {}^{24}_{12}Mg + {}^{0}_{-1}\beta$;
$^{60}_{27}Co^* \rightarrow {}^{60}_{27}Co$

28 $^{1}_{1}H + {}^{7}_{3}Li \rightarrow {}^{8}_{4}Be \rightarrow 2{}^{4}_{2}He$

29 Three neutrons; the heaviest elements have the largest neutron excess to remain stable; the two lighter fission fragments have a higher fractional neutron excess; hence some are 'left over'

30 4.5×10^{16} J

31 2.0×10^{-13} J

32 4.67×10^{-13} kg

33 4.8 MeV

34 7.3 MeV

35 1010 MeV

36 **a** 236.132u; 235.918 u
b 0.214 u
c 199.3 MeV
d 8.14×10^{9} J

37 ${}^{2}_{1}H + {}^{1}_{1}p \rightarrow {}^{3}_{2}He$; 4.96 MeV

$2\,{}^{3}_{2}He \rightarrow {}^{4}_{2}He + {}^{1}_{1}p + {}^{1}_{1}p$; 13.92 MeV

38 28 MeV

39 4.05 MeV

40 0.01 MeV

8 Energy, power and climate change

1 About 35%

3 **a** 1 kg of water at 35 °C
b The source of energy must be hotter than its surroundings
c Because all the things that surround us are at similar temperatures; the temperature differences are not great enough for efficient transfers of energy to work

4 **a** 12.1 MJ kg^{-1}
b Higher; some thermal energy transferred from the burning fuel was spread into the surroundings, rather than into the water

5 **a** 1.1×10^{11} J
b 3600 W
c 27%
d Factories, schools, offices, shops, transportation, etc.

6 **a** 20 g
b 2.7 g s^{-1}
c 710 J g^{-1}

7 **a** **i** 31 kg
ii 1.9×10^{7} kg
b 130 kg

9 Oil and renewables sources would have a lower percentage; the others would be higher

11 57 kg

12 **a** 5.0 kg
b 2.9×10^{5} N
c 1.0×10^{5} kg

13 42 MJ kg^{-1}

14 **a** 34%
b 6.5×10^{4} kg s^{-1}
c 9300 kg

16 **a** 1.4×10^{7} m s^{-1}
b 1.4×10^{4} m s^{-1}

17 Their physical properties depend on their masses; there is only about a 1% difference in mass between the two nuclides; in a mixture, there will be considerable overlap of kinetic energies and momenta of nuclides from the different isotopes

18 **a** 7.6×10^{19}
b 2.6 kg

19 **a** 7.8×10^{13} J kg^{-1}
b 2.3×10^{12} J kg^{-1}
c Uranium is about 10^{5} times more energy dense than coal

20 ${}^{1}_{0}n + {}^{235}_{92}U \rightarrow {}^{236}_{92}U \rightarrow {}^{138}_{55}Cs + {}^{96}_{37}Rb + 2\,{}^{1}_{0}n$ + photons

21 **a** 3.2×10^{10} J
b 0.4 g

22 The three naturally occurring isotopes of uranium have very long half-lives; the half-life of uranium-238 is comparable to the age of the Earth

23 The answer, of course, depends on what level is considered to be safe; for example, after about 60 years the level will still be about double the acceptable safety level

24 **a** 100 000 years is a long time and the area may become prone to earthquakes, volcanoes or other unexpected natural disasters or dramatic changes in climate
b There would be a small risk that something might go wrong during the launch; resulting in the radioactive material being scattered over a wide area of the Earth

25 **a** 9.2 N
b 2.8×10^{27} m s^{-2}

30 **a** 5.9×10^{24} J
b The total amount received is of the order of 10^{4} times greater than the consumption

33 Variation in the Sun's activity; variation in the distance between the Earth and the Sun

34 The metal pipes are good conductors of thermal energy; the glass cover prevents convection currents; black surfaces are good absorbers of radiation

36 **a** 0.023 W
b 0.031 A

c 2200
d 0.39 m^2

39 **a** 9.4×10^7 J
b 33 °C
c All of the energy was transferred to the water in the tank; none was transferred to the surroundings; or retained in the solar panel

40 **c** Greater angle of incidence and radiation has more atmosphere to pass through

41 2.2×10^6 W

42 1.7×10^{14} J

43 **a** 98 cm
b All the rain that falls over the area flows into the lake; there is no evaporation from the lake; these are unreasonable assumptions

44 6.1 kg

47 The energy wasted in a pumped storage system is significantly less than the energy wasted if a power station is operated for many hours every night well below its most efficient power

48 **b** 8.9×10^{12} J

49 A large wind generator is needed to produce a relatively small power

50 2.4×10^5 W

51 **a** 10 $m\,s^{-1}$
b 320 kW

52 4.4 m

53 **a** Diameter = 140 m (assuming efficiency = 25%, effective average wind speed = 8 $m\,s^{-1}$, density of air = 1.3 $kg\,m^{-3}$)
b A larger generator would be expected to be more efficient; but design and construction problems (and the cost) of such a large structure may be too great

54 **a** 400
b 36 km^2
c So that each generator does not affect the flow of wind to the others

55 Calculations involve the wind speed cubed, and the cube of the average wind speed is a lot less than the average of the speeds cubed

57 The predicted value is too high; a water wave with a sine wave shape has less mass displaced

58 **a** 0.17 Hz
b 0.32 $m\,s^{-1}$
c 340 W
d 1.9×10^9 J
e It is dissipated as thermal energy into the surroundings

59 13 kW

62 **a** 593 $W\,m^{-2}$
b 463 $W\,m^{-2}$

63 About 1.4 billion km; Saturn

64 0.49 $W\,m^{-2}$

65 **a** Difference in the growth of trees and other plants; snow and ice in winter; variation in cloud cover; variation in angle of incidence
b Variation in cloud cover; variation in angle of incidence

66 600 $W\,m^{-2}$

67 **a** 1.8×10^{25} W
b It was a perfect black body
c 0.25 $W\,m^{-2}$

68 **a** 1.5 m^2
b 700 W
c Conduction from the skin into the air; convection of warm air currents away from the body
d The body will also receive thermal energy from the surroundings

69 **a** 44%
b 75 °C

70 2500 °C

71 **a** 410 K
b It acts as a black body; and has no atmosphere
c The received power and the radiated power would both be double

72 0.87

73 **a** 0.029 °C
b In practice, a smaller temperature rise will occur and there will be some dissipation of thermal energy overnight; it takes a long time for large bodies of water to warm up as the weather gets hotter

76 Carbon dioxide: about 15%; nitrous oxide: about 9%; methane: about 14%

77 **a** A 50 m × 50 m × 50 m lump of ice would have a mass of about 115 000 tonnes, and surface area of 15 000 m^2
b 3.8×10^{13} J
d About 1.6 years

78 **a** 7 cm
b No consideration has been given to water entering or leaving the lake; by evaporation, seepage into the land, rainfall, rivers, etc.

79 **a** 5.8×10^{24} J
b About one year

80 **a** 2.26×10^6 J are needed to turn 1 kg of water into steam at 100 °C
b 1.1×10^{12} J
c The gravitational energy is very much smaller
d About an hour

9 Motion in fields

1 7.69 $m\,s^{-1}$ and 27.3 $m\,s^{-1}$

2 **a** 17 $km\,h^{-1}$
b 2 min

3 **a** 15 $m\,s^{-1}$
b **i** Same
ii Less

6 **a** 0.167 s
b 0.137 m below the centre

7 **a** 46 m above sea
b 30 $m\,s^{-1}$ downwards
c 4.7 s

d 96 m
e 36 m s^{-1} at 56° to horizontal

8 528 m

9 **a** 23°; v_V = 1.5 m s^{-1}, v_H = 3.5 m s^{-1}
b 98 cm

10 **b** There was no air resistance or friction
c Larger masses will experience greater gravitational forces, but the same accleration, $a = F/m$ (assuming there is no friction or air resistance)

11 28 m s^{-1}

12 –2.0 × 10^8 J

13 Although the distance is greater, the average component of force is less

14 **a** 1.3 × 10^{11} J
b The engines are not efficient; the object will also have kinetic energy; energy is needed to raise the fuel and its tanks; energy will be used in overcoming air resistance

15 Because there is no component of force in the direction of motion

16 **a** –5.9 × 10^7 J kg^{-1}
b 7.0 ×10^{11} J

17 **a** 6.3 ×10^7 J would have to be given to 1 kg to move it to infinity from the Earth's surface
b P: –1.6 × 10^7 J; Q: –6.25 × 10^6 J; R: –3.1 × 10^6 J
c 7.6 × 10^9 J

18 5.0 N kg^{-1}

20 **a** Closer to the Earth

19 **b** 6.25 × 10^6 J kg^{-1}

21 20 N kg^{-1}

25 **a** 2.4 km s^{-1}
b An object escaping from the Moon also has to escape the Earth's gravitational field
c The Earth is nearer than infinity and will attract the vehicle

26 10.9 km s^{-1}

27 6.3 × 10^{10} m

28 **a** 3.0 × 10^3 m
b The Sun's radius is roughly 200 000 times bigger

29 The escape speed increases by about 1%; which may be considered insignificant

30 4.1 × 10^3 m s^{-1}

31 4.3 × 10^6 V

32 14 V

33 68 cm

34 +5.6 × 10^{-7} C

35 **a** + 4.5 × 10^{-7} C
b 14 000 V m^{-1}

38 **a** –4.17 × 10^{-8} C
b Circular lines with radii 50, 75 and 150 cm
c Radial lines pointing inwards

39 R: –540 V; S: 0 V; T: +300 V

40 2.0 × 10^{-9} C

42 **a** 5.0 × 10^4 V m^{-1}
b 0.039 N
c Downwards
d 3000 V
e 2.3 × 10^{-3} J
f Gains kinetic energy

43 250 km: 9.1 N kg^{-1}; 7800 m s^{-1}; 5400 s
1000 km: 7.3 N kg^{-1}; 7300 m s^{-1}; 6300 s
10 000 km: 1.5 N kg^{-1}; 4900 m s^{-1}; 21 000 s
30 000 km: 0.30 N kg^{-1}; 3300 m s^{-1}; 69 000 s

44 $g/4$; $F/4$; $2c$; $v/\sqrt{2}$; $T\sqrt{8}$

45 **a** 5400 s; 7.7 × 10^3 m s^{-1}
b Because it is close to the Earth communication is easy; it requires less energy to be placed in orbit; it is better placed to observe the Earth's surface
But it may be affected by a very small amount of air resistance; it does not stay above the same location on Earth throughout its orbit

46 1.5 × 10^{11} m

47 **d** Yes; the gradient (r^3/T^2) is constant
e The gradient should equal $\frac{3}{2}$

48 44 days

49 **a** Decreases
b $E_T = -\frac{1}{2}Gm_1m_2/r$: if E_T is decreasing it must be changing to a larger negative number, so that r must be getting smaller
c The satellite will gain kinetic energy as it loses gravitational potential energy; as it goes faster it will encounter greater air resistance

50 The work done against air resistance is transferred to internal energy of the satellite; which is destroyed because it gets so hot that it vaporizes and/or chemically reacts with the air

51 **a** –4.9 × 10^{10} J
b +2.4 × 10^{10} J
c –2.4 × 10^{10} J

10 Thermal physics

1 1.71 × 10^{-2} m^3; = 1.71 × 10^4 cm^3

2 39 °C

3 130 mol

4 **a** 32 g
b 0.60 m^3

5 2.1 × 10^6 Pa

6 **a** 2.51 × 10^5 Pa
b Molecules collide with the walls more frequently

7 **a** 7.56 × 10^6 Pa
b On average the molecules are travelling faster; collide with the walls more frequently; and with greater force
c Helium may not behave like an ideal gas at low temperature and high pressure

8 1.7 m^3

9 8/3

10 307 °C

11 When the gas is burned, hot gases enter the balloon and displace cooler, denser gas out of the bottom; in this way the overall weight of the balloon can be adjusted to be less than the upthrust

12 Internal energy is the total of all the kinetic and potential energies of the particles in a substance
Temperature is a measure of the average kinetic energy of the particles
Thermal energy is the non-mechanical transfer of energy from hotter to colder

13 The most work is done in an isobaric change; the least work is done during an adiabatic change; this is shown by the different sizes of the areas under the respective P–V diagrams

14 70 J

15 When molecules collide with the inwards moving surface they gain kinetic energy

16 **a** Work was done by the gas
b 3000 J

17 **a** Temperature and internal energy are constant; the pressure decreases.
b The process was isothermal

18 For the same increase in volume, an adiabatic expansion finishes at a lower pressure; because the temperature falls

19 **a** 2.33×10^5 Pa
d 3.7×10^5 J

20 AB: Isothermal expansion, $\Delta Q = \Delta W$
BC: Adiabatic expansion, $-\Delta U = \Delta W$
CD: Isothermal compression, $-\Delta Q = -\Delta W$
DA: Adiabatic compression, $\Delta U = -\Delta W$

21 **a** AB
b 0.26 mol
c 650 K
d 620 J; this is the work done by the gas in one cycle

22 **e** During process **d**
f 340 J

23 Internal energy remains the same; entropy increases

24 The molecules spread out and become more disordered; entropy increases

25 It is much easier to mix things up than to separate them; the particles were more ordered before they were mixed

26 Only a very simple process in which no significant amount of energy was transferred to the surroundings during the time shown; perhaps a few swings of a good pendulum

27 First law: energy cannot be created
Second law: energy is always dissipated

Chapter 11 Wave phenomena

1 **a** $58.6\,\text{m s}^{-1}$
b 71.4 Hz
c 0.492 m

2 **a** **i** 0
ii π
d 21.6 cm; 127 Hz
e $27.5\,\text{m s}^{-1}$

3 A stretched string has its own natural frequencies at which it can vibrate freely. If an oscillating force is continually applied to it at one of these frequencies, energy will be transferred into the system and the amplitude will increase

4 **a** The wave speed will increase because of the larger forces in the system
b The fundamental frequency increases because $f = v/\lambda$ and v is higher but λ has not changed
c The oscillating string will accelerate more slowly because it is more massive
d The fundamental frequency decreases because $f = v/\lambda$ and v is smaller but λ has not changed

6 360 Hz

7 $338\,\text{m s}^{-1}$

8 0.94 m

9 **a** 1.49 m
b 342 Hz
c 0.745 m
d To produce the same fundamental frequency, they can be half the length of pipes closed at both ends

10 413 Hz

11 $338\,\text{m s}^{-1}$

12 $31\,\text{m s}^{-1}$

13 **a** The sound will become louder as the train gets closer to P (allowing for a time delay for the sound to reach the observer). A pitch higher than that emitted by the train will be heard, but it will gradually fall as the train approaches (because the component of velocity towards the observer is decreasing). These processes are reversed as the train moves past P
b The pitch and loudness will remain constant

14 59 Hz

15 **a** 3000 Hz
b More easily absorbed and scattered in air; diffract and spread out more; slower speed

16 **a** $260\,\text{m s}^{-1}$; 12.6 km
b Make the plane's surface scatter radiation (rather than reflect it); travel close to the ground

18 Moving away with a speed of $8.45 \times 10^6\,\text{m s}^{-1}$

19 a Ultraviolet
b By fluorescence
c 5.04×10^{-3} rad

20 5.7×10^{-7} m

21 0.085 mm

22 a Like Figure 11.26 – minima should occur at angles of $\pm 7.8 \times 10^{-3}$ and $\pm 15.6 \times 10^{-3}$ rad
b The spacing of the diffraction pattern for blue light should be less

23 a There is less diffraction with bigger lenses; they receive more light
b It will be more difficult for a larger lens to focus all the light in the right places on the image

24 Blue is near the short wavelength end of the visible spectrum and diffracts less

25 A larger pupil at night means that diffraction is reduced, suggesting that resolution improves; but the much lower light intensity will reduce the quality of the image

26 12 km

27 0.15 m

28 a 1.4×10^{14} m
b A line joining the stars is perpendicular to a line joining them to Earth

29 Yes; the angle subtended at the telescope by the writing is 1.6×10^{-5} rad and this is much bigger than $1.22\lambda/b$ (about 5×10^{-7})

30 About 100 km

32 Look through the sunglasses at light reflected from glass or water; if they are Polaroid®, the intensity of the image will change as the glasses are rotated

33 6.7%

34 63°

35 The sky appears blue because blue light is scattered from air molecules; simple scattering, like reflection, can result in polarization

36 a 57°
b 53°
c 37°

38 a $0.41\,\mathrm{W\,m^{-2}}$
b $0\,\mathrm{W\,m^{-2}}$
c $0.090\,\mathrm{W\,m^{-2}}$

Chapter 12 Electromagnetic induction

1 Because it has no free electrons; that can move to cause a charge separation

2 a Into, or out of, the plane of the paper
b Perpendicularly between the poles of the magnet, or along the line of the wire
c It is not in a circuit

3 A coil connected to a galvanometer should be moved quickly close to, or surrounding, the magnet

4 a Use a stronger magnet; move it quicker
b Reverse the motion; reverse the polarity

5 Current flows in the opposite direction

6 The induced emf increases as the speed of the falling magnet increases and also as the magnetic field passing through the coil gets stronger. The emf reverses direction when the magnet leaves the coil. The second peak is higher and quicker than the first because the speed is greater

7 $0.35\,\mathrm{m\,s^{-1}}$

8 0.028 T

9 a 0.19 V
b The vertical component of the Earth's magnetic field is greater
c No; any connecting leads would have the same voltage induced across them

10 1.3×10^{-5} Wb

11 7.2×10^{-4} Wb

12 4.5×10^{-4} Wb

13 a i The pointer on the galvanometer will deflect and then quickly return to zero. At the moment the current in the solenoid is switched on, a changing magnetic field through the loops of wire induces a current. There is no induction when the current in the solenoid is steady
ii As in **i** but the deflection on the meter is in the opposite direction.
b Zero; no deflection

14 Increase the number of turns; move the coils closer together; place an iron core through the coils
Increase the magnitude of the current in P

15 a 8.8×10^{-3} Wb
b 3.0×10^{-2} V

16 The amplitude and frequency would double; the maximum induced emf doubles because the current in A (and the resulting magnetic flux) is changing at twice the rate

17 a 4.6×10^{-4} Wb
b The magnetic field strength at A is neligible compared to the field at B
c 0.12 Wb
d 0.083 V

18 a The plane of the coil should be parallel to the direction of the Earth's magnetic field
b 9.0×10^{-5} V

19 a 760
b The coil and the solenoid should have the same axis

20 a The change of flux linkage
b The areas are equal; the flux linkage when the magnet enters the coil is equal to the flux linkage as the magnet leaves the coil

22 a The oscillations will be approximately simple harmonic
b Alternating voltage with the same frequency as **a**; voltage peaks when the magnet is passing through the middle of an oscillation
c The motion will be damped because kinetic energy of the magnet will be transferred to the induced current

23 Currents are induced in the tube which set up a magnetic field opposing the motion of the falling magnet

24 a The changing magnetic field passing through the loop induces a current which sets up a magnetic field in opposition to the original field; the resulting force on the loop causes it to decelerate
b North pole
c Into the plane of the paper
d The magnetic flux through the loop is not changing
e The directions of current and induced magnetic field are opposite to those when the loop entered the magnetic field
f Transferred to internal energy in the loop because of the current in it

25 a Its movement is parallel to the magnetic field
b Faster movement; stronger field; wind into a coil of many turns; place coil on an iron core

28 a 7 W
b 89 Ω

29 a 325 V
b 13.2 A
c 9.35 A
d 492 W

30 240 V

31 a 469 W
b 1880 W
c Resistor was ohmic and did not overheat

32 a Peak power = 4.0 kW; V_0 = 350 V; I_0 = 11 A
b Similar to Figure 12.27
Period of voltage and current variations = 0.0167 s

33 a 30 turns
b 25 Ω
c 5.8 W
d 0.024 A

34 a Turns ratio = $\frac{22}{1}$
b 0.072 A

35 a 1640
b 0.012 A; 60 W

36 Resistance of the wires; cost; magnetic field strength in the core

37 a 3.1×10^5 W
b 4.9 W

38 More power transferred (for a given current)

39 So that most of the transmission line is at high voltage

40 a The cables will have less weight; are easier to support on pylons
b Unreactive; does not react with air and/or water

41 a 2280
b 4500 V
c No voltage drop along the transmission line; transformer is 100% efficient

42 a Hysteresis effects increase
b Lower rate of change of magnetic flux

43 a 14 400 V
b 1.92×10^5 W
c 13 A
d 7700 W
e 29 W

44 a 6000 km
b The spreading (diffraction) of a wave is greatest when the object (circuit) and the wavelength are about the same size; household appliances are very much smaller than 6000 km

45 a 3.3×10^{-32} J
b The energy carried by all the individual ELF photons would have to be absorbed in a single process; this amount of energy is very, very small and much less than that required to produce known reactions in the body
c Wave–particle duality

47 Higher current and higher frequency both result in a greater rate of change of magnetic flux and induced emf; a longer time results in a greater total amount of energy being transferred

Chapter 13 Quantum physics and nuclear physics

1 1.38×10^{24}

2 2 : 1

3 100

4 a 5×10^{-15} J; 3×10^4 eV

5 a 4.91×10^{14} Hz
b 3.26×10^{-19} J
c 3.54×10^{-19} J
d No
f 5.34×10^{14} Hz

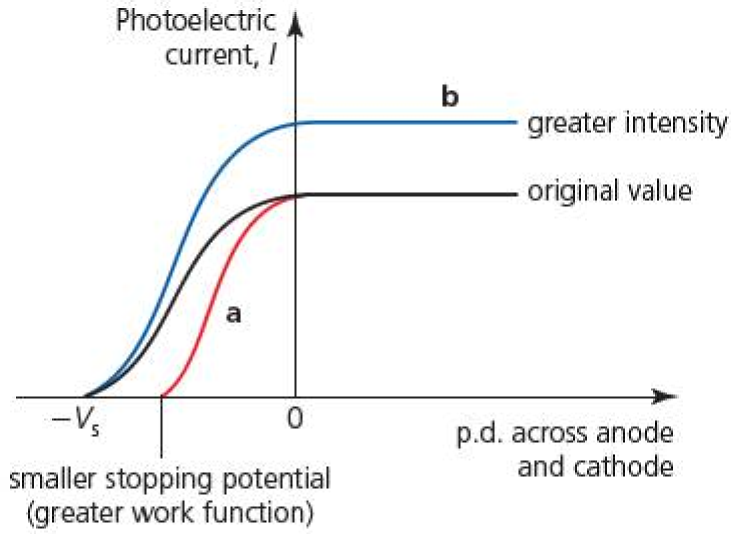
7 2.0×10^{19} J

8 a 3.62×10^{-19} J
b 5.5×10^{-7} m; yellow light
c Red

9 1.44×10^{15} Hz

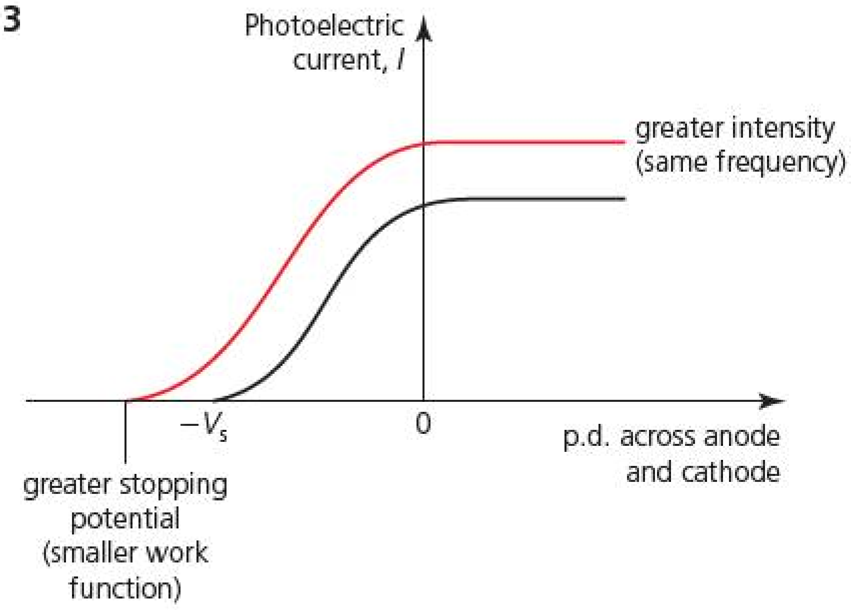
10 3.8×10^{-19} J; 2.4 eV

11

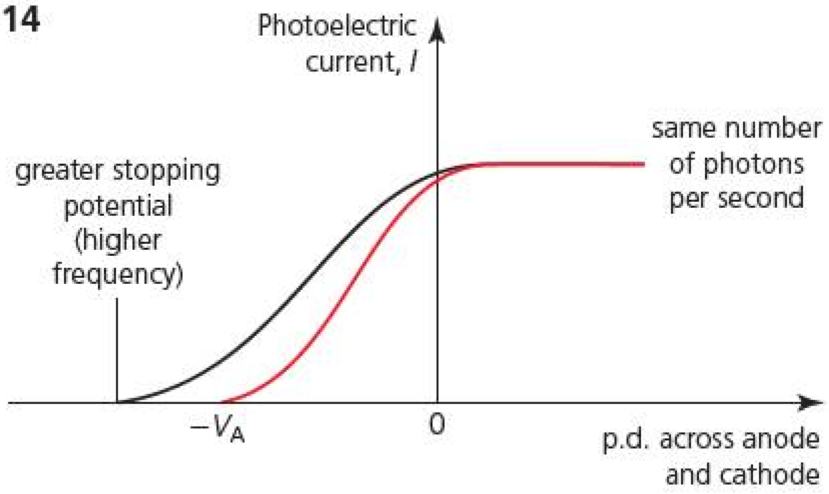


12 6.81×10^{-34} J s

13



14



16 **b** If you double the kinetic energy, the speed increases by the square root of 2. Thus, since the de Broglie wavelength is inversely proportional to speed for a non-relativistic particle, the wavelength will decrease by a factor of square root of 2

c If you double the speed, the de Broglie wavelength will decrease by half

20 3.55×10^{-11} m

21 4.95×10^{2} m s^{-1}

22 9.0×10^{-7} m

23 105 V

24 **a** The wavelength of the de Broglie wave associated with a particle is inversely proportional to the mass of the particle. The mass of an electron is less than that of a proton, so the wavelength of the de Broglie waves associated with the electron would be higher

b Electron, neutron, alpha particle, gold nucleus

25 An airplane has a relatively large mass; hence the wavelength of the de Broglie waves associated with it is too small to be observed and measured

29 **a** 15.3 eV

b $\lambda = 8.13 \times 10^{-8}$ m

c Shorter; there is an inverse relationship between energy and wavelength

30 The energy of an electron at rest outside the atom is taken as zero; when an electron 'falls' into the atom, energy is lost as electromagnetic radiation

32 $n = 1$: 6.0×10^{-18} J
$n = 2$: 2.4×10^{-17} J

34 1.22×10^{-7} m

38 5.8×10^{6} m s^{-1}

39 1.46×10^{-33} m

40 3.7×10^{-19} eV

43 2.0×10^{7} m s^{-1}

46 0.068 m

47 **a** Nuclear energies are significantly larger than electron energies.

b 1.4×10^{-11} m

c Gamma radiation

50 6.36×10^{23} atoms

51 **a** 126 Bq

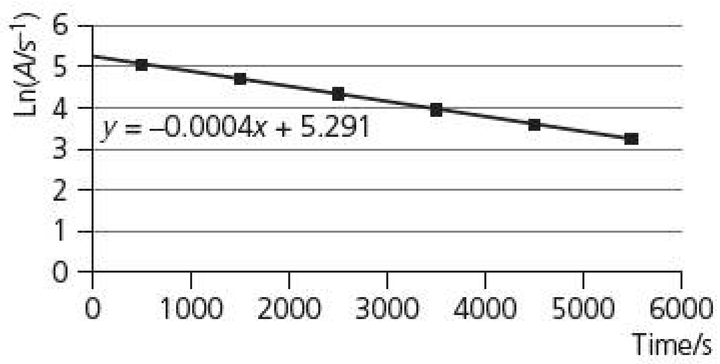
b 805 nuclei

52 **a** 1.201×10^{-4} yr^{-1}

b 0.30

53 1620 years

55 λ is 0.0004; $T_{1/2} = 0.693/0.0004 = 1700$ s.



56 Let T_A and T_B be the half-lives of elements A and B.

$$A \xrightarrow[T_A]{} \tfrac{1}{2}A \xrightarrow[T_A]{} \tfrac{1}{4}A \xrightarrow[T_A]{} \tfrac{1}{8}A$$

$$B \xrightarrow[T_B]{} \tfrac{1}{2}B \xrightarrow[T_B]{} \tfrac{1}{4}B$$

$3T_A = 2T_B$; therefore $\frac{T_A}{T_B} = \frac{2}{3}$

Chapter 14 Digital technology

4 **a** 00001000
b 00001110
c 00010001
d 01000100
e 01111101

5 **a** 6
b 37
c 50
d 63

7 01000011 01000001 01000010

17 To achieve destructive interference, the path difference between light from the top of the pit the bottom of a land should be $\frac{1}{2}\lambda$. Because the light travels there and back, the pit height must be $\frac{1}{4}\lambda = \frac{620\text{ nm}}{4} = 155\,\text{nm}$

18 If v is constant this means that $f \propto \frac{1}{r}$, so if the radius is trebled the frequency is reduced to a third of its original value, namely, 200 rpm

20 **a** 3.6×10^9 pits
b 7.1×10^9 bits
c 890 MB

21 **a** 3.1×10^9 bits
b 1.86×10^{10} bits
c 2217 MB

26 0.25 V

27 120 μC

28 8×10^{-10} V

30 25 μF

31 10 μF

33 80%

34 60%

36 **a** 4.2 mm
b 1.2×10^5 pixels
c 1.2×10^{-3}

37 **a** 1.7×10^4
b 12 500

38 0.125 s

39 240 000 pixels

Answers to examination questions in Chapters 1 to 14

1 Physics and physical measurement

Paper 1

Q1 B

Q2 B

Q3 C

Q4 A

Q5 C

Q6 C

Q7 A

Q8 A

Q9 B

Q10 C

2 Mechanics

Paper 1

Q1 B

Q2 B

Q3 C

Q4 D

Q5 C

Q6 A

Q7 B

Q8 C

Q9 A

Q10 C

Q11 C

Q12 C

Q13 A

Q14 B

Q15 C

Paper 2

Q1 **a** The equation can only be used for motion which has constant acceleration; the force on the bullet changes, so the acceleration cannot be constant

b **i** $6.6 \times 10^4\,\mathrm{m\,s^{-2}}$

c **i** $280\,\mathrm{m\,s^{-1}}$

ii 0.26 MW

d Newton's third law states that for every force there is an equal and opposite force acting on a different body; the forces of the bullet and gun are a pair of Newton's third law forces; the forward force on the bullet is equal to the backwards force on the gun

Q2 **a** Impulse is the product of a force and the time for which it acts ($F\Delta t$) (impulse also equals the change of momentum)

c 3.50 N

Q3 **a** **i** Zero

ii Diagram should show normal reaction forces acting vertically upwards from the ground on both wheels; weight of bicycle acting downwards from its (approximate) centre of mass; force of foot pushing down on pedal (a downwards push of the hands may also be included); the total lengths of the upwards and downwards vector arrows should be about the same

iii The forwards force on the bicycle is equal and opposite to the resistive forces (air resistance/drag/friction); so the resultant force is zero and therefore there is no acceleration

b 320 W

c **i** $0.57\,\mathrm{m\,s^{-2}}$

ii 56 m

iii The total resistive force is *not* constant (as assumed in the calculation); air resistance decreases with speed (the braking force may also change)

3 Thermal physics

Paper 1

Q1 C

Q2 B

Q3 D

Q4 A

Q5 A

Q6 B

Q7 A

Q8 D

Q9 A

Q10 D

Paper 2

Q1 **a** **i** Evaporation occurs at any temperature (and involves a drop in temperature); boiling occurs at a constant temperature

ii Evaporation occurs at the surface of a liquid, so that a larger area increases the overall rate of evaporation; boiling occurs throughout the liquid

b $3.8 \times 10^5\,\mathrm{J\,kg^{-1}}$

c Thermal energy will be transferred to the calorimeter and the surroundings; so less energy is available to boil the liquid (than used in the calculation)

Q2 **a** **i** Internal energy is the total potential energy and kinetic energy of the copper atoms; heating is the transfer of thermal energy to the copper from something hotter

ii $240\,\mathrm{J\,kg^{-1}\,K^{-1}}$

b **i** The molecules gain energy and move faster

ii On average, each molecular collision with the walls exerts a greater force; so for the pressure to remain constant, the frequency of collisions must go down; which means that the volume must increase

4 Oscillations and waves

Paper 1

Q1 C

Q2 C

Q3 B

Q4 B

Q5 A

Q6 C

Q7 C

Q8 D

Q9 C

Paper 2

Q1 **a** **i** A wave which transfers energy
ii Amplitude = 4.0 mm; wavelength = 2.4 cm; frequency = 3.3 Hz; speed = $8.0\,\mathrm{cm\,s^{-1}}$
b **i** 63°

Q2 **a** 1: The graph is a straight line through the origin, indicating that acceleration is proportional to displacement
2: The graph has a negative gradient indicating that acceleration and displacement are in opposite directions
c **i** The wave is progressive because it is transferring energy; the oscillations of the medium are parallel to the direction of energy transfer
ii 0.94 m

5 Electric currents

Paper 1

Q1 B

Q2 B

Q3 A

Q4 D

Q5 D

Q6 D

Q7 D

Q8 C

Q9 D

Q10 B

Paper 2

Q1 **a** The cell should be connected across the ends of the variable resistor (potential divider); the lamp and the ammeter should be connected in series between one end of the variable resistor and its sliding contact; the voltmeter should be connected in parallel across the lamp
b A curve passing through the origin (may be straight at first), with equal increases in voltage producing smaller and smaller increases in current
c 0.24 A
d 1.4×10^{-4} m
e Total resistance of circuit is 3.8 ohms; current through each lamp is 0.2 A, which is less than the 0.24 A needed for normal operation; lamps will be dim

Q2 **a** **ii** 54 Ω
b There are many possibilities; for example, one heater in series with a switch connected to the supply, and the other two heaters in parallel with each other, connected across the supply with a switch in series with the combination

6 Fields and forces

Paper 1

Q1 D

Q2 C

Q3 C

Q4 A

Q5 B

Q6 C

Q7 A

Q8 C

Q9 C

Q10 C

Q11 D

Paper 2

12 **b** **i** $g = GM/R^2$
ii 1.9×10^{27} kg

13 **a** Y: both field strength
K: gravitational constant and Coulomb constant
X: mass and charge
s: distance from mass and distance from charge
b 2.3×10^{39}

7 Atomic and nuclear physics

Paper 1

Q1 D

Q2 C

Q3 B

Q4 C

Q5 D

Q6 D

Q7 A

Q8 C

Q9 C

Q10 D

Paper 2

Q1 a i Natural radioactive decay is the emission of particles; and/or electromagnetic radiation from an unstable nucleus; it is not affected by temperature or the environment
Radioactive decay is a spontaneous random process with a constant probability of decay (per unit time); the activity/ number of unstable nuclei in a sample reduces exponentially

b i Fission

ii U shown near the right-hand end of the line; Sr and Xe shown between U and the peak with Sr to the left of Xe

iii 7.60 MeV (1.22×10^{-12} J)

iv The binding energy of the neutrons is zero because neutrons are separate particles

Q2 a i $^{220}_{86}\text{Rn} \rightarrow {}^{216}_{84}\text{Po} + {}^{4}_{2}\text{He}$

ii 1.01×10^{-12} J

iii $E_K = \frac{1}{2}mv^2$; $1.01 \times 10^{-12} = \frac{1}{2} \times 4 \times 1.66 \times 10^{-27} \times v^2$; $v = 1.74 \times 10^7\,\text{m}\,\text{s}^{-1}$

b i The polonium nucleus moves in the opposite direction to the α-particle

ii By the conservation of momentum, $m_\alpha v_\alpha = m_p v_p$; $v_p = \frac{4}{216} \times 1.74 \times 10^7 = 3.22 \times 10^5\,\text{m}\,\text{s}^{-1}$

iii The polonium nucleus has momentum forward; if the α-particle is emitted along the line of motion, there is no change in direction; but if it is emitted in any other direction, the polonium nucleus will deviate from its original path

c i Decay is a random process; so it is not possible to state when nuclei will decay

8 Energy, power and climate change

Paper 1

Q1 B

Q2 A

Q3 D

Q4 C

Q5 C

Q6 D

Q7 A

Q8 C

Q9 C

Q10 C

Q11 B

Q12 D

Q13 C

Q14 A

Q15 D

Paper 2

Q1 a Coal, oil, gas

b The natural rate of production of the fuels is much lower than the rate of usage; so they will run out

c Using the width of the arrows; efficiency = 5/14 = 0.36 or 36%

d High energy density; readily available; cheap source of electrical energy; used widely by transport, etc.

Q2 a The kinetic and potential energy of the wave is transferred to an oscillating water column (or similar); which is used to drive a turbine connected to an electrical generator

c 5.4 kW m^{-1}

d A sinusoidal waveform will have less water in each peak than a square waveform; so less power will be transferred

Q3 a i Infrared

ii Nitrogen dioxide absorbs infrared radiation radiated from the Earth's surface; which it then re-radiates in random directions; thus reducing the net radiated energy transferred away from the Earth

d i 154 W m^{-2}

ii 402 W m^{-2}

e 2 K

9 Motion in fields

Paper 1

Q1 C

Q2 B

Q3 B

Q4 B

Q5 C

Q6 B

Q7 A

Q8 A

Q9 C

Q10 B

Paper 2

Q1 a The work done per unit mass; in bringing a small test mass from infinity; to that point

b 7.8 N kg^{-1}

c $7.9 \times 10^3\,\text{m}\,\text{s}^{-1}$

Q2 a 1.8 m

b The initial velocity remains horizontal; the curve is steeper than the original because the horizontal distance travelled at any particular height becomes less and less; compared to that without air resistance

10 Thermal physics

Paper 1

Q1 B

Q2 A

Q3 A

Q4 C

Q5 B

Q6 A

Paper 2

Q1 b i 0.18 mol
ii 1.9×10^6 Pa
c i A smooth, upwardly rising curve going from A to a lower volume; followed by a vertical line to a higher pressure; then a steeper curve back to A
ii Change **b**
iii The work done is equal to the area; within the cycle drawn on the diagram

11 Wave phenomena

Paper 1

Q1 A

Q2 D

Q3 D

Q4 C

Q5 B

Q6 C

Q7 C

Q8 B

Q9 C

Q10 B

Paper 2

Q1 a The wavefronts should all be circular, with their centres moving between S and P
b The diagram should show that the wavelength moving towards P is shorter than that moving towards Q; so that P hears a higher frequency
c i Doppler effect
ii Change in frequency of the sound from a car moving past an observer

Q2 a Polarized light has all its electric field vectors oscillating in only one direction (or magnetic field vectors); the liquid crystal changes this plane
b i Nothing would be seen because the optical axes of the polarizer and the analyser are at right angles to each other; light passing through P_1 cannot pass through P_2
ii When the liquid crystal is in place (but no p.d. is applied to the electrodes), light is transmitted because the plane of polarization has been rotated; when a p.d. is applied to the electrodes, the resulting electric field across the liquid crystal prevents rotation of the plane of polarization and nothing will be seen; but this only occurs in the shape of the electrode

Q3 a l is equal to the path difference between the rays from the edges of the slit; if $l = \lambda$, these rays will interfere constructively, but rays from the top of the slit and rays from X will interfere destructively because their path difference will be half a wavelength; in a similar way, pairs of rays across the slit width can be paired-off to produce destructive interference; so when $l = \lambda$ the first minimum of the diffraction pattern will be produced
b The waveform is identical to the first, but the central maximum of one coincides with the first minimum of the other
c The smaller the value of λ/b, the better the resolution. Since the radio wavelengths received are relatively large, the width of the receiving aperture (dish) needs to be as large as possible
d Calculations show that they cannot be resolved; the angle subtended by the sources at Earth is smaller than $1.22\lambda/b$

12 Electromagnetic induction

Paper 1

Q1 A

Q2 C

Q3 B

Q4 C

Q5 B

Q6 D

Q7 B

Paper 2

Q1 a Faraday's law states that an induced emf is proportional to the rate of change of flux (linkage); in this example, the alternating current in the cable produces a changing magnetic field around it; so that a (time) changing magnetic flux passes through the coil
b The graph should have the same shape and frequency as the current graph, but show a phase shift of $\pi/2$
c The size of the induced emf is proportional to the rms value of the current *if* the coil is always placed the same distance from the cables; with the same orientation

13 Quantum physics and nuclear physics

Paper 1

Q1 A

Q2 D

Q3 C

Q4 C

Q5 A

Q6 B

Q7 A

Paper 2

Q1 **a** There is a (probability) wave associated with all particles; the wavelength is given by Planck's constant divided by the particle's momentum

b The de Broglie wavelength of the ball is

$l = \frac{6.63 \times 10^{-34}}{0.05 \times 15} \approx 9 \times 10^{-34}\,\text{m}$

This is much smaller than the 0.5 m gap so wave properties will not be observed

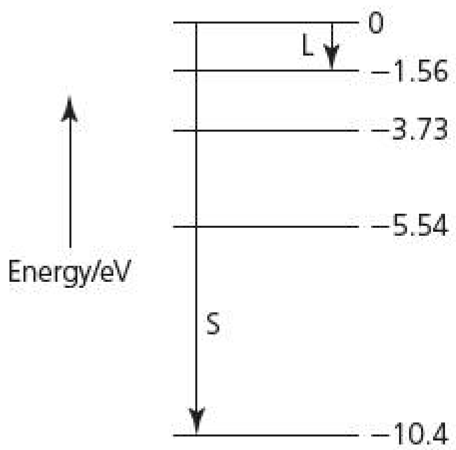
c $\frac{p^2}{2m} = eV$ or $\frac{1}{2}mv^2 = eV$;

$p = \sqrt{2meV}$,

so $\lambda = \frac{h}{\sqrt{2meV}}$

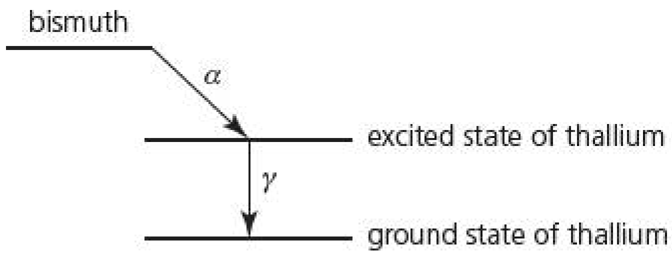
d 1.7×10^{-10} m

Q2 **a** **i, ii**

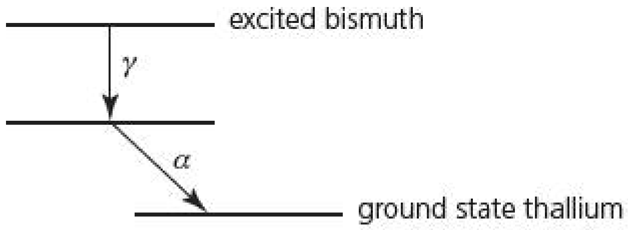


b $E = \frac{hc}{\lambda}$; $\lambda = \frac{6.6 \times 10^{-34} \times 3 \times 10^8}{10.4 \times 1.6 \times 10^{-19}} = 1.20 \times 10^{-7}$ m

c



Or



d $A = A_0 e^{-\lambda t}$; $\frac{1.13}{2.80} = e^{-80\lambda} = 0.404$;

$\lambda = 0.0113$ (min^{-1});

$T_{1/2} = \frac{0.693}{0.0113} \approx 61$ min

Q3 **a** Light consists of photons; the energy of each photon is hf, where h is the Planck constant; a certain amount of energy, called the work function ϕ, is required to remove an electron from the metal surface; if f is less than ϕ/h, no electrons will be emitted

b **i** 1.1×10^{15} Hz

ii $E_K = hf - \phi = Ve$, so slope of graph $= \frac{h}{e}$

Slope = 4.2 (±0.4) × 10^{-15}

giving h = 4.2 (±0.4) × 10^{-15} × 1.6 × 10^{-19}
= 6.7 (±0.4) × 10^{-34} J s

iii $\phi = hf$; using the value of h from **b ii**,

$\phi = 6.7 \times 10^{-34} \times 1.1 \times 10^{15} = 7.4 \times 10^{-19}$ J

Or

From the intercept on (the extended) E_K-axis, $\phi = 4.5(\pm 0.2)$ eV

Q4 **a** **i** 1×10^{-10} m

ii The wavelength is smaller/frequency is higher; hence the kinetic energy is greater

b **i** The electron is attracted to the nucleus, so work must be done (on the electron) to move it away from the nucleus; and so electrical potential energy increases as the distance from the nucleus increases

Or:

The potential due to the nucleus is $V = k\frac{Q}{r}$ where Q is the nuclear charge;

because $E_p = -V|e|$ the potential energy of the electron becomes less negative as the electron moves away

ii The total energy of the electron is constant, so the kinetic energy decreases as the distance increases (because the potential energy increases)

iii Because the kinetic energy decreases, the wavelength must increase as the electron moves away from the nucleus

14 Digital technology

Paper 1

Q1 C

Q2 D

Q3 D

Q4 C

Q5 D

Q6 A

Q7 A

Q8 B

Q9 C

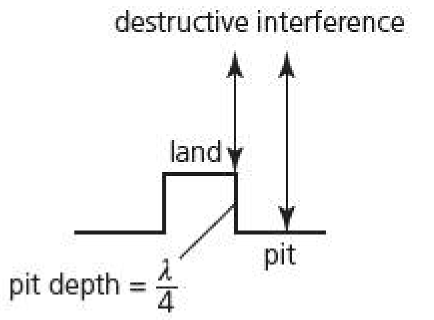
Paper 2 and 3

Q1 a i 12 V

ii 1100

iii The least significant bit has value 0; it is the last digit on the right of the number; and has the least effect on the value. The value 0 shows that the voltage is an even denary number

b



Light is reflected from the land surface and from the edge of pit. If there is just the right pit depth $\left(\frac{1}{4}\lambda\right)$, light reflected from the edge of a pit land and have a path difference of $\frac{1}{2}\lambda$ and interfere destructively; such that there is zero intensity at the detector; giving a digital 0 signal

Q2 a The indentations will be resolved if their images are two pixels apart

$$\text{magnification} = \frac{\text{image length}}{\text{object length}}$$

$$= \left(\sqrt{\frac{1.0 \times 10^{-3}}{2.5 \times 10^{-3}}}\right) = 0.20$$

length of a pixel = $\sqrt{2.3 \times 10^{10}} = 1.5 \times 10^{-5}$ m

Separation of 5.0×10^{-4} m on the object corresponds to a separation of $5.0 \times 10^{-4}\,\text{m} \times 0.2 = 1.0 \times 10^{-4}$ m on the CCD;

this is equal to $\frac{1.0 \times 10^{-4}}{1.5 \times 10^{-5}} = 6.7$ pixels, so the indentations are resolved

b i The ratio of the number electrons emitted to the number of photons incident on the pixel

ii Number of electrons emitted
$= 0.8 \times 5.5 \times 10^{4}$
Amount of charge generated on each pixel,
$Q = 0.8 \times 5.5 \times 10^{4} \times 1.6 \times 10^{-19}$
$= 7.0 \times 10^{-15}$ C

$V = \frac{Q}{C}$ and $C = 40$ pF

Therefore $V = \frac{7.0 \times 10^{-15}}{4.0 \times 10^{-11}} = 0.18$ mV

c The variation of p.d. across the collection area is a 'map' of the image of the object on the collection area; each p.d. can be converted to a digital signal; these digital signals can be converted to an image on an LCD/screen

Glossary

Standard level

The glossary contains key words and terms from the IB Physics Diploma course. This section takes words from Chapters 1–8. Entries that are **IB syllabus-required definitions and statements** are coloured in blue.

A

Absolute zero Temperature at which (almost) all molecular motion has stopped.

Absorption spectrum A spectrum of dark lines across a continuous spectrum produced when light passes through a gas at low pressure.

Acceleration, *a* Rate of change of velocity with time, $a = \Delta v/\Delta t$ (unit: $m\,s^{-2}$). Can be determined from the gradient of a velocity–time graph. Acceleration is a vector quantity.

Accuracy A single measurement is described as accurate if it is close to the correct result. A series of measurements of the same quantity can be described as accurate if their mean is close to the correct result. Accurate results have small systematic errors.

Air resistance Resistive force opposing the motion of an object through air. It depends on the cross-sectional area and shape of the object, and it is approximately proportional to the object's speed squared. Sometimes called drag.

Albedo Defined as the total scattered or reflected power/total incident power (on part of a planet's surface). Albedo depends on the nature of the surface, cloud cover and inclination of the radiation to the surface. The Earth's annual mean albedo is 30%.

Alpha particle A fast moving helium-4 nucleus emitted by a radioactive nucleus, consisting of two protons and two neutrons tightly bound together.

Alternating current (ac) A flow of electric charge which changes direction periodically.

Ammeter An instrument which measures electric current.

Ampere, A SI unit of electric current (fundamental). One ampere is that current which, when flowing in two parallel wires 1 m apart in a vacuum, produces forces of 2×10^{-7} N per metre. $1\,A = 1\,C\,s^{-1}$.

Amplitude, x_0 Defined as the maximum displacement of a wave or oscillation.

Analogue instrument Measuring instrument which has a continuous scale.

Angular frequency, ω Change of angle/change of time, $\omega = \Delta\theta/\Delta t = 2\pi/T = 2\pi f$

Anode An electrode into which (conventional) current flows.

Anthropogenic Caused by humans.

Antineutrino Low mass and very weakly interacting particle emitted during beta-negative decay.

Artificial transmutation An artificially induced nuclear reaction caused by the bombardment of a nucleus with subatomic particles or small nuclei.

Atomic energy level One of the stable states of a bound electron. According to quantum theory, only certain discrete energy levels are possible.

Avogadro's constant The number of particles in one mole of a substance. Defined as the number of carbon atoms in 12 g of carbon-12.

B

Background radiation Radiation from radioactive materials in rocks, soil and building materials, as well as cosmic radiation from outer space.

Beta particle A high speed electron which is released from the nucleus during beta negative decay.

Binding energy The energy released when a nucleus is formed from its constituent nucleons. Alternatively, it is equal to the work required to completely separate the nucleons. Binding energy is the energy equivalent of the mass defect.

Binding energy per nucleon The binding energy per nucleon for an atomic nucleus equals the binding energy for the nucleus divided by the number of nucleons. It is a measure of the stability of a nucleus.

Black body An idealized object which absorbs all the electromagnetic radiation that falls upon it.

Black-body radiation Radiation emitted from a 'perfect' emitter. The characteristic ranges of different radiations emitted (spectra) at different temperatures are commonly shown in graphs of intensity against wavelength (or frequency).

C

Calibrate Put numbered divisions on a scale.

Carbon (dioxide) capture and storage Removal of carbon dioxide from the gases released when fossil fuels are burned at a power station and its subsequent storage, often in the form of solid compounds.

Carbon fixation Removal of carbon dioxide from the air and the formation of solid carbon compounds, principally by the process of photosynthesis.

Cathode An electrode out of which (conventional) current flows.

Cell (electric) Device which transfers chemical energy to the energy carried by an electric current.

Celsius (scale of temperature) Temperature scale based on the melting point (0°C) and boiling point (100°C) of water.

Centre of mass Average position of all the mass of an object. The mass of an object is distributed evenly either side of any line through its centre of mass.

Centripetal acceleration and force Any object moving in a circular path has an acceleration towards the centre of the circle, called its centripetal acceleration. It can be calculated from the expression $a = v^2/r$. The force producing this acceleration is called a centripetal force.

Chain reaction (nuclear) If, on average, one (or more) of the neutrons produced in a nuclear fission process causes further fission, the process will not stop and will become a self-sustaining chain reaction.

Charge Fundamental property of some sub-atomic particles which makes them experience electric forces when they interact with other charges. Charge may be positive or negative.

Coherent waves Waves that have the same frequency and a constant phase difference.

Collision Two (or more) objects coming together and exerting forces on each other for a short length of time. In an elastic collision the total kinetic energy before and after the collision is the same. In an inelastic collision the total kinetic energy is reduced after the collision.

Combined heat and power (CHP) Power station in which some of the thermal energy (that would otherwise spread into the environment) is transferred (using hot water) to keep local homes and offices, etc., warm.

Combustion (of fuels) Burning. Release of thermal energy from a chemical reaction between the fuel and oxygen in the air.

Components of a vector For convenience, a single vector quantity can be considered as having two parts (components), usually perpendicular to each other. The combined effect of these components is exactly the same as the single vector. *See also* resolve.

Compression (force) Force that tries to squash an object or material.

Compressions (in a longitudinal wave) Places where there are increases in the density and pressure of a medium as a wave passes through it.

Condensation Change from a gas (vapour) to a liquid.

Conduction (heat) Passage of thermal energy through substances as energy is transferred from molecule to molecule.

Conductor (electrical) A material through which an electric current can flow, because it contains significant numbers of mobile charges.

Conservation laws
Charge: the total charge in any isolated system remains constant.
Energy: the total energy in any isolated system remains constant. Energy cannot be created or destroyed.
Momentum: the total momentum in any isolated system remains constant. The total (linear) momentum of a system is constant provided that no external forces are acting on it.

Contact (normal) forces Forces that occur because surfaces are touching each other. Contact forces are perpendicular (normal) to the surfaces.

Continuous spectrum A spectrum in which all possible wavelengths are present. A continuous visible spectrum shows a smooth and uninterrupted change from one colour to another.

Control rods Used for adjusting the rate of fission reactions in nuclear reactors by absorbing more or fewer neutrons.

Controlled and uncontrolled nuclear fission In a nuclear power station the number of neutrons in the reactor core is carefully controlled in order to maintain the rate of the nuclear reactions. In nuclear weapons the number of neutrons is uncontrolled.

Convection Passage of thermal energy through liquids and gases due to the movement of the substance because of differences in density.

Conventional current The direction of flow of a direct current is always shown as being from the positive terminal of the power source, around the circuit, to the negative terminal. Conventional current is opposite in direction to electron flow.

Coulomb, C The unit of measurement of electric charge. One coulomb of charge passes a point in one second if the current is one amp.

Coulomb constant, *k* The constant that occurs in the Coulomb's law equation. $k = 8.99 \times 10^9\,\text{N m}^2\,\text{C}^{-2}$. $k = 1/(4\pi\varepsilon_0)$, where ε_0 is the electrical permittivity of free space.

Coulomb's law There is an electric force between two point charges, q_1 and q_2 given by $F = kq_1q_2/r^2$, where r is the distance between them and k is the Coulomb constant. This is sometimes called the Coulomb interaction. The law may also be applied to charged spheres that are relatively far apart.

Crest Highest part of a transverse mechanical wave.

Critical damping When an oscillating system quickly returns to its equilibrium position without oscillating.

Critical mass The minimum mass needed for a self-sustaining nuclear chain reaction.

Current (electric) A flow of electric charge. Current is defined in terms of the force per unit length between two parallel current-carrying conductors. *See also* ampere. Also equal to the amount of charge passing a point in unit time: $I = \Delta q/\Delta t$

D

Damping When resistive forces act on an oscillating system, dissipating energy and reducing amplitude. Damping may be described according to its degree: heavy damping, light damping, or critical damping.

Decay series A series of nuclides linked in a chain by radioactive decay. Each nuclide in the chain decays to the next until a stable nuclide is reached.

Deceleration Negative acceleration. Reduction in the magnitude of a velocity (speed).

Deforestation Removal of large numbers of trees from an area of land, for example clearance of trees in order to plant food crops.

Degraded energy Energy that has spread into the surroundings and cannot be recovered to do useful work.

Derived units Units of measurement that are defined in terms of other units.

Diffraction The spreading of waves as they pass obstacles or travel through apertures (gaps).

Diffusion Movement of randomly moving particles from a place of high concentration to places of lower concentration.

Digital instrument Measuring instrument that displays the measurement only as digits (numbers).

Dipole Two close electric charges (or magnetic poles) of equal magnitude but of opposite sign (or polarity).

Direct current (dc) A flow of electric charge which is always in the same direction.

Disperse (light) Separate into different colours (e.g. to form a spectrum).

Displacement, *x* Defined as the distance from a reference point in a specified direction; a vector quantity.

Dissipate Spread out so that it cannot be recovered.

E

Earth (ground) connection A good conductor connected between a point on a piece of apparatus and the ground. This may be part of a safety measure, or to ensure that the point is kept at 0 V.

Edge effect The strength of an electric field between parallel plates is reduced at the edges of the plates.

Efficiency Defined as the ratio of the useful energy (or power) output from a device to the total energy (or power) input; often expressed as a percentage. Typical (approximate) efficiencies of power stations: gas 45%, oil 40%, coal 35%.

Einstein mass–energy equivalence relationship In nuclear reactions the equation $E = \Delta mc^2$ is used to calculate the amount of energy released, where Δm represents the mass defect and c represents the speed of light.

Einstein model of light The Einstein model consists of three postulates: electromagnetic radiation travels as photons; each photon has energy, $E = hf$; the superposition of a sufficiently large number of photons has the characteristics of a continuous electromagnetic wave.

Elastic strain potential energy Form of energy that is stored in a material which has been deformed elastically. The energy is transferred when the material returns to its original shape.

Electric field strength, *E* Defined as the electric force per unit charge that would be experienced by a small test charge placed at that point. $E = F/q$ (unit: N C^{-1}).

Electric forces Fundamental forces that act across space between all charges. The forces between opposite charges are attractive. The forces between similar charges are repulsive.

Electrode Conductor used to make an electrical connection to a non-metallic part of a circuit.

Electromagnetic spectrum Electromagnetic waves of all possible different frequencies, displayed in order. In order of increasing frequency: radio waves, microwaves, infrared, visible light, ultraviolet, X-rays, gamma rays. The visible spectrum in order of increasing frequency: red, orange, yellow, green, blue, indigo, violet.

Electromagnetic waves Combined electric and magnetic waves that can travel across free space at $3.0 \times 10^8 \text{ m s}^{-1}$. *See also* electromagnetic spectrum.

Electromotive force (emf), ε Defined as the total energy transferred in a source of electrical energy per unit charge passing through it.

Electron Sub-atomic particle with a negative charge (-1.6×10^{-19} C) and mass of 9.110×10^{-31} kg; present in all atoms and located in energy levels outside the nucleus

Electron transition The movement of an electron between energy levels. A photon is emitted when an electron makes a transition to a lower energy level. The energy of the photon is equal to the difference in energy of levels involved.

Electronvolt, eV Defined as the amount of energy transferred when an electron is accelerated by a p.d. of 1 V. $1 \text{ eV} = 1.6 \times 10^{-19}$ J

Emission spectrum Line spectrum associated with emission of electromagnetic radiation by atoms or molecules, resulting from electron transitions from higher to lower energy states.

Emissivity, *e* The power radiated by an object divided by the power radiated from a black-body of the same surface area and temperature.

Energy density The energy transferred from a unit mass of fuel. It is measured in joules per kilogram, J kg^{-1}.

Equations of motion Equations which can be used to make calculations about objects that are moving with uniform acceleration.

$v = u + at$

$s = (u+v)t/2$

$v^2 = u^2 + 2as$

$s = ut + \frac{1}{2}at^2$

Equilibrium (translational) An object is in translational equilibrium if there is no resultant force acting on it, so that it remains at rest or continues to move with a constant velocity.

Error When a measurement is not exactly the same as the correct value.

Error bars Vertical and horizontal lines drawn through each data point on a graph to represent the uncertainties in the two values.

Evaporation The change from a liquid to a gas at a temperature below the boiling point of the liquid.

Excitation The addition of energy to a system, changing it from its ground state to an excited state. Excitation of a nucleus, an atom, or a molecule can result from absorption of photons or from inelastic collisions with other particles.

Expansion coefficient, γ Liquids expand in all three dimensions when heated, so we usually refer to a volume expansion, rather than a linear expansion. The coefficient of volume expansion is defined as the fractional change in volume, $\Delta V/V_0$, per unit change in temperature, ΔT. Its unit is K^{-1} (or $°C^{-1}$).
$\gamma = \Delta V/V_0 \Delta T$

Expansion of water (anomalous) The expansion of water is unusual (anomalous): when the temperature rises between 0 °C and 4 °C it contracts (rather than expands). When heated above 4 °C it expands, but not linearly with temperature.

F

Feedback When the past results of a continuing process are able to change future events in the same process. If results cause increased effects, it is described as positive feedback. If results cause reduced effects, it is called negative feedback.

Field (gravitational, electric or magnetic) A field is a region of space in which a mass (or a charge, or a current) experiences a force due to the presence of one or more other masses (charges, or currents – moving charges).

Field lines and patterns Fields can be represented in drawings by a pattern of lines. Each line shows the direction of force on a mass (in a gravitational field), or on a positive charge (in an electric field), or of the north pole (in a magnetic filed). A field is strongest where the lines are closest together. *See also* uniform field and radial field.

Filament lamp Lamp which emits light from a very hot metal wire. Also called incandescent lamp.

Fission fragments The nuclei produced by nuclear fission.

Fluorescent lamp Lamp which produces light by passing electricity through mercury vapour at low pressure.

Force constant Ratio of force to extension for a stretched material or spring. (Sometimes called the spring constant.)

Forced oscillations Oscillations of a system produced by an external periodic force.

Fossil fuels Naturally occurring fuels that have been produced by the effects of high pressure and temperature on dead organisms (in the absence of oxygen) over a period of millions of years. Coal, oil and natural gas are all fossil fuels.

Frame of reference (for motion) Location to which observations and measurements of motion are compared. For example, the speed of a car may be $10\,m\,s^{-1}$ compared to the Earth's surface.

Free electrons Electrons (most commonly in metals) which are not attached to individual atoms. Also called delocalized electrons. They provide the mobile charges that are needed for an electric current to flow in solid conductors.

Free-body diagram Diagram showing all the forces acting on a single object, and no others.

Free fall Motion through the air under the effects of gravity but without air resistance. In common use free fall can also mean falling towards Earth without an open parachute. Acceleration of free fall, $g = 9.81\,m\,s^{-2}$ for objects falling in a vacuum near the Earth's surface.

Free space Place where there is no air (or other matter). Also called a vacuum.

Frequency, f Defined as the number of oscillations per unit time or number of waves passing a point per unit time (usually per second). $f = 1/T$.

Frequency (driving/forcing) The frequency of an external periodic force acting on a system.

Frequency (natural) The frequency at which a system oscillates when it is disturbed and then left to oscillate on its own.

Frequency response graph Graph used to show how the amplitude of a system's oscillations responds to different forcing frequencies.

Friction Resistive forces opposing (relative) motion, particularly between solid surfaces. Static friction prevents movement, whereas dynamic friction occurs when there is motion.

Fuel A store of energy (chemical or nuclear) which can be transferred to do useful work (for example, generate electricity or power vehicles).

Fuel enrichment Increasing the percentage of ^{235}U in uranium fuel in order to make it of use in a nuclear power station or for a nuclear weapon.

Fundamental units Units of measurement that are not defined as combinations of other units.

Fusion (thermal) Melting.

G

Gamma radiation Electromagnetic radiation (photons) emitted during radioactive decay and having an extremely short wavelength.

Gas laws Laws of physics relating the temperature, pressure and volume of a fixed mass of an ideal gas: Boyle's law, Charles' law and the pressure law.

Geiger–Marsden experiment The scattering of alpha particles by a thin sheet of gold foil which demonstrated that most particles passed through the foil completely undeflected, while a few were deflected at extremely large angles. This demonstrated that atoms consist of mostly empty space with a very dense core (the nucleus).

Geiger–Muller tube The Geiger–Muller tube is used to detect the radiation from a radioactive sample. It is a type of ionization chamber that detects alpha, beta or gamma radiation. The output of the tube can be connected either to a scaler that gives an accumulated total of counts or to a ratemeter.

Global warming Increasing average temperatures of the Earth's surface, atmosphere and oceans.

Graviational field strength, g Defined as the gravitational force per unit mass that would be experienced by a small test mass placed at that point. $g = F/m$ (unit: N kg^{-1})

Gravitational forces Fundamental attractive forces that act across space between all masses. *See also* weight.

Gravitational potential energy, E_p Energy that a mass has because of its position in a gravitational field. Changes in gravitational potential energy in a uniform field can be calculated from $\Delta E_p = mg\Delta h$.

Greenhouse effect The natural effect that a planet's atmosphere has on reducing the amount of radiation emitted into space, resulting in a warmer planet (warmer than it would be without an atmosphere).

Greenhouse effect (enhanced) The reduction in radiation emitted into space from Earth due to an increasing concentration of greenhouse gases in the atmosphere (especially carbon dioxide) caused by human activities; believed by most scientists to be the cause of global warming.

Greenhouse gases Gases which absorb and emit infrared radiation and thereby affect the temperature of the Earth. The principal greenhouse gases are water vapour, carbon dioxide, methane and nitrous oxide. Atmospheric concentrations of the last three of these have been increasing significantly in recent years.

Ground state The lowest energy state of an atom.

H

Half-life (radioactive) The time taken for the activity or count rate from a source, or the number of radioactive atoms, to be reduced by half; the half-life is constant for a particular radioisotope.

Heat exchanger Equipment designed to efficiently transfer thermal energy from one place to another.

Hertz, Hz Unit of measurement of frequency. 1 Hz = 1 oscillation per second.

Hybrid vehicle Vehicle with two power systems that can be used to provide the energy for motion.

Hydroelectric power (HEP) The generation of electrical power from falling water.

I

I–V characteristic Graph of current–p.d. representing the basic behaviour of an electrical component.

Ideal gas The kinetic model of an ideal gas makes the following assumptions. **i** The molecules are identical. **ii** The molecules are point masses with negligible size or volume. **iii** The molecules are in completely random motion. **iv** There are negligible forces between the molecules, except when they collide. **v** All collisions are elastic, that is, the total kinetic energy of the molecules remains constant.

Ideal meters Meters which have absolutely no effect on the electrical circuits in which they are used.

Impact Collision involving relatively large forces over a short time. The effect of such an impact may be greater than from the same impulse (Ft) delivered by a smaller force over a longer time.

Impulse Defined as the product of force and the time for which the force acts. It is equal to the change of momentum. Unit: N s.

Incident wave Wave (or ray) arriving at a boundary.

Inclined plane Flat surface at an angle to the horizontal (but not perpendicular). A simple device which can be used to reduce the force needed to raise a load, sometimes called a ramp.

Insulator (electrical) A non-conductor. A material through which a (significant) electric current cannot flow, because it does not contains many mobile charges. *See also* conductor.

Insulator (thermal) Material used to reduce the flow of thermal energy into or out of an object.

Intensity, I Defined as wave power/area: $I = P/A$. The intensity of a wave is proportional to its amplitude squared, $I \propto x_0^2$.

Interference Effect that may be produced when similar waves meet; most important for waves of the same frequency and similar amplitude. Waves arriving in phase will interfere constructively. Waves completely out of phase will interfere destructively. *See also* path difference.

Internal energy Total potential energy and random kinetic energy of the molecules of a substance.

Internal resistance, r Sources of electrical energy, for example, batteries, are not perfect conductors. Therefore, they have resistance in themselves, which we call internal resistance.

Ionization The process by which an atom gains or loses one or more electrons, thereby becoming an ion.

Ionizing radiation Radiation with enough energy to cause ionization.

Isothermal Occurring at constant temperature.

Isotope Two or more atoms of the same element with different numbers of neutrons (and therefore different masses).

J

Joule, J Derived SI unit of work and energy. 1 J = 1 N m.

K

Kelvin scale of temperature Temperature scale based on absolute zero (0 K) and the melting point (273 K) of water. The kelvin, K, is the fundamental SI unit of temperature. $T/K = t/°C + 273$.

Kilogram, kg SI unit of mass (fundamental).

Kinetic energy, E_K Energy of moving masses, calculated from $\frac{1}{2}mv^2$.

Kyoto protocol An international agreement aimed at stabilizing the amount of greenhouse gases in the atmosphere and reducing the effects of global warming (now agreed by about 190 countries). It set some targets for the years up to 2012 and has recently been extended.

L

Latent heat Thermal energy that is transferred during any change of physical phase. *See* specific latent heats of fusion and vaporization.

Left-hand rule (Fleming's) Rule for predicting the direction of the magnetic force on moving charges, or a current in a wire.

Light-dependent resistor (LDR) A resistor which has less resistance when placed in light of greater intensity.

Longitudinal wave A wave in which the oscillations are parallel to the direction of transfer of energy, for example sound waves.

Lost volts Term sometimes used to describe the voltage drop that occurs when a source of electrical energy delivers a current to a circuit.

M

Magnetic field strength, *B* Defined as the force acting per unit length on unit current moving across the field at an angle θ: $B = F/(IL \sin \theta)$. Unit: tesla, $1\,T = 1\,N\,A^{-1}\,m^{-1}$.

Magnetic forces Fundamental forces that act across space between all moving charges, currents and/or permanent magnets. The forces are perpendicular to the direction of the current. *See* left-hand rule.

Magnitude Size.

Mass defect The difference in mass between a nucleus and the total masses of its constituent neutrons and protons when separated.

Mechanics Study of motion and the effects of forces on objects.

Medium (of a wave) Substance through which a wave is passing (plural: media).

Meltdown (thermo-nuclear) Common term for the damage to the core and reactor vessel which results from overheating following some kind of accident or malfunction at a nuclear power station.

Metre, m SI unit of length (fundamental).

Moderator Material used in a nuclear reactor to slow down neutrons to low energies.

Molar mass Defined as the mass of a substance which contains one mole of its defining particles.

Mole, mol SI unit of amount of substance (fundamental). Defined as the amount of a substance that contains as many of its defining particles as there are atoms in exactly 12 g of carbon-12.

Momentum (linear), *p* Defined as mass times velocity: $p = mv$ (unit = $kg\,m\,s^{-1}$).

N

Natural gas Naturally occurring mixture of gases (mainly methane) that can be used as a fuel. May be either a fossil fuel, or produced more recently by biological processes.

Neutral Uncharged.

Neutron Neutral sub-atomic particle with a mass of 1.675×10^{-27} kg. The number of neutrons in a nucleus is called the neutron number (N).

Neutron capture Nuclear reaction in which a neutron interacts with a nucleus to form a heavier nucleus.

Newton, N Derived SI unit of force. $1\,N = 1\,kg\,m\,s^{-2}$.

Newton's laws of motion
First law: an object will remain at rest, or continue to move in a straight line at a constant speed, unless a resultant force acts on it.
Second law: acceleration is proportional to resultant force, $F = ma$ or $F = \Delta p/\Delta t$.
Third law: whenever one body exerts a force on another body, the second body exerts exactly same force on the first body, but in the opposite direction.

Newton's universal law of gravitation There is a gravitational force between two point masses, m_1 and m_2 given by $F = Gm_1m_2/r^2$, where r is the distance between them and G is the universal gravitation constant. The law may also be applied to spherical masses which are relatively far apart.

Non-proliferation treaty (nuclear weapons) International agreement to limit the spread of nuclear weapons.

Normal Perpendicular to a surface.

Nuclear equation An equation representing a nuclear reaction. The sum of nucleon numbers (A) on the left-hand side of the nuclear decay equation must equal the sum of the nucleon numbers on the right-hand side of the equation. Similarly with proton numbers (Z).

Nuclear fission (reaction) A nuclear reaction in which a massive nucleus splits into smaller nuclei whose total binding energy is greater than the binding energy of the initial nucleus, with the simultaneous release of energy.

Nuclear fuel The source of energy used in a nuclear power station, most commonly uranium-235 or plutonium-239.

Nuclear fusion (reaction) The combination of two (or more) light nuclei to form a heavier nucleus whose binding energy is greater than the combined binding energies of the initial nuclei, thereby releasing energy.

Nuclear potential energy Energy transferred when there are changes in the nucleus of an atom.

Nuclear transmutation The process of changing one nuclide into another. May be natural or artificial.

Nuclear waste Radioactive materials associated with the production of nuclear power that are no longer useful, and which must be stored safely for a long period of time.

Nucleon A particle in a nucleus, either a neutron or proton. The nucleon number (A) is the total number of protons and neutrons in the nucleus.

Nucleus (of an atom) The central part of an atom containing protons and neutrons (except for hydorogen). A nucleus is described by its atomic number and nucleon number.

Nuclide The general term for a unique atom.

O

Ohm, Ω The derived unit of measurement of electric resistance. $1\,\Omega = 1\,\text{V/A}$

Ohmic (and non-ohmic) behaviour The electrical behaviour of an ohmic component follows Ohm's law.

Ohm's law States that the current in a conductor is proportional to the potential difference across it, provided that the temperature is constant.

Order of magnitude When a value for a quantity is not known precisely, we can give an approximate value by quoting an order of magnitude. This is the estimated value rounded to the nearest power of ten. For example, 400 000 has an order of magnitude of 6 (10^6) because $\log_{10} 400\,000 = 5.602$, which is nearer to 6 than 5.

Oscillating water column (OWC) energy converter A device for transferring the energy of ocean waves to generate electricity.

Oscillation Repetitive motion about a fixed point.

P

Parallax error Error of measurement that occurs when reading a scale from the wrong position.

Parallel connection Two or more components connected between the same two points, so that they have the same p.d. across them.

Particle beams Streams (flows) of very fast moving particles, most commonly charged particles (electrons, protons or ions), moving across a vacuum. Properties of the individual particles can be investigated by observing the behaviour of the beams in electric and/or magnetic fields.

Path difference The difference in distance of two sources of waves from a particular point. If the path difference between coherent waves is a whole number of wavelengths constructive interference will occur.

Penetrating power The penetrating power of nuclear radiation depends upon the ionizing power of the radiation. The radiation continues to penetrate matter until it has lost all of its energy. The greater the ionization per cm, the less penetrating power it will possess.

Period (time period), T Defined as the time taken for one complete oscillation; the time taken for one complete wave to pass a point.

Permanent magnet Magnetized material which creates a significant and persistent magnetic field around itself. Permanent magnets are made from ferromagnetic materials, like certain kinds of steel. Soft iron cannot be magnetized permanently.

Permittivity of free space, ε_0 Fundamental constant which represents the ability of a vacuum to transfer an electric force and field, $\varepsilon_0 = 8.85 \times 10^{-12}\,\text{C}^2\,\text{N}^{-1}\,\text{m}^{-2}$.

Phase (of matter) A system (substance) in which all the physical and chemical properties are uniform.

Phase (oscillations) Oscillators are in phase if they have the same frequency and they are doing exactly the same thing at the same time.

Phase difference When oscillators which have the same frequency are out of phase with each other, the difference between them is defined by the angle (usually in radians) between the oscillations. Phase difference can be between 0 and 2π radians.

Photon A quantum ('packet') of electromagnetic radiation, with an energy (E) given by $E = hf$. *See* Planck relationship.

Photosynthesis Chemical process that produces plant sugars (chemical energy) from carbon dioxide and water using the radiant energy from the Sun.

Photovoltaic cell Device which converts electromagnetic radiation (mainly light) into electrical energy. Also called a solar cell.

Plane waves Waves with parallel wavefronts, which can be represented by parallel rays.

Planck relationship The frequency of electromagnetic radiation, f, emitted or absorbed when an electron undergoes a transition between two energy states is given by $E = hf$, where h represents Planck's constant and E is the difference in energy levels.

Plasma State of matter which is similar to a gas, but which contains a significant proportion of charged particles (ions).

Point mass or charge Theoretical concept used to simplify the discussion of forces in gravitational and electric fields.

Polarity Separation of opposite electric charges or opposite magnetic poles, which produces uneven effects in a system.

Potential difference (electric), p.d. The electrical potential energy transferred (work done) as a unit charge moves between two points, $V = W/q$. Commonly referred to as voltage.

Potential divider Two resistors used in series with a fixed potential difference across them. When one resistance is changed, the p.d.s across each resistor will change, and this can be used for controlling another part of the circuit.

Potentiometer Variable resistor (with three terminals) used as a potential divider.

Power, *P* Defined as energy transferred/time taken ($P = \Delta E/\Delta t$), or, for mechanical energies: work done/time taken ($P = \Delta W/\Delta t$). Unit: watt. $1\,W = 1\,J\,s^{-1}$.

Power (electrical) The rate of dissipation of energy in a resistance, $P = VI = I^2R = V^2/R$.

Precision A measurement can be described as being precise if a similar result would be obtained if the measurement was repeated. Precise measurements have small random errors.

Prefixes (for units) Used immediately before a unit to represent powers of ten. For example, 'milli-' in front of a unit represents $\times 10^{-3}$, as in millimeter.

Pressure, *P* Defined as force acting normally per unit area, pressure = force/area. The unit is pascal, Pa ($1\,Pa = 1\,N\,m^{-2}$).

Primary (prime) energy source Natural source of energy that has not been converted to another form (for example not to electricity or hydrogen).

Prism A regularly shaped piece of transparent material (such as glass) with flat surfaces which is used to refract and disperse light.

Progressive wave A wave that transfers energy from place to place. Sometimes called a travelling wave.

Propagation (of waves) Transfer of energy by waves.

Proton Sub-atomic particle with a positive charge ($+1.6 \times 10^{-19}\,C$) and mass of $1.673 \times 10^{-27}\,kg$. The number of protons in a nucleus is called the proton number (Z).

Pulse (wave) A progressive wave of short duration.

Pump storage (HEP) Large quantities of water are pumped up to a higher location using excess electrical power. When the water is allowed to fall down again, electricity can be regenerated.

R

Radial field Field that spreads out from a point equally in all directions.

Radian Unit of measurement of angle. There are 2π radians in 360°.

Radiation sickness The condition associated with intense exposure to ionizing radiation.

Radioactive decay (radioactivity) Spontaneous disintegration of an unstable nuclide, accompanied by the emission of ionizing radiation in the form of alpha particles, gamma rays or beta particles.

Radioactive decay curve A graph of the number of undecayed nuclei (or the activity or count rate from a source) against time.

Random errors Measurements which may be bigger or smaller than the correct value and are scattered randomly around that value.

Rarefactions (in a longitudinal wave) Places where there are reductions in the density and pressure of a gas as a wave passes through it.

Raw data Measurements made during an investigation.

Ray A line showing the direction in which a wave is transferring energy. Rays are perpendicular to wavefronts.

Reaction force *See also* Newton's third law: forces always occur in pairs and these forces are sometimes described as action and reaction. For example, your force pressing down on the ground could be described as the action force and the normal contact force pushing up could be described as the reaction force.

Reaction time The time delay between an event occurring and a response. For example, the delay that occurs when using a stopwatch.

Reflection (waves) Change of direction that occurs when waves meet a boundary between two media, such that the waves return into the medium from which they came.

Refraction Change of direction that can occur when a wave changes speed (most commonly when light passes through a boundary between two different media).

Refractive index, *n* Defined as the ratio of the speed of waves in vacuum to the speed of waves in a given medium. Also: n = sine of angle in air / sine of angle in medium.

Regenerative braking Decelerating a vehicle by transferring (some of) its kinetic energy into a form which can be of later use (rather than dissipating the energy into the surroundings). For example, by generating an electric current that charges a battery.

Relative velocity The difference between two velocities (measured in the same frame of reference). To determine a relative velocity, one velocity is subtracted from the other, remembering that velocity is a vector quantity.

Renewable energy sources Sources that will continue to be available for our use for a very long time; that cannot be used up or 'run out' (except in billions of years, when the Sun reaches the end of its lifetime).

Resistance (electrical) Ratio of p.d. across a conductor to the current flowing through it $R = V/I$.

Resistive force Any force which opposes motion, for example, friction, air resistance, drag.

Resistivity, *ρ* Resistance of a specimen of a material which has length of 1 m and cross-sectional area of $1\,m^2$.

Resistor A resistance made to have a specific value or range of values.

Resolve (vector) To express a single vector as components (usually two perpendicular to each other). *See also* components.

Resonance The increase in amplitude that occurs when a system is acted upon by an external periodic force which has the same frequency as the natural frequency of the system.

Restoring force Force acting in the opposite direction to motion, returning an object to its equilibrium position.

Resultant The single vector which has the same effect as the combination of two or more separate vectors.

Resultant force The vector sum of the forces acting on an object, sometimes called the 'unbalanced force' or 'net' force.

Rheostat Variable resistance used to control current.

Ripple tank A tank of water used for investigating wave properties.

Rising sea levels (mean) Mean sea level is a measurement of the average position of the ocean at a particular location (typically half way between the levels of high tide and low tide). Rising mean sea levels is one possible effect of the enhanced greenhouse effect.

Rotational kinetic energy of molecules Kinetic energy of molecules because of their rotation (spin).

S

Sankey (energy flow) diagram Diagram representing the flow of energy through a system. The widths of the arrows are proportional to the amounts of energy (or power). Degraded energy is directed downwards in the diagram.

Scalars Quantities that have only magnitude (no direction).

Scattering Irregular deflection of waves (most commonly light) from their original path by interactions with matter.

Scientific notation Every number is expressed in the following form: $a \times 10^b$, where a is a decimal number larger than 1 and less than 10 and b is an exponent (integer).

Scintillation Burst of light of short duration caused by an individual energetic particle.

Second, s SI unit of time (fundamental).

Semiconductor Material (such as silicon) with a resistivity between that of conductors and insulators. Such materials are essential to modern electronics.

Sensor An electrical component which responds to a change in a physical property with a corresponding change in an electrical property (usually resistance). Also called a transducer.

Series connection Two or more components connected such that there is only one path for the electrical current, which is the same through all the components.

Short circuit An unwanted (usually) electrical connection which provides a low resistance path for an electric current. It may result in damage to the circuit, unless the circuit is protected by a fuse or circuit breaker.

SI system of units International system of units of measurement (from the French 'Système International') which is widely used around the world. It is based on seven fundamental units and the decimal system.

SI format for units For example, the SI unit for momentum is kilogram metre per second and should be written as kg m s^{-1} (not kg m/s).

Significant figures (digits) All the digits used in data to carry meaning, whether they are before or after a decimal point.

Simple harmonic motion (SHM) Defined as oscillations in which the acceleration, a, is proportional to the displacement, x, and in the opposite direction: $a \propto -x$; $a = -\omega^2 x$ (where ω is the angular frequency).

Sinusoidal In the shape of a sine wave (usually equivalent to a cosine wave).

Slow neutrons Low energy neutrons (typically less than 1 eV) that are needed to sustain a chain reaction. They are slowed down in a nuclear reactor by the use of a moderator. They are sometimes called thermal neutrons because they are in approximate thermal equilibrium with their surroundings.

Snell's law (of refraction) Connects the sines of the angles of incidence and refraction to the wave speeds in the two media (or the refractive indices): $n_1/n_2 = \sin\theta_2/\sin\theta_1 = v_2/v_1$

Soft iron Form of iron (pure or nearly pure) which is easily magnetized and demagnetized. Soft iron cores are used in a wide variety of electromagnetic devices.

Solar flare Relatively sudden increase in the brightness of a part of the Sun's surface, with an increase in radiated energy.

Solar heating panel Devices for transferring radiated thermal energy from the Sun to internal energy in water.

Solenoid Long coil of wire with turns which do not overlap (helical). Solenoids are often used because of the strong uniform magnetic fields inside them.

Solubility Measure of the amount of a substance that can be dissolved in a given volume.

Sound Longitudinal waves in air, or other media, that are audible (can be heard).

Specific heat capacity Defined as the amount of energy needed to raise the temperature of 1 kg of a substance by 1 K.

Specific latent heat Defined as the amount of energy needed to melt (fusion) or vaporize (vaporization) 1 kg of a substance at constant temperature.

Spectroscopy The production and analysis of spectra using instruments called spectroscopes or spectrometers.

Spectrum The components of radiation displayed in order of their wavelengths, frequencies or energies (plural: spectra).

Speed, v Average speed is defined as distance travelled/time taken, $v = \Delta s/\Delta t$. Instantaneous speed is determined over a very short period of time, during which it is assumed that the speed does not change. It may also be determined from the gradient of a distance–time graph. Compare speed, a scalar quantity, with velocity.

Stefan–Boltzmann law Equation that can be used to calculate the total power radiated from a surface, $P = e\sigma AT^4$. σ is known as the Stefan–Boltzman constant and e is the emissivity.

Strain (linear) As a result of stress: change in length/original length.

Strain gauge (resistance) A resistor which has greater resistance when strained (made longer).

Streamlined Having a shape that reduces the resistive forces acting on an object which is moving through a fluid (gas or liquid).

Stress (linear) Tensile force/cross-sectional area.

Strong nuclear force The strong nuclear force holds protons and neutrons together. It is a short range attractive force (the range is about 10^{-15} m). For longer distances, the force becomes too small or negligible, but for very small distances it is repulsive, and hence it prevents the nucleus from collapsing.

Superposition (principle of) The resultant of two or more waves arriving at the same point can be determined by the vector addition of their individual displacements.

Surface heat capacity, C_S Defined as the amount of energy needed to raise the temperature of unit area of a surface (of a very large mass) by one kelvin, $C_S = Q/A\Delta T$ (unit: $\text{J m}^{-2}\,\text{K}^{-1}$).

Surroundings Everything apart from the system that is being considering; similar to the 'environment'.

System The object(s) that we are considering.

Systematic errors A reading with a systematic error is always either bigger or smaller than the correct value by the same amount, for example, a zero-offset error.

T

Temperature Determines the direction of thermal energy transfer. It is a measure of the average random kinetic energy of the molecules of a substance.

Tension (force) Force which tries to stretch an object or material.

Terminal speed (velocity) The greatest downwards speed of a falling object which is experiencing resistive forces (for example, air resistance). It occurs when the object's weight is equal in magnitude to the sum of resistive forces.

Tesla, *T* Unit of measurement of magnetic field strength. $1\,\text{T} = 1\,\text{N A}^{-1}\,\text{m}^{-1}$.

Test charge (or mass) Idealized model of a small charge (or mass) placed in a field in order to determine the properties of that field, but without affecting those properties.

Thermal capacity Defined as the amount of energy needed to raise the temperature of something by one kelvin.

Thermal contact Objects can be considered to be in thermal contact if thermal energy (of any kind) can be transferred between them.

Thermal energy (heat) The non-mechanical energy transfer between two or more bodies at different temperatures (from hotter to colder).

Thermal equilibrium All temperatures within a system are constant.

Thermistor (negative temperature coefficient) A resistor which has less resistance when its temperature increases.

Thermostat Component that is used with a heater or cooler to maintain a constant temperature.

Tidal water storage Hydroelectric power generation involving water stored behind artificial dams (barrages) at locations where there is a large difference in water level between the high tide and low tide.

Transmittance The fraction of incident radiation (of a specified frequency) that is transmitted through a sample of a substance.

Transparent Describes a medium which transmits light without scattering.

Transverse wave A wave in which the oscillations are perpendicular to the direction of transfer of energy, for example light waves.

Trough Lowest point of a transverse mechanical wave.

Turbine Device which transfers the energy from a moving fluid (gas or liquid) to do mechanical work and cause (or maintain) rotation.

U

Ultrasonic Relating to frequencies of 'sound' above the range which can be heard by humans (approximately 20 kHz).

Uncertainty (random) The range (±), above and below a stated value, over which we would expect any repeated measurements to fall. Uncertainty may be expressed in absolute, fractional or percentage terms.

Unified atomic mass unit A unit of mass used to express the mass of atoms and molecules equal to one-twelfth of the mass of the nucleus of a carbon-12 atom (at rest and in the ground state).

Uniform field Field of constant strength, represented by parallel field lines.

Universal gravitation constant, G The constant which occurs in the Newton's universal law of gravitation, $G = 6.67 \times 10^{-11}\,\text{N m}^2\,\text{kg}^{-2}$.

V

Vaporization Change from a liquid to a gas by boiling or evaporation.

Variable Quantity that may change during the course of an investigation. It may be measurable (quantitative) or just observable (qualitative). A quantity being deliberately changed is called the independent variable and the measured or observed result of those changes occurs in a dependent variable. Usually all other variables will be kept constant (as far as possible); they are called the controlled variables.

Variable resistor A resistor (usually with three terminals) which can be used to control currents and/or p.d.s in a circuit.

Vector Quantity which has both magnitude and direction.

Velocity, v Defined as rate of change of displacement with time, $v = \Delta s/\Delta t$. Velocity is a vector quantity and can also be considered as speed in a specified direction. Velocity may also be determined from the gradient of a displacement–time graph. If the velocity (speed) of an object changes during a period of time t, the initial velocity (speed) is given the symbol u and the final velocity (speed) is given the symbol, v.

Vibration Mechanical oscillation.

Volt Derived unit of measurement of potential difference, $1\,V = 1\,J\,C^{-1}$.

Voltmeter An instrument used to measure potential difference (voltage).

W

Watt, W Derived SI unit of power, $1\,W = 1\,J\,s^{-1}$.

Wavefront A line connecting adjacent points moving in phase (for example, crests). Wavefronts are one wavelength apart and perpendicular to the rays that represent them.

Wavelength, λ Defined as the shortest distance between two points moving in phase (for example, the distance between adjacent crests).

Wave speed Defined as the speed at which energy is transferred by a wave. $v = f\lambda$.

Weigh Determine the weight of an object. In everyday use the word 'weighing' usually means quoting the result as the equivalent mass: 'my weight is 60 kg' actually means I have the weight of a 60 kg mass.

Weight, W Gravitational force acting on a mass, $W = mg$.

World energy resources Approximate proportions of world energy use: oil 38%, coal 25%, natural gas 23%, nuclear 6%, renewables 8%.

Work, W Work is the energy transfer that occurs when an object is moved with a force. More precisely, work done = force × displacement in the direction of the force: $W = Fs\cos\theta$, where θ is the angle between the direction of movement and the direction of the force.

Z

Zero offset error A measuring instrument has a zero offset error if it records a non-zero reading when it should be zero.

Higher level

The glossary contains key words and terms from the IB Physics Diploma course. This section takes words from Chapters 9–14. Entries that are **IB syllabus-required definitions and statements** are coloured in blue.

A

Activity (radioactivity) The number of disintegrations per second that occur in a radioactive source. The SI unit of activity is the becquerel (Bq). *See* decay law.

Adiabatic Occurring without thermal energy being transferred into or out of the system.

American Standard Code for Information Interchange *See* ASCII code.

Amplifier A device that increases the amplitude (power) of a signal. The resulting signal is a reproduction of the input signal.

Analogue-to-digital converter A device which translates continuous analogue signals into proportional discrete digital signals.

Analogue signal A continuous signal that is proportional to the physical property which created it.

Antinodes The positions in a standing wave where the amplitude is greatest. *See also* nodes.

ASCII code An encoding system for the transfer of digital data, in which all the letters of the alphabet, numbers and other characters are given a 8-bit binary equivalent.

Attenuation The process by which intensity is reduced during transmission of energy through a medium.

B

Bainbridge mass spectrometer A type of mass spectrometer in which a beam of positive ions passes through perpendicular electric and magnetic fields. For particles with velocity equal to $v = E/B$ the forces due to these two fields are equal and opposite, so that they do not experience a resultant force. They then pass through a slit into another magnetic field, moving in a semicircular path and striking a detection plate. Since different isotopes have different masses, they trace out different radii.

Becquerel, Bq SI unit of (radio) activity, equal to one nuclear decay every second.

Beta-positive decay Nuclear decay accompanied by the emission of a positron (β^+ particle) and a neutrino from the nucleus.

Binary numbers Numbers represented using only two digits: binary one (1) and binary zero (0).

Bit The smallest unit of data in computing, with a value of either binary 0 or binary 1.

Bohr model A theory of atomic structure that explains the spectrum of hydrogen atoms. It assumes that the electron orbiting around the nucleus can exist only in certain energy states.

Boundary condition (wave equation) A requirement to be met so that a wave equation can be solved.

Boundary conditions (standing wave system) The conditions at the end of a standing wave system, which will encourage either a node or an antinode at those places.

Brewster's angle, ϕ The angle of incidence for which all reflected light is plane polarized parallel to the reflecting surface.

Brewster's law The tangent of the Brewster angle, ϕ, is equal to the refractive index, n, of the material into which the light is entering: $n = \tan\phi$

Byte A set of eight binary digits.

C

Capacitance, C The ratio of charge on an electrically charged, isolated conductor to the potential difference across it. For a parallel plate capacitor this is equal to the ratio of the electric charge transferred between the plates to the resulting potential difference:

$$\text{capacitance} = \frac{\text{charge stored on one plate}}{\text{potential difference between plates}}$$

Capacitor An electric circuit component used to store small amounts of energy temporarily, often consisting of two metallic plates separated and insulated from each other.

Capacity (data storage) The amount of data that can be stored on a storage device. In a communications system it is the amount of data delivered per second.

Cassette tape A compact case containing magnetic tape that runs between two reels; used for recording or playing in a tape recorder.

Cathode ray oscilloscope (CRO) An instrument for displaying and measuring voltages which change with time.

CD-ROM Compact Disc Read-Only Memory. An optical disc used to store computer data.

Charge-coupled device (CCD) A semiconducting device in which incident light causes the build up of electric charge in individual pixels, so producing an image of the object. The amount of charge is proportional to the intensity of light.

Conjugate quantities A pair of physical variables describing a quantum-mechanical system such that either of them, but not both, can be precisely specified at any given time.

Coulomb scattering An interaction between similarly charged particles in which the Coulomb repulsive force is the dominant interaction.

Cycle (thermodynamic) A series of thermodynamic processes that return the system to its original state (for example, the Carnot cycle).

D

Davisson–Germer experiment An experiment which verified the wave properties of matter by showing that a beam of electrons is diffracted by a crystal at an angle dependent upon the velocity of the electrons.

De Broglie's hypothesis and equation All particles exhibit wave-like properties which can be described by the de Broglie equation: $\lambda = \frac{h}{mv} = \frac{h}{p}$ where λ represents the wavelength, h represents Planck's constant, m represents the mass of a particle, moving at a velocity v and p represents the momentum of the particle.

Decay constant, λ The decay constant is defined as the probability of decay of a nucleus per unit time. The unit is: s^{-1}. The decay constant is linked to the half-life by the following equation: $T_{\frac{1}{2}} = \frac{\ln 2}{\lambda}$.

Decay law The law of radioactive decay states that the number of nuclei that will decay per second ($\Delta N/\Delta T$ = the activity, A, of the source) is proportional to the number of atoms, N, still present that have not yet decayed: $\frac{N}{t} = -\lambda N$, where λ represents a constant, known as the decay constant.

Digital signal A coded signal that can have one or two values (binary 0 or binary 1).

Digital Versatile Disc (DVD) An optical disc used to store computer data. A DVD can contain audio, video or data.

Digitize Convert an analogue signal into digital format.

Discrete energy levels Energy which does not vary continuously but that is restricted to specific values.

Doppler effect When there is relative motion between a source of waves and an observer, the emitted frequency and the received frequency are not the same. Sometimes called a Doppler shift.

E

Eddy currents Circulating currents induced in solid pieces of metal when changing magnetic fields pass through them, for example in the iron core of a transformer.

Electric potential, V The electric potential at a point is defined as the work done per unit charge in bringing a small test charge from infinity to that position, $V = E_p/\text{charge}$. The electric potential at a distance r from a point charge, q, can be determined from $V = -kq/r$.

Electric potential difference, ΔV The electric potential difference, ΔV, between two points is the work done when moving a unit charge between those two points, $\Delta V = \Delta E_p/q$.

Electric potential energy, E_p. Defined as the work done when bringing all the charges of a system to their present positions, assuming that they were originally at infinity. The electric potential energy, E_p stored between two point charges, q_1 and q_2, separated by a distance r can be calculated from $E_p = kq_1q_2/r$.

Electric potential gradient, $\Delta V/\Delta r$ The electric potential difference between two points divided by their separation. The magnitude of the potential gradient is equal to the electric field strength: $E = -\Delta V/\Delta r$.

Electromagnetic induction Production of an emf across a conductor which is experiencing a changing magnetic flux. This may be as a result of moving a conductor through a magnetic field, moving a magnetic field through a conductor, or a time-changing magnetic flux passing from one circuit to another (without the need for any physical motion).

'Electron in a box' model A simplified quantum mechanical system that confines an electron to a one-dimensional region of space. The energy levels of the electron are explained in terms of standing waves that fit into that space.

Entropy A measure of the disorder of a thermodynamic system.

Equation of state for an ideal gas $PV = nRT$

Equipotential line (or surface) Line (or surface) joining points of equal potential. Equipotential lines are always perpendicular to field lines.

Escape speed (velocity) Minimum theoretical speed that an object must be given in order to move to an infinite distance away from a planet (or moon, or star), $v_{esc} = \sqrt{2Gm_1/r}$, where m_1 is the mass of the planet.

Exponential decay Represented by the equation $A = -\Delta N/\Delta T = \lambda N$. This equation has a mathematical solution in terms of an exponent: $N = N_0 e^{-\lambda t}$. N_0 is the number of undecayed nuclei at the start and N is the number remaining after time t.

Extra-low-frequency radiation (ELF) Electromagnetic radiation of frequency between about 1 and 300 Hz; principally that transmitted into the environment by the generation, transmission and use of mains electricity (at 50 Hz or 60 Hz).

F

Farad, F The unit of capacitance in the SI system equal to the capacitance of a capacitor having an equal and opposite charge of one coulomb on each plate and a potential difference of one volt between the plates.

Faraday's law of electromagnetic induction The magnitude of an induced emf is equal to the rate of change of magnetic flux linkage, $\varepsilon = -N\Delta\Phi/\Delta t$. For an explanation of the negative sign, *see* Lenz's law.

Ferromagnetic materials Materials that have excellent magnetic properties (high magnetic permeability). Many contain iron.

First harmonic Another name for the fundamental mode of vibration.

First law of thermodynamics If an amount of thermal energy, $+\Delta Q$, is transferred into a system, then the system will gain internal energy, $+\Delta U$, and/or the system will expand and do work on the surroundings, $+\Delta W$: $\Delta Q = \Delta U + \Delta W$. (This is an application of the principle of conservation of energy.)

Floppy disk A removable data storage disk using a flexible magnetic medium in a plastic case.

Fundamental mode The simplest mode of vibration, with the fewest number of nodes and antinodes.

G

Galvanometer Ammeter that measures very small currents.

Generator (ac) A device containing coils that rotate in a magnetic field, transfering kinetic energy to the energy carried by an alternating electric current.

Geostationary A satellite is described as geostationary if it remains 'above' the same location on the Earth's surface. This can be very useful for communications. Geostationary satellites orbit in the same plane as the equator and with the exactly same period as the Earth's oscillation on its axis (one day).

Gravitational potential, V The gravitational potential at a point is defined as the work done per unit mass in bringing a small test mass from infinity to that position. $V = E_p/\text{mass}$. The gravitational potential at a distance r from a point mass, m, can be determined from $V = -Gm/r$.

Gravitational potential difference, ΔV The gravitational potential difference, ΔV, between two points is the work done when moving a unit mass between those two points.

Gravitational potential energy, E_p Defined as the work done when bringing all the masses of a system to their present positions, assuming that they were originally at infinity. The gravitational potential energy, E_p stored between two point masses, m_1 and m_2, separated by a distance r can be calculated from $E_p = -Gm_1m_2/r$.

Gravitational potential gradient, $\Delta V/\Delta r$ The gravitational potential difference between two points divided by their separation. The magnitude of the potential gradient is equal to the gravitational field strength: $g = -\Delta V/\Delta r$.

H

Hard disk A flat, circular, rigid plate with a magnetizable surface on one or both sides, used to store data.

Heat engine A system that performs useful mechanical work from a transfer of thermal energy.

Heisenberg uncertainty principle A fundamental principle of quantum mechanics which states that it is impossible to measure simultaneously the momentum and the position of a particle with infinite precision, $\Delta x \Delta p \geq h/4\pi$.

The principle also applies to measurements of energy and time, $\Delta E \Delta t \geq h/4\pi$.

I

Irreversible process A process in which entropy increases; all real processes are irreversible.

Isobaric Occurring at constant pressure.

Isochoric Occurring at constant volume. Also called isovolumetric.

K

Kepler's third law For all the planets orbiting the Sun, the average radius cubed is proportional to the period squared (that is, r^3/T^2 is a constant). This law can also be applied to any group of satellites (such as planets and moons) orbiting the same centre.

Kilobyte 1024 bytes. This is approximately 1000 bytes.

L

Land A reflective area of an optical medium that reflects the laser into a sensor to register it as a binary one.

Least-significant bit (LSB) The bit in a binary number that is the least important, or having the least weight. It is the right-most bit of a binary number.

Lenz's law (of electromagnetic induction) The direction of an induced emf is such that it will oppose the change that produced it. This is represented mathematically by the negative sign in the equation representing Faraday's law.

Liquid crystals State of matter with properties between that of a liquid and a solid. Certain kinds of liquid crystal can rotate the plane of polarization of light, dependent on a small p.d. applied across them. Such crystals are widely used in displays and computer screens (LCDs).

Logic gate A device, usually an electrical circuit, that performs a logical operation on one or more input signals (1s or 0s). Logic gates are the building blocks of digital technology.

Long-playing disc (LP) An analogue audio recording pressed in vinyl plastic that rotates at 33.3 revolutions per minute (rpm). The sound is encoded in a spiral groove, starting at the outer edge of the disc.

M

Magnetic flux, Φ Defined as the product of an area, A, and the component of the magnetic field strength perpendicular to that area, $B\cos\theta$, $\Phi = BA\cos\theta$

Magnetic flux density, B The term commonly used at Higher Level for magnetic field strength (*see* magnetic field strength, p. 530)

Magnetic flux linkage Defined as the product of magnetic flux and the number of turns in the circuit. Magnetic flux linkage = $N\Phi$

Magnification (of a CCD) The ratio of the length of the image on the CCD to the length of the object.

Malus' law Used for calculating the intensity of light transmitted by a polarizing filter: $I = I_0\cos^2\theta$. I_0 is incident intensity and θ is the angle between the polarizer axis and the plane of polarization of the light.

Mass spectrometer A device which can measure the masses and relative abundance of gaseous ions.

Matter waves Waves that represent the behaviour of an elementary particle, atom or molecule under certain conditions. *See* de Broglie's hypothesis and equation.

Megabyte A megabyte contains 1048 576 bytes (1024 × 1024 bytes). This is approximately 1 million bytes.

Modes of vibration The different ways in which a standing wave may be set up in a given system.

Most-significant bit (MSB) In a binary number, the MSB is the most important bit. It is the left-most bit of a binary number.

N

Neutrino Low mass and very weakly interacting particle emitted during beta-positive decay. The antiparticle of the antineutrino. The emission of neutrinos and antineutrinos results in a continuous distribution of beta particle energies.

Node The positions in a standing wave where the amplitude is zero. *See also* antinodes.

Noise Random, unwanted electrical disturbances that reduce the integrity of electronic signals. Noise can come from a variety of sources, including heat, other users of the same radio frequency, nearby electrical currents and bad connections.

Nuclear energy level Discrete energy levels within the nucleus of an atom whose properties are governed by the rules of quantum mechanics.

Nuclear transition A change in nuclear energy level which results in the emission of a high energy photon.

O

Optical media (storage) Type of data storage medium that uses light, usually in the form of a laser, to read and/or write data. Most commonly involves patterns on discs, such as DVDs.

Optically active substances Substances which rotate the plane of polarization of light which is passing through them.

Orbit Curved path of a mass revolving around a much larger mass, for example a planet around a star, or a moon around a planet. For simplicity, the paths are often considered to be approximately circular, with gravitation providing the centripetal force.

P

PV diagram A graphical way of representing changes to the state of a gas during a thermodynamic process.

Parabolic trajectory In the absence of air resistance, projectiles move in parabolic trajectories. This is the combination of a constant horizontal component of speed and a constant vertical acceleration.

Peak value (electrical) The maximum value of an alternating current or voltage (compare with rms value).

Pit A succession of pits form a track on a CD. The depth of pits are usually one-quarter of the laser wavelength and reflection from the edge of a pit causes cancellation of the beam by interference.

Pitch The sensation produced by sounds of different frequency.

Pixel Picture element. The smallest element in a CCD (or visual display).

Photoelectric effect Ejection of electrons from a substance by incident electromagnetic radiation, especially by ultraviolet radiation. Sometimes called photo emission. The ejected electrons are called photoelectrons.

Photoelectric equation The maximum kinetic energy of an emitted photoelectron is the difference between the photon's energy and the work function: $KE_{max} = hf - \phi$ or $\frac{1}{2}mv_{max}^2 = hf - \phi$ where ϕ represents the work function, hf represents the energy present in the incident photon and KE_{max} represents the maximum kinetic energy of the ejected photoelectrons.

Plane of polarization The plane in which all oscillations of a plane polarized wave are occurring.

Polarimeter Instrument for measuring the angle of rotation of the plane of polarization that occurs when light is transmitted through an optically active substance.

Polarization (plane) A property of some transverse waves in which the electric field (and magnetic field) oscillations are all in the same plane.

Polarizing filter A filter which transmits light which is polarized only in one plane. A filter used to produce polarized light from unpolarized light is called a polarizer. A polarizing filter which is rotated in order to analyse polarized light is called an analyzer. Crossed filters prevent all light from being transmitted.

Portability Stored data is portable if it can be easily transported on the storage device or medium.

Positron The antiparticle of the electron.

Potential hill The potential in a field where the force exerted on a particle is such as to oppose the movement of the particle.

Primary and secondary coils (transformers) The coils to which the input and output of a transformer are connected.

Probability distribution The square of the absolute value of the wavefunction.

Projectile Object that has been projected (thrown or fired) and then moves through the air (usually) without being powered or controlled in any way.

Pulse amplitude modulation (PAM) The transmission of data by varying the amplitudes (voltage or power levels) of the individual pulses in a regularly timed sequence of electrical pulses.

Q

Quality of data A measure used to define the maximum allowable number of bits in error in a digital communication system, such that if the number of bits rose above this threshold, the quality of data would be unacceptable.

Quantization (atoms) The possible values of a quantity are limited to a discrete set of values by quantum mechanical rules.

Quantization levels (data) The discrete values an analogue signal can take when digitally sampled. With n-bit words used in the sampling, the number of quantization levels is 2^n.

Quantum efficiency (of a CCD) The ratio of the number of electrons emitted to the number of incident photons on a pixel.

Quantum mechanics A branch of physics that describes the behavior of objects of atomic and subatomic size.

R

Range (of projectile) Horizontal distance between the point of release and the point of impact with the ground.

Rating Alternative name for the rms value of an electrical quantity.

Rayleigh's criterion Two point sources can just be resolved if the first minimum of the diffraction pattern of one occurs at the same angle as the central maximum of the other. This means that if the sources are observed through a narrow slit, they will just be resolved if they have an angular separation of $\theta = \lambda/b$. For a circular aperture $\theta = 1.22\lambda/b$.

Rectify Change alternating current (ac) into direct current (dc).

Register (CCD) The area of the CCD which receives the binary information for processing and storage before passing it to the analogue-to-digital converter.

Reproducibility (data) The ability of a computer system or electronic device to produce the same result, given the same input conditions and operating in the same environment.

Reservoir (thermal) Part of the surroundings of a thermodynamic system which is kept at approximately constant temperature and is used to encourage the flow of thermal energy.

Resolution The ability of an instrument (or an eye) to detect separate details.

Resolution (CCD) Two points on an object will be resolved if their images are at least two pixels apart.

Retrieve (data) To recover or make newly available stored digital information from a computer system. The retrieval speed is an important measure of how quickly a device can gain access to stored data.

Reversible process A process in which entropy is constant; an impossible ideal.

Root mean squared (rms) value The effective value of an alternating current (or voltage), also called its rating. It is equal to the value of the direct current (or voltage) which would dissipate power in a resistor at the same rate, $I_{rms} = I_0/\sqrt{2}$ and $V_{rms} = V_0/\sqrt{2}$, where I_0 and V_0 are the peak values.

S

Sampling Digitizing an analogue waveform by measuring its amplitude at precisely timed intervals.

Schrödinger wave equation An equation which mathematically represents the Schrödinger model of the hydrogen atom by describing electrons using wavefunctions. The square of the amplitude of the wavefunction gives the probability of finding the electron at a particular point.

Second law of thermodynamics The overall entropy of the universe is always increasing. This implies that energy cannot spontaneously transfer from a place at low temperature to a place at high temperature.

Secondary waves The propagation of waves in two or three dimensions can be explained by considering that each point on a wavefront is a source of secondary waves.

Signal generator Piece of laboratory electrical equipment used to supply alternating voltages over a wide range of frequencies.

Slip rings and brushes In an ac generator these are used for connecting the rotating coil to the external circuit.

Spectral line Line of a single colour on a spectrum, corresponding to a precise wavelength of light emitted (or absorbed) by an atom, ion (or molecule).

Standing wave The kind of wave which may be formed by two similar travelling waves moving in opposite directions. The most important examples are formed when waves are reflected back upon themselves. The wave pattern does not move and the waves do not transfer energy. Also known as stationary waves.

State of a gas Specified by quoting the pressure, P, temperature, T, and volume, V, of a known amount, n, of gas.

Stopping potential The minimum voltage required to reduce the photoelectric current to zero. $V_s = \frac{h}{e}f - \frac{\phi}{e}$, where V_s represents the stopping potential, h represents Planck's constant, f represents the frequency of the incident photons, ϕ represents the work function and e represents the charge on the electron.

T

Thermodynamics The branch of physics which studies the transfer of thermal energy as it is used to do mechanical work.

Threshold frequency, f_0 The minimum frequency of a photon that can eject an electron from the surface of a metal.

Time-base The control on a CRO that moves the spot horizontally across the screen from left to right at a constant speed, and which can be controlled by the user.

Trajectory Path followed by a projectile.

Transducer Device which converts one form of energy to another. The word is most commonly used with devices that convert to or from changing electrical signals.

Transformer A device that transfers electrical energy from one circuit to another using electromagnetic induction between coils wound on an iron core. Transformers are used widely to transform one alternating voltage to another of different magnitude. Step-up transformers increase voltages; step-down transformers decrease voltages.

Transmission of electrical power Electrical power is sent (transmitted) from power stations to different places around a country along wires (cables) which are commonly called transmission lines. These lines are linked together in an overall system called the transmission grid.

Turns ratio (transformer) The ratio of turns in the primary and secondary coils of a transformer controls the output voltage: $N_P/N_S = V_P/V_S$

U

Universal (molar) gas constant The constant, R, that appears in the equation of state for an ideal gas ($PV = nRT$). $R = 8.31\,\text{J K}^{-1}\,\text{mol}^{-1}$.

W

Wavefunction Mathematical function of space and time which describes the quantum state of a subatomic particle, such as an electron. It is a solution to the Schrödinger wave equation.

Wave–particle duality Matter and light exhibit the behaviours of both waves and particles, depending upon the circumstances of the experiment. However, it is not possible to observe both properties simultaneously.

Weightlessness True weightlessness can only occur if a mass is in a position where there is no gravitational field, or where the field is insignificantly small – such as in 'deep space'. Apparent weightlessness occurs when an object is accelerating freely under the effects of gravity, for example in an orbiting satellite which is also freely accelerating, so that there is no contact between the object and the 'floor'.

Weak nuclear force A short range force acting within the nucleus, responsible for beta decay.

Word The largest number of *bits* the processor (CPU) of an electronic device can process at one time.

Work function, ϕ The minimum amount of energy required to free an electron from the attraction of atoms in a metal's surface is called the work function, ϕ. Since the energy of the incident photons is equal to hf, $hf_0 = \phi$, where f_0 represents the threshold frequency and h represents Planck's constant.

Working substance The substance (usually a gas) used in thermodynamic processes to do useful work.

Acknowledgements

Photo credits

p.1 *l* "Courtesy Zettl Research Group, Lawrence Berkeley National Laboratory and University of California at Berkeley." All rights reserved © 2009, *r* NASA, ESA, HEIC, and The Hubble Heritage Team (STScI/AURA); **p.4** Jacques Brinon/AP Photo/Press Association Images; **p.6** sciencephotos/Alamy; **p.22** © Schlegelmilch/Corbis; **p.25** *l* ESA–AOES-Medialab, *r* The Photos – Fotolia; **p.26** *t* © Photodisc/Getty Images, *b* The Art Archive/Alamy; **p.27** *t* NASA Goddard Space Flight Center; *b* Andy Lyons/Getty Images; **p.33** Markus Schreiber/AP/Press Association Images; **p.34** Edward Kinsman/Science Photo Library; **p.37** NASA Marshall Space Flight Center (NASA-MSFC); **p.39** KeystoneUSA-ZUMA/Rex Features; **p.40** *t* David Talbot, *b* © 1996 R.Strange/Photodisc/Getty Images; **p.42** *t* Christopher Talbot, *b* Tomohiro Ohsumi/Bloomberg via Getty Images; **p.43** *t* tingimage/Alamy, *b* Takeshi Takahara/Science Photo Library; **p.49** © Argus – Fotolia; **p.50** Ian Walton/Getty Images; **p.51** © imagebroker/Alamy; **p.54** Koji Sasahara/AP/Press Association Images; **p.55** RSBPhoto/Alamy; **p.57** *t* © Vasily Smirnov – Fotolia.com, *b* Johner Images/Alamy; **p.58** *t* Tim Graham/Alamy, *b* Yoshikazu Tsuno/AFP/Getty Images; **p.59** ITAR-TASS Photo Agency/Alamy; **p.62** Cameron Spencer/Getty Images; **p.66** Photolink Ltd/Alamy; **p.68** Julija Sapic – Fotolia; **p.71** fStop/Alamy; **p.72** *t* David Gee/Alamy, *b* Andrew Jankunas/Alamy; **p.73** Christopher Talbot; **p.79** magann – Fotolia; **p.81** *r* Alban Egger – Fotolia, *b* Sergio Azenha/Alamy; **p.89** Tony McConnell/Science Photo Library; **p.92** SuperStock/Alamy; **p.94** Anne Trevillion; **p.99** © Laurence Gough – Fotolia.com; **p.101** Bert de Ruiter/Alamy; **p.102** *tl* Siemens AG, *tr* © 2003 by Daniel Tückmantel/Fotolia.com, *b* SSPL/Getty Images; **p.119** natureshots – Fotolia; **p.122** © ELEN – Fotolia.com; **p.132** lkordela – Fotolia; **p.133** *t* Peter Tsai Photography/Alamy, *b* © Imagestate Media (John Foxx); **p.135** *t* Sipa Press/Rex Features, *br* Pavel Svoboda – Fotolia, *bl* © patrickw – Fotolia; **p.137** Laurence Cartwright – Fotolia; **p.139** Konstantin Sutyagin – Fotolia; **p.142** sciencephotos/Alamy; **p.144** INTERFOTO/Alamy; **p.145** Christopher McGowan/Alamy; **p.147** alexandre zveiger – Fotolia; **p.149** *tl* Robert Fried/Alamy, *tr* © Imagestate Media (John Foxx), *c* sandy young/Alamy; **p.153** *t* Art Photo Picture – Fotolia, *b* jarous – Fotolia; **p.154** APTN/AP Photo/Press Association; **p.169** David J. Green – electrical/Alamy; **p.178** Xuejun li – Fotolia; **p.182** *l*, *cr* and *r* Alx – Fotolia, *cl* Dmitry Kolmakov – Fotolia; **p.184** *l* David J. Green – electrical/Alamy, *r* Martyn F. Chillmaid/Science Photo Library; **p. 185** David J. Green – technology/Alamy; **p.186** Trevor Clifford Photography/Science Photo Library; **p.187** bloomua – Fotolia; **p.198** sciencephotos/Alamy; **p.201** Leslie Garland Picture Library/Alamy; **p.202** © Christopher Curry/Alamy; **p.209** *t* Matt Writtle/PA Archive/Press Association Images, *b* yuri4u80 – Fotolia; **p.212** Phil Degginger/Alamy; **p.213** Phil Degginger/Alamy; **p.219** CERN; **p.220** Joshua Strang, USAF, Wikipedia; **p.231** Cesar Reyes; **p.232** Christopher Talbot; **p.243** David Talbot; **p.265** *l* RIA Novosti/Alamy; **p.268** © lifeonwhite.com – Fotolia.com; **p.270** © Ashley Cooper/Alamy; **p.271** U.S. Coast Guard/Science Photo Library; **p.272** © alohaspirit/iStockphoto.com; **p.274** Ria Novosti/Science Photo Library; **p.277** GAMMA/Gamma-Rapho via Getty Images; **p.278** *t* Mere Words/Science Photo Library, *b* David Boily/AFP/Getty Images; **p.279** *bl* Sipa Press/Rex Features, *br* Barbara Sax/AFP/Getty Images; **p.280** EFDA; **p.282** *t* Friedrich Saurer/Science Photo Library, *c* F1online digitale Bildagentur GmbH/Alamy; **p.283** LMR Media/Alamy; **p.286** *t* Mike Goldwater/Alamy; **p.289** *t* Birute Vijeikiene – Fotolia, *c* © seraphic06 – Fotolia.com; **p.292** Oceanlinx Limited; **p.295** Kevin Schafer/Alamy; **p.309** *t* Ian Patrick/Alamy, *b* Li jianshu – Imaginechina/Corbis; **p.311** *t* Paul Springett/Still Pictures/Specialiststock.com, *b* izmostock/Alamy; **p.321** © Dr. Loren Winters/Visuals Unlimited/Corbis; **p.325** *t* Nathaniel S. Butler/NBAE via Getty Images, *b* Peter Parks/AFP/Getty Images; **p.328** Courtesy NSSDC, NASA; **p.329** KeystoneUSA-ZUMA/Rex Features; **p.334** Chun Mei Kin; **p.336** NASA; **p.344** Universal History Archive/Getty Images; **p.347** Jenny Thompson – Fotolia; **p.348** NASA; **p.353** *tl* Railpix – Fotolia, *tr* Deborah Benbrook Photography, *bl* arquiplay77 – Fotolia, *br* Panos – Fotolia; **p.354** *t* David Talbot, *b* NASA Marshall Space Flight Center (NASA-MSFC); **p.356** *t* © Photodisc/Getty Images/World Commerce & Travel 5, *b* ©

Photodisc/Getty Images/Business & Industry 1; **p.363** Stuart Forster/Rex Features; **p.367** David Talbot; **p.368** © Shelley Gazin/Corbis; **p.372** Edward Kinsman/Science Photo Library; **p.374** © 1997 Doug Menuez/Photodisc/Getty Images; **p.378** Steen Jeberg/http://commons.wikimedia.org/wiki/File:Nykirke_orglet.jpg/http://creativecommons.org/licenses/by-sa/2.5/deed.en; **p.382** © Danita Delimont/Alamy; **p.383** © dave stamboulis/Alamy; **p.384** *tl* © Christopher Pasatieri/Reuters/Corbis, *b* Edward Kinsman/Science Photo Library; **p.385** © sciencephotos/Alamy; **p.390** Mark Williamson – Fotolia; **p.391** © Kim Steele/Photodisc/Getty Images; **p.393** *l* Illustrated London News, *r* artSILENSEcom – Fotolia; **p.400** Pasieka/Science Photo Library; **p.401** *t* David Talbot, *b* Deklofenak – Fotolia; **p.417** Lucky Dragon – Fotolia; **p.424** David Talbot; **p.425** Gan Boon Hean; **p.426** Galina Semenko – Fotolia; **p.427** Lester Lefkowitz/Taxi/Getty Images; **p.430** Himmelssturm – Fotolia; **p.431** © Islemount Images/Alamy; **p.446** Andrew Lambert Photography/Science Photo Library; **p.455** Cesar Reyes; **p.480** *l* jamdesign – Fotolia, *r* Susumu Nishinaga/Science Photo Library; **p.481** © Frank Wright/iStockphoto.com; **p.483** Los Angeles County Museum of Art, CA, USA/Giraudon/The Bridgeman Art Library/© ADAGP, Paris and DACS, London 2012; **p.489** David Talbot; **p.496** *l* Christopher Talbot, *b* NASA, ESA and the Hubble SM4 ERO Team.

t = top, *c* = centre, *b* = bottom, *l* = left, *r* = right

Computer hardware and software brand names mentioned in this book are protected by their respective trademarks and are acknowledged.
Every effort has been made to trace all copyright holders, but if any have been inadvertently overlooked the Publishers will be pleased to make the necessary arrangements at the first opportunity.

Artwork credits

p.93 Fig 3.8, **p.151** Fig 4.56, **p.203** Fig 6.19, **p.355** Fig 10.5 and **p.414** Fig 12.17 all Jon Homewood; **p.328** Fig 9.14 Terri Harwood.

Examination questions

Examination questions have been reproduced with kind permission from the International Baccalaureate Organization.

Index

Option

A

Chapter 15
Sight and wave phenomena

STARTING POINTS

- Light can be refracted when it travels from air into a transparent medium (in which it travels more slowly).
- The refractive index of a medium is the ratio of the speed of light in vacuum (or air) to the speed of light in the medium.
- Refraction by convex (converging) lenses is used to focus light to form images.
- Refraction through a prism disperses white light into the colours of the visible spectrum.
- Displacement, amplitude, frequency, wavelength, speed, intensity, phase and phase difference are all important terms used to describe waves.
- The wavelength, λ, frequency, f, and speed, v, of a wave are connected by the equation $v = f\lambda$.
- Progressive (travelling) waves transfer energy and they can be transverse or longitudinal.
- Electromagnetic waves (like light) are transverse. Sound is a longitudinal wave.
- Two-dimensional waves are represented in diagrams using wavefronts.
- Waves are diffracted when they pass obstacles or go through gaps. Diffraction is most significant when the gap and the wavelength are the same size.
- When waves meet, the result can be determined using the principle of superposition.
- The results of superposition at a point depend on the path difference and phase difference between the waves arriving at that point. When waves arrive in phase constructive interference occurs. If waves arrive out of phase the effect is described as destructive interference.
- A system can be made to resonate by an external frequency equal to its natural frequency.

A1 The eye and sight

A.1.1 Describe the basic structure of the human eye.

Basic structure of the human eye

In this chapter we will discuss some important properties of light, so it is appropriate to begin with an explanation of the way in which the eye functions.

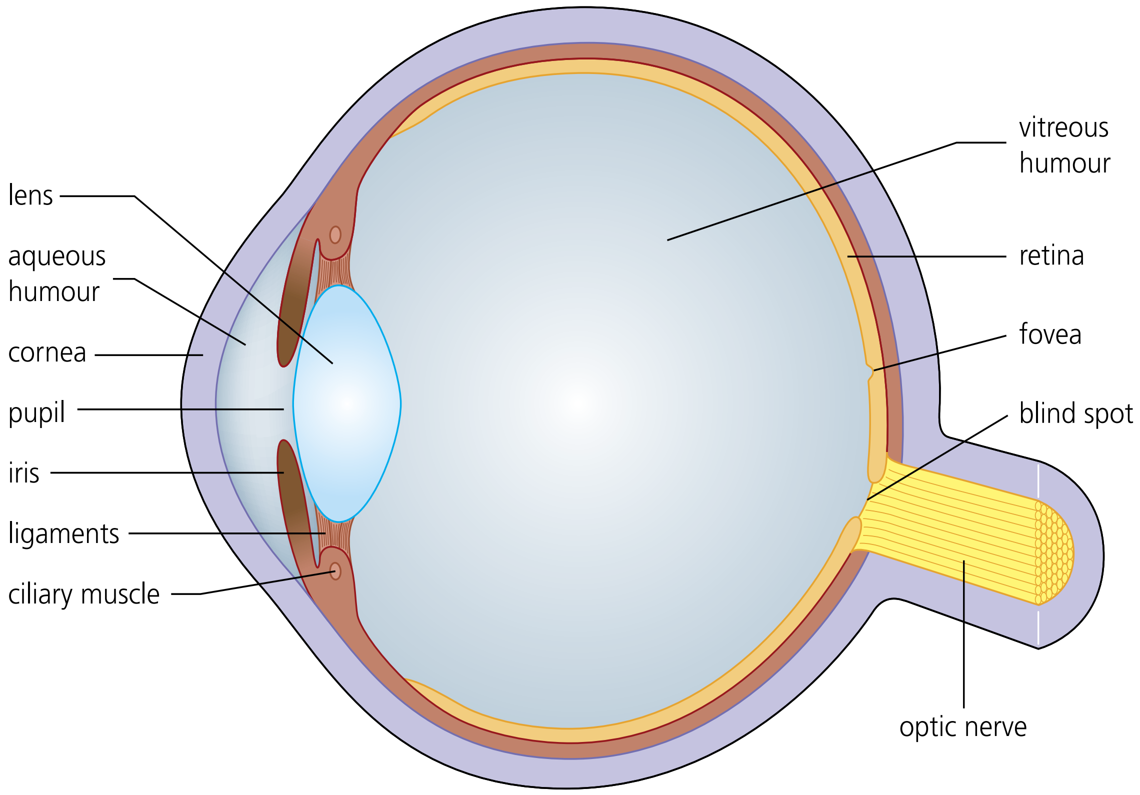


Figure 15.1 Physical features of the human eye

Figure 15.1 shows the basic structure of the human eye. Light rays are refracted as they pass into the eye through the **cornea**. Further refraction then takes place at the **lens**. As a result, the rays are focused onto the back of the eyeball (the **retina**), where an image is formed. This is shown in Figure 15.2.

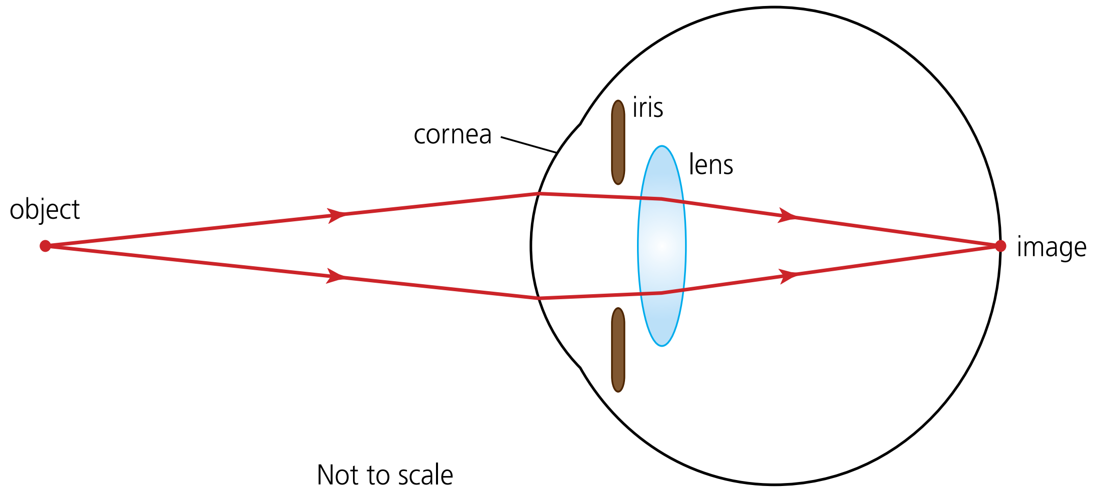


Figure 15.2 Light rays focused in the eye

Figure 15.3 Iris and pupil of the human eye

The **iris** controls the amount of light entering the eye. The aperture (opening) through which the light passes before entering the lens is called the **pupil**, shown in Figure 15.3. In bright light, the iris decreases the size of the pupil to protect the eye, whilst at night the pupil gets larger (**dilates**) so that more light can be received by the retina in order to see clearly. The **aqueous humour** is a watery liquid between the cornea and the lens; the **vitreous humour** is a clear gel between the lens and the retina.

A.1.2 **State** and **explain** the process of depth of vision and accommodation.

Accommodation and depth of vision

Figure 15.2 shows an eye focusing rays from a *nearby* object. If an eye is looking at an object which is further away, the rays incident on the cornea will not be diverging as much. If the object is a very long way away, the incident rays will be (almost) parallel. So the eye must adapt in some way to keep rays from objects at different distances still focused on the retina. To do this, the **ciliary** muscles change the shape of the lens in a process which is called **accommodation**.

In Figure 15.4a the eye is looking at a distant object and the parallel rays are focused on the retina; the eye muscles are relaxed and the lens is relatively thin. Figure 15.4b shows what would happen if the object was brought closer to the lens but the lens was unable to accommodate the change and stayed the same shape. The image is not focused on the retina. In Figure 15.4c the eye is observing a close object and is able to focus the image because the ciliary muscles have strained to make the lens fatter. (For simplicity, the ray diagrams show rays refracting in the middle of the lens. More detailed drawings would show refraction at the surfaces of both the cornea and the lens.)

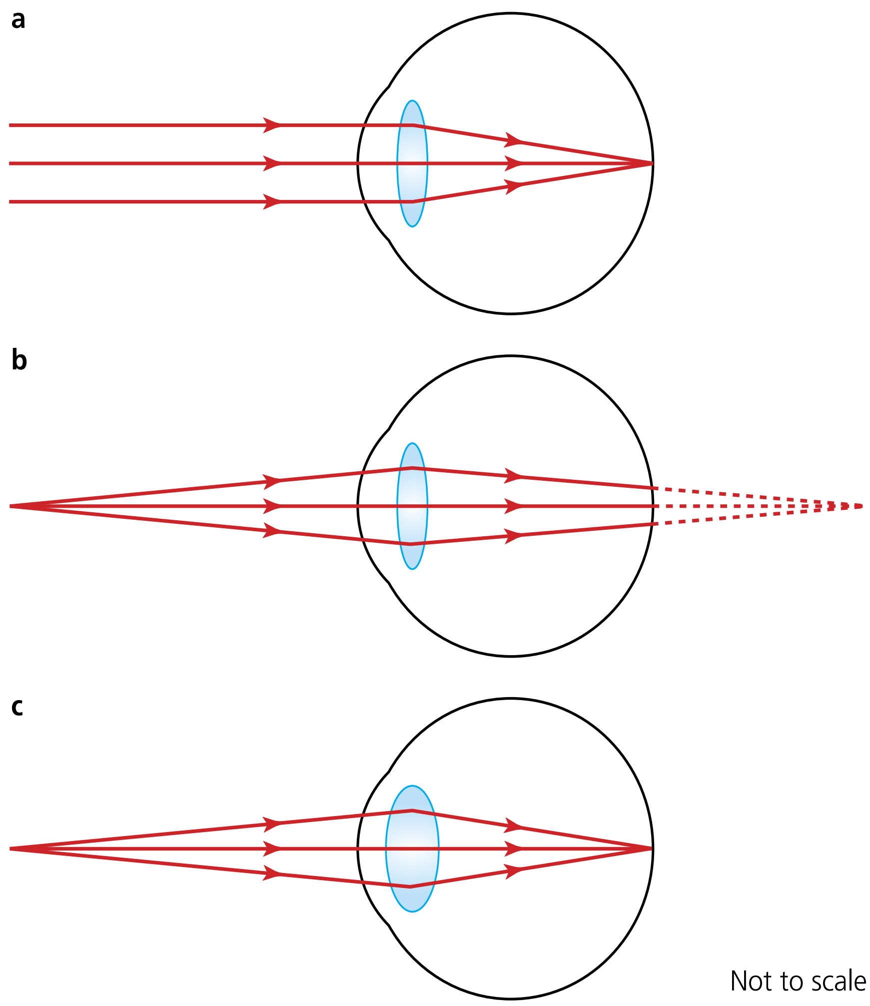


Figure 15.4 The shape of the lens changes in order to keep an image in focus

The **near point** is the closest point to the eye at which an object can be focused without excessive strain. The near point of a normal human eye (without aid) is about 25 cm from the eye. The **far point** is the furthest distance at which objects can be focused clearly. For normal vision, the distance to the far point is assumed to be infinity. An object a long way away will be focused but the image will be too small to be seen clearly. The ciliary muscles are most relaxed when looking at objects a long way away and most strained when looking at objects at the near point.

The eye cannot focus perfectly on two or more objects which are different distances away *at the same time*. However, it is possible for the quality of the focused image to be acceptable (although not perfect) for objects at different distances from the eye.

Figure 15.5 Neither the eye nor the camera can focus objects at different distances away

Depth of vision is the range of distances from the eye over which objects can be acceptably focused. Depth of vision is small for objects close to the eye but increases for objects which are further away. Depth of vision increases in brighter light, when the pupil becomes smaller. Figure 15.5 shows the limited depth of focus produced by a camera lens, which has similar depth of focus problems.

To see our surroundings in three dimensions (3D), we need to be able to see objects clearly at different distances away. Because we have two eyes we get two slightly different images (especially for nearby objects) and our brain is able to judge distance more accurately than with a single image. We also judge distance by using our previous experience and knowing how big most objects are likely to be.

Rods and cones

A.1.3 **State** that the retina contains rods and cones, and **describe** the variation in density across the surface of the retina.

When light is focused on the retina, tiny electrical signals (caused by the flow of ions across membranes) are sent to the brain along the **optic nerve**. This is done by two types of light-sensitive cells in the retina, which are known as **rods** and **cones**. Cone cells are used mainly in normal lighting conditions and they produce the appearance of detailed, coloured images. This is called **photopic** vision. Rod cells are especially useful at night and for other poor lighting conditions because they are more sensitive to light of low intensity. This is called **scotopic**

A.1.4 **Describe** the function of the rods and of the cones in photopic and scotopic vision.

vision. Images obtained mainly from rod cells have very little colour and lack detail, but rod vision is good at detecting motion.

The two types of cells are not distributed equally across the retina, as shown in Figure 15.6. The concentration of cones is at its greatest in a small area known as the **fovea**, which is at the centre of the retina. When we look directly at an object, the image is formed at the fovea and is of high quality because of the large concentration of cones (this normally leads to *acute* vision). There are no rod cells in the fovea; rod cells are distributed around the rest of the retina. The eye contains many more rod cells than cones. Vision using the rod cells is good at detecting movement from the side (*peripheral* vision).

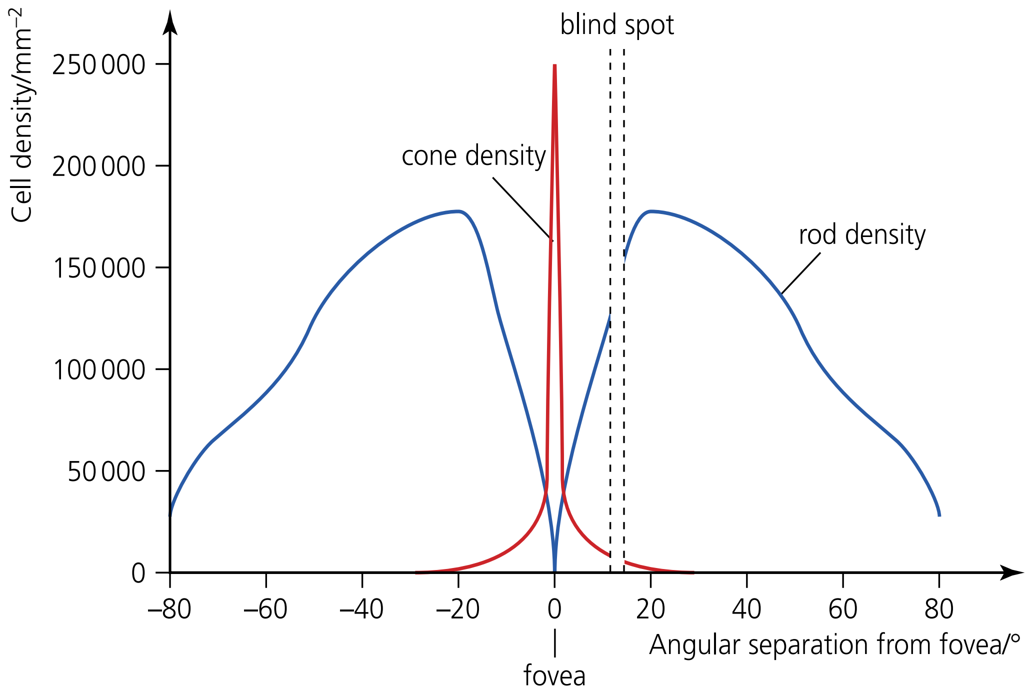


Figure 15.6 Distribution of rods and cones across the retina

There are no light-sensitive cells at the point where the optic nerve joins the retina. An image formed at this point will not be detected. This point is called the **blind spot**. An image formed at the blind spot of one eye will not fall on the blind spot of the other, so that it can still be seen.

Additional Perspectives

The fovea

The fovea occupies less than 1% of the total area of the retina and only receives light from the central two degrees of the visual field. However, it is the part of the retina where the image is formed when we look directly at a specific point. Approximately 50% of all the signals from the retina to the brain come from the fovea.

The cones in the fovea are smaller than those at other places on the retina, which means that they can be packed more closely together and thereby produce a detailed image (acute vision). The fovea has several distinct regions. Its centre has a diameter of approximately 1.5 mm, with the cones about 2×10^{-3} mm apart.

Questions

1 Estimate how many cones are located in the central fovea.

2 Figure 15.7 shows an eye looking directly at a square object which is 1 m by 1 m and located at a distance of 20 m from the eye.

a What is the angle, θ, subtended by the object at the eye?

b What are the dimensions of the image on the retina?

c Assuming the image falls on the central fovea, approximately how many cells are involved in sending the image to the brain?

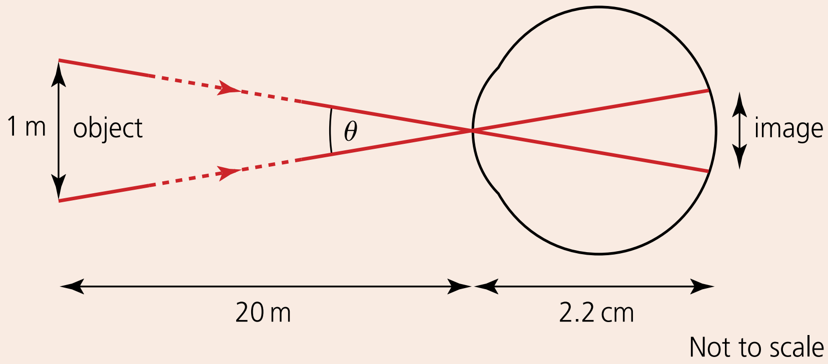


Figure 15.7

1 a A photographer makes the aperture on her camera smaller. This is called 'stopping down' the lens. Suggest two possible reasons for doing this.
b In what ways are a human eye and a camera similar?

2 a Explain why it is not possible to see clearly under water.
b Explain why wearing a mask or goggles corrects the problem.

3 Use information from Figure 15.6 to estimate the total number of rod cells in the human retina.

4 a Explain why it is difficult to see colour in objects viewed at night.
b Make a chart which compares the principal characteristics of photopic and scotopic vision.

5 A group of rod cells are connected together to send a single signal to the brain (this is called retinal summation). Suggest one advantage and one disadvantage of this (compared with each rod cell sending an individual signal).

Colour vision

The brain interprets different wavelengths of light as different colours. However, it is unlikely that a point on the retina will receive light of just one single wavelength, or even a very narrow range of wavelengths. (Such light can be provided by lasers and is described as being **monochromatic**.) Usually any point on the retina will receive a mixture of different wavelengths, but the brain will still interpret this as *one* single colour. To explain this we need to understand more about how the eye and brain work together to create the perception of colour.

The retina contains three different types of cone cell. Each responds to a range of wavelengths, but not in the same way. Between the three types, they respond to all visible light wavelengths and their peak responses are in the short, medium or long wavelength regions of the visible spectrum. They are often referred to as blue, green or red cone cells. (This is misleading if it suggests that they detect only that colour.)

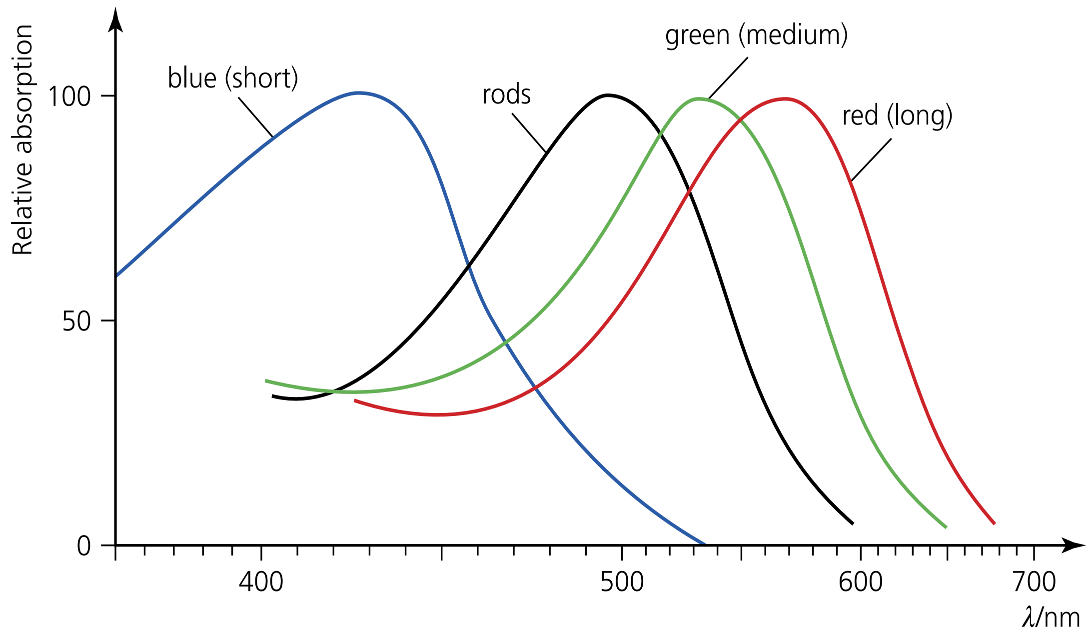


Figure 15.8 Spectral response curves for rod cells and red, green and blue cone cells

The spectral response curves for the three types of cone cells are shown in Figure 15.8. The response of the rod cells has also been included for comparison.

Taken on its own, each individual cone cell gives no indication of colour. It is only when the retina and the brain combine the responses from the different types of cone cell that a particular colour is interpreted. For example, consider the cone cells in an eye looking at monochromatic light of wavelength 600 nm. The red cells would respond much more than the green cells, and the response from the blue cells would be negligible. This combination would be interpreted by the brain as orange–red.

We see a few objects because they emit light but, more commonly, we see objects because of the light that they scatter or reflect into our eyes. White light contains a spectrum of colours and an object viewed in white light might appear to be green, for example, because only green light is reflected from its surface (the other colours have been absorbed). However, it is more probable that it appears green because that is the combined effect on the eye and the brain of a range of different received wavelengths (see Figure 15.9 overleaf).

If, for any reason, one type of cone cell is not functioning correctly, the ability to detect different colours is impaired because there are only two types of cell providing signals to the

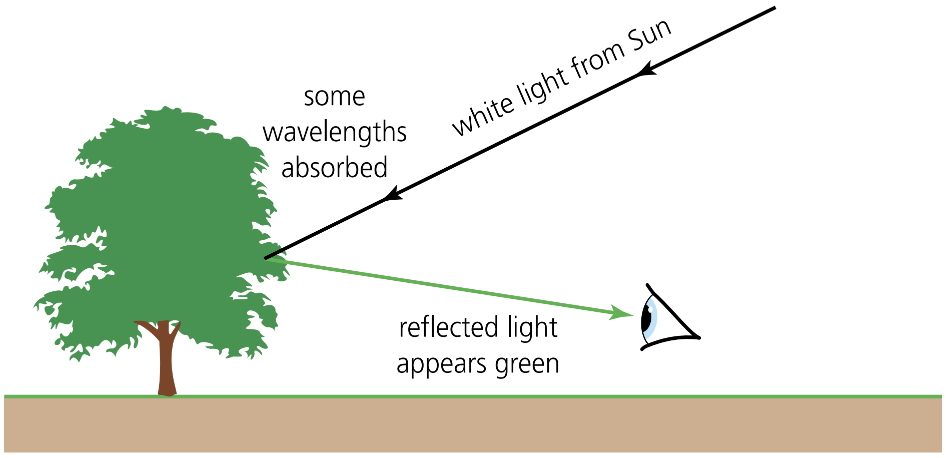


Figure 15.9 The tree appears green because the leaves have absorbed some of the wavelengths of white light

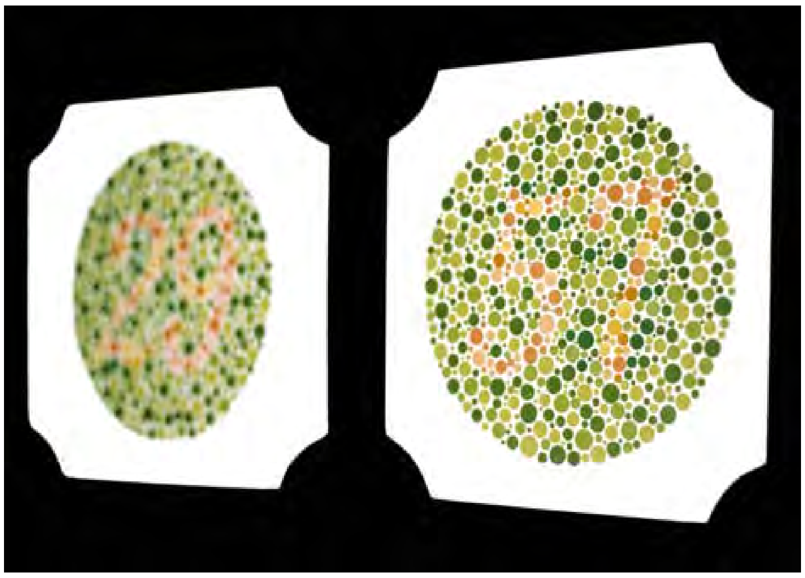


Figure 15.10 Colour blindness tests. Colour-blind people will not be able to see the numbers

brain, rather than three. This is known as (partial) **colour blindness**. It is usually an inherited condition that is more common among males than females. The most common example of colour blindness results in a difficulty distinguishing between red and green. A test for colour blindness is shown in Figure 15.10. Total colour blindness is very rare and will occur if only one of the three kinds of cone cell is working correctly.

TOK Link: Is colour real?

We may say that yellow light (for example) has a frequency of 5×10^{14} Hz, but in no sense is the electromagnetic wave 'yellow'. Colour is not a property of the wave. The concept of different colours is the way in which (the eye and) the brain distinguish and describe different wavelengths or combinations of wavelengths.

There is more about perception in the TOK Link on page 581.

Questions

1. We may all agree that the colour of a leaf on a tree is green, but is there any way of determining if the 'green' colour that you see is the same as the 'green' colour seen by a friend?
2. Is there any reliable way of explaining the colour orange, for example (other than comparing it to something else of the same colour)?

Creating different colours by addition and subtraction

A.1.5 **Describe** colour mixing of light by addition and subtraction.

The eye and the brain can distinguish a very large range of different colours by combining the responses from just three different types of cone cell. In a similar way, light of the three colours red, green and blue can be added to produce any other colour, by mixing in various proportions. Approximately equal proportions of red, green and blue light will produce the effect of white light. For this reason, red, green and blue are often described as the **primary colours** (although this is *not* a fundamental property of the light). Figure 15.11 shows the effect of mixing the three primary colours in equal proportions. Note that white light is formed where all three primary colours overlap. Magenta, yellow and cyan are known as the **secondary colours**. These colours are formed where only two of the three primary colours are mixed. It is not possible to make one of the primary colours by mixing the other two.

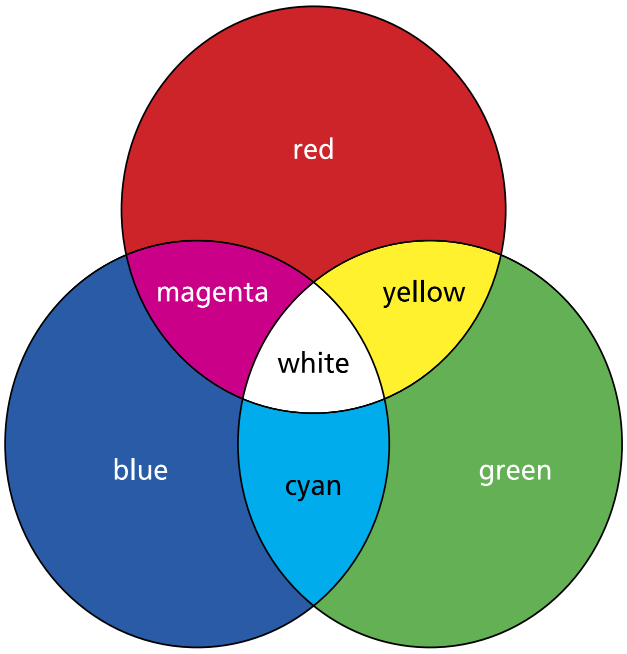


Figure 15.11 Primary and secondary colours

Various kinds of LED displays, TVs, and the screens of mobile phones, laptops and tablets, etc., use the mixing of the three primary colours to achieve a vast range of different colours.

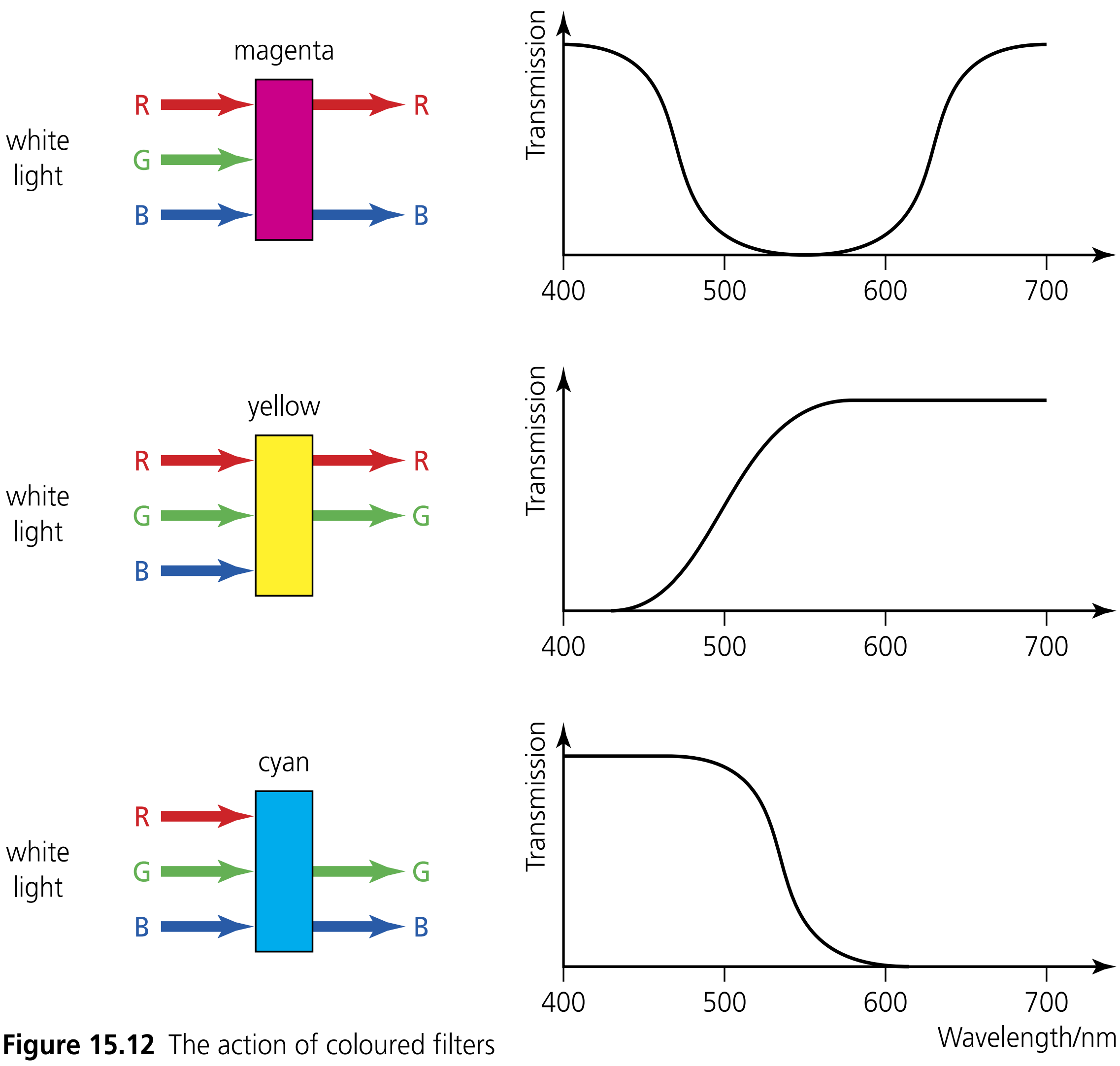


Figure 15.12 The action of coloured filters

The most common way of producing light of a certain colour is by passing white light through a colour **filter** and thereby removing (subtracting) certain colours. Figure 15.12 shows the action of three filters (producing the three secondary colours). For example, a yellow filter absorbs certain wavelengths so that the transmitted light appears yellow. This might be because only pure yellow light is transmitted or, more likely (as shown), the transmitted wavelengths have a combined effect of yellow (blue light is absorbed and red and green are transmitted).

Computer simulations are a good way to illustrate colour addition and subtraction, as well as the other visual effects described in this section.

Using lighting to create visual effects

A.1.6 Discuss the effect of light and dark, and colour, on the perception of objects.

When we look at something, what we 'see' has as much do with our brains as with our eyes. The term visual **perception** is used to describe how the brain interprets the signals sent to it by the eye. Our perception of objects develops with experience and is very much influenced by what we expect to see, based on a lifetime of previous observations. Sometimes, this can mean that our perceptions can be very wrong. Figure 15.13 shows an interesting example of how our perceptions can easily be confused.

For centuries, artists, architects and designers have understood and used the effects of shading and shadows to create strong impressions of size, shape and depth. The floodlighting of buildings at night is used to enhance these dramatic effects (Figure 15.14).

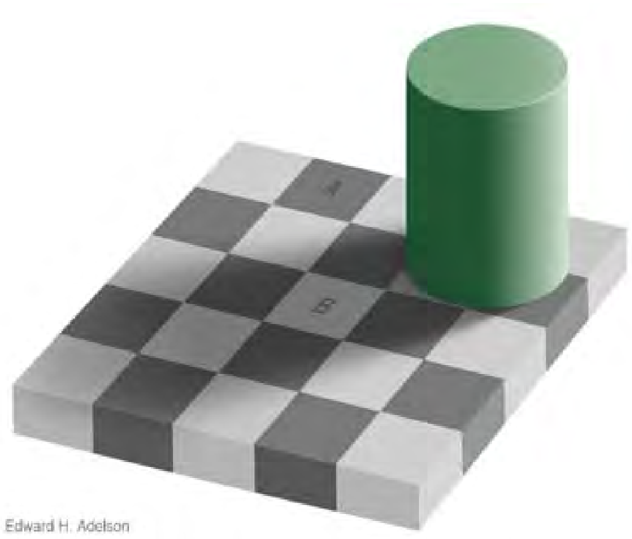


Figure 15.13 Optical illusion: squares A and B are exactly the same shade of grey!

Figure 15.14 Lighting effects can give an impression of size and depth

The drawings of the Dutch graphic artist Maurits Cornelis Escher (1898–1972) provide another interesting example of how the brain can be tricked into false perceptions (Figure 15.15).

Colours can also have a strong psychological effect on perception. In interior design pale colours are used to give a feeling of space; for example, a room can be made to appear larger if the ceiling is lighter than the walls. Pale colours towards the red/orange end of the spectrum are often described as warm and comforting. Alternatively, the blues and violets at the other end of the spectrum are considered to be 'cold'.

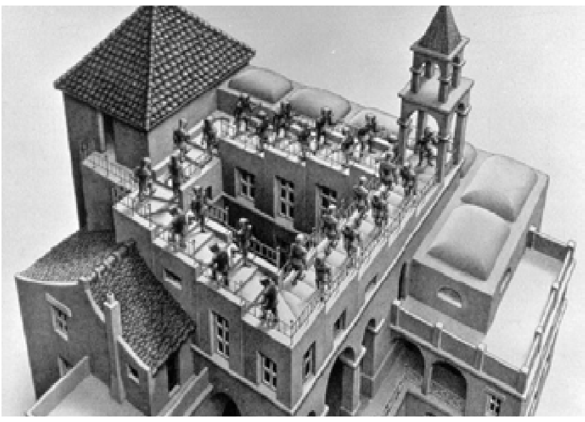


Figure 15.15 Ascending and Descending M.C. Escher's © 2012 The M.C. Escher Company – Baarn – The Netherlands. All rights reserved. www.mcescher.com

6 a Use the graph shown in Figure 15.8 to compare the responses of the four different types of light-sensitive cells to light of wavelength 490 nm.
b What colour is this wavelength?

7 Explain why we can usually see an object most clearly by looking directly at it, but at night-time we may be able to see it better by looking in a slightly different direction.

8 An object reflecting pure red light is viewed under
a bright light and
b poor lighting conditions.
Explain the differences in the appearance of the object under these conditions.

9 a What colour will a white object appear when it is illuminated by red and green light at the same time?
b What colours would be seen by mixing these colours of light?
i cyan and magenta **ii** yellow and magenta

10 a What colour is seen when white light is passed through both yellow and magenta filters?
b What two secondary filters used together will produce green light?

11 Suggest some jobs for which colour blindness would be a significant handicap.

A2 Standing (stationary) waves

Describing standing waves

A.2.1 **Describe** the nature of standing (stationary) waves.

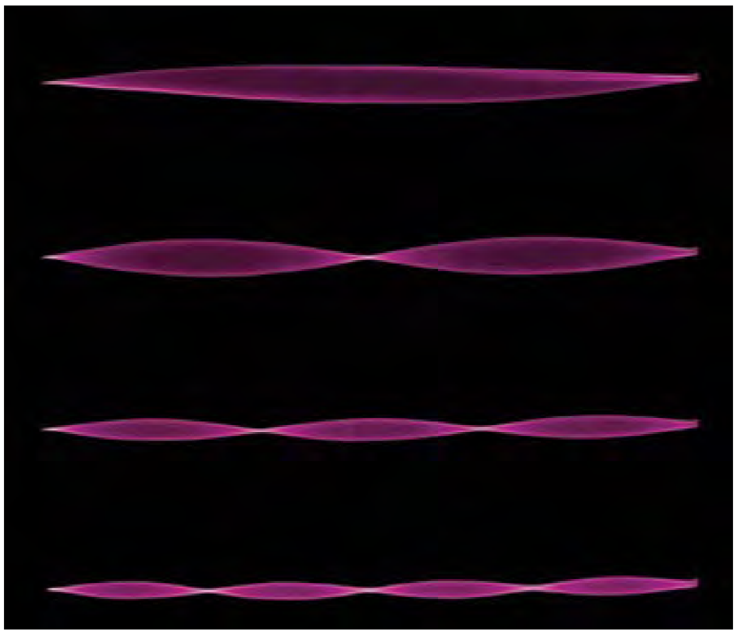


Figure 15.17 Standing waves on a stretched string

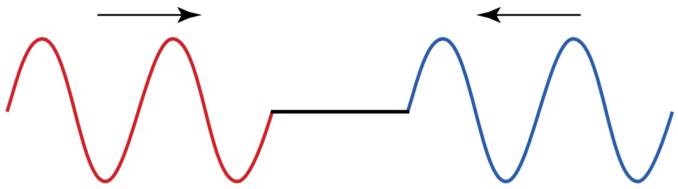


Figure 15.16 Two sinusoidal waves travelling towards each other

Consider two travelling waves of the same shape, frequency, wavelength and amplitude moving in opposite directions, such as shown in Figure 15.16, which could represent transverse waves on a string or rope.

As these waves pass through each other they will combine to produce an oscillating wave pattern that does not change its position. Such patterns are called **standing waves** (sometimes stationary waves) and typical examples are shown in Figure 15.17. Note that a camera produces an image over a short period of time (not an instantaneous image) and that is why the fast-moving string appears blurred. This is equally true when we view such a string with our eyes.

There are points in a standing wave where the displacement is *always* zero. These points are called **nodes**. At positions between the nodes, the oscillations are all **in phase** but the amplitude of the oscillations will vary. Midway between nodes, the amplitude is at its maximum. These positions are called **antinodes**. Figure 15.18 represents the third wave in the photograph in Figure 15.17 diagrammatically. Note that the distance between alternate nodes (or antinodes) is one wavelength.

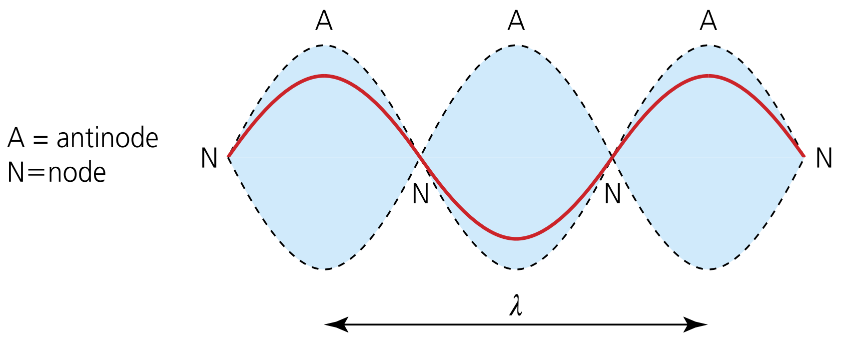


Figure 15.18 Nodes and antinodes in a standing wave on a stretched string. The solid line represents a possible position of the string at one moment

There is energy associated with a standing wave and, without dissipative forces, the oscillations would continue for ever. Energy is *not* transferred by a standing wave.

Explaining standing waves

A.2.2 Explain the formation of one-dimensional standing waves.

When waves similar to those shown in Figure 15.16 meet, **interference** occurs. We can explain the standing wave pattern by determining the resultant at any place and time, and we can do this by using the *principle of superposition* (Chapter 4). The overall displacement is the sum of the two individual displacements at that moment. Nodes occur at places where the two waves are *always* out of phase. At other places, the displacements will oscillate between zero and a maximum value which depends on the phase difference. At the antinodes the two waves are always perfectly in phase. (Students are recommended to use a computer simulation to illustrate this time-changing concept.)

Standing waves are possible with any kind of wave moving in one, two or three dimensions. For simplicity, discussion has been confined to one-dimensional waves, such as transverse waves on a stretched string.

Modes of vibration of transverse waves on strings

Stretched strings

A.2.3 Discuss the modes of vibration of strings and air in open and closed pipes.

Standing waves occur most commonly when waves are repeatedly reflected back from boundaries in a confined space, like waves on a string stretched between two fixed ends. If a stretched string is struck or plucked, it can usually only vibrate if it sets up a standing wave with nodes at both fixed ends. The simplest way in which it can vibrate is shown at the top of Figure 15.19. This is known as the **fundamental mode** of vibration (also called the **first harmonic**). It is usually the most important mode of vibration, but a series of other harmonics is possible and can occur at the same time. Some of these harmonics are also shown in Figure 15.19.

The wavelength, λ_0, of the fundamental mode (first harmonic) is $2l$, where l is the length of the string. The speed of the wave, v, along the string depends on the tension and the mass per unit length. The fundamental frequency, f_0, can be calculated from:

$$f_0 = \frac{v}{\lambda_0}$$

$$= \frac{v}{2l}$$

For a given type of string under constant tension, the fundamental frequency is inversely proportional to the length.

The wavelengths of the harmonics are, starting with the longest, $2l, \frac{2l}{2}, \frac{2l}{3}, \frac{2l}{4}$… and so on.

The corresponding frequencies, starting with the lowest, are f_0, $2f_0$, $3f_0$, $4f_0$ … and so on.

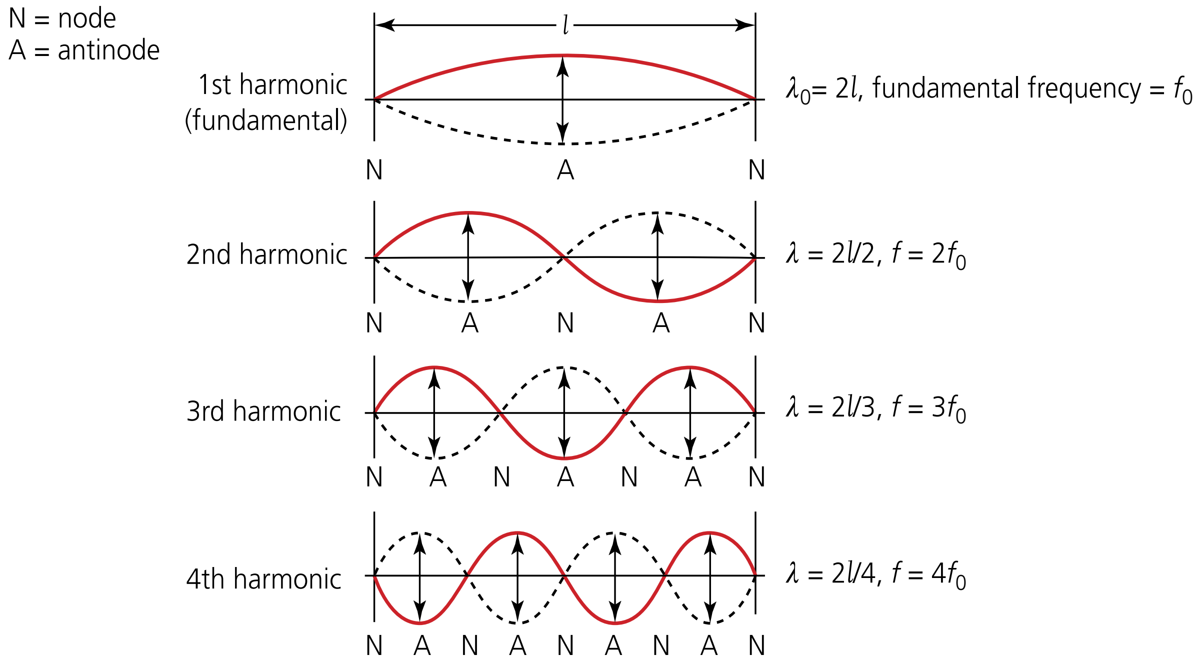


Figure 15.19 Modes of vibration of a stretched string

Additional Perspectives

Stringed musical instruments

When musical notes are played on stringed musical instruments, like guitars, cellos (see Figure 15.20b) and pianos, the strings vibrate mainly in their fundamental modes, but various other harmonics will also be present; this is one reason why each instrument has its own, unique sound. Figure 15.20a shows a range of frequencies that might be obtained from a vibrating guitar string. The factors affecting the fundamental frequency of a note are the length of the string, the tension and the mass per unit length, for example middle C has a frequency of 261.6 Hz. The standing transverse waves of the vibrating strings are used to make the rest of the musical instrument oscillate at the same frequency and, when the vibrating surfaces strike the surrounding air, travelling longitudinal sound waves propagate away from the instrument to our ears.

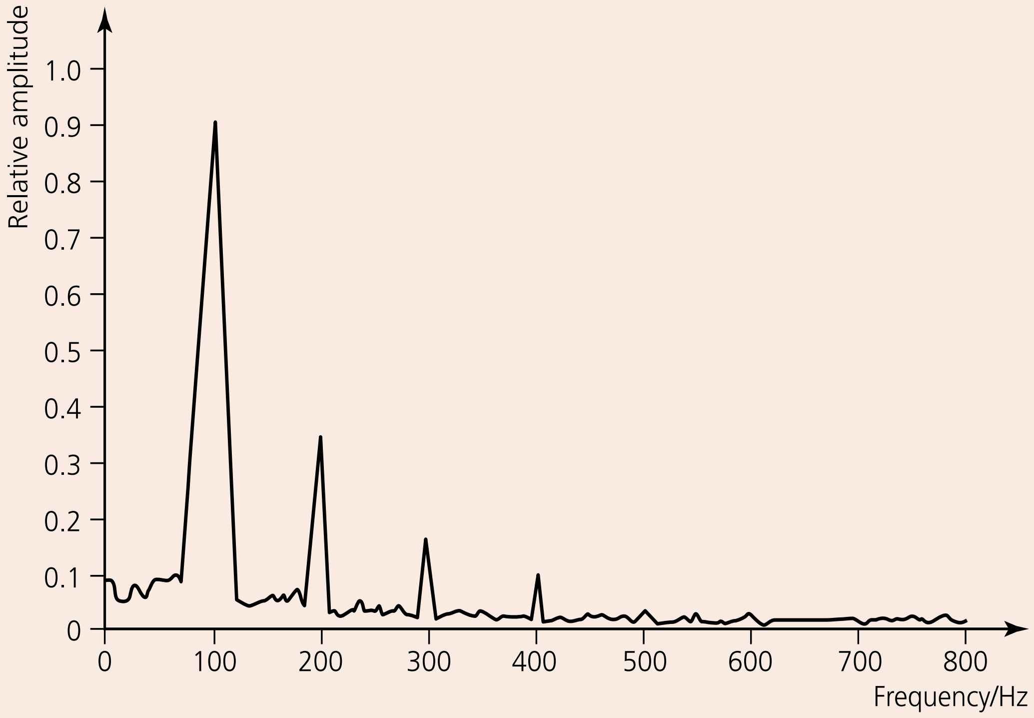


Figure 15.20a Frequency spectrum from a guitar string

Figure 15.20b Creating standing waves on a cello

Worked example

1 A string has a length of 1.2 m and the speed of transverse waves on it is $8.0\,m\,s^{-1}$.

a What is the wavelength of the fundamental mode (first harmonic)?
b Draw sketches of the first four harmonics.
c What is the frequency of the third harmonic?

a $\lambda_0 = 2l = 2 \times 1.2 = 2.4\,m$

b See Figure 15.19.

c $\lambda = \frac{2.4}{3} = 0.8\,m$

$f = \frac{v}{\lambda} = \frac{8.0}{0.8} = 10\,Hz$

Longitudinal sound waves in pipes

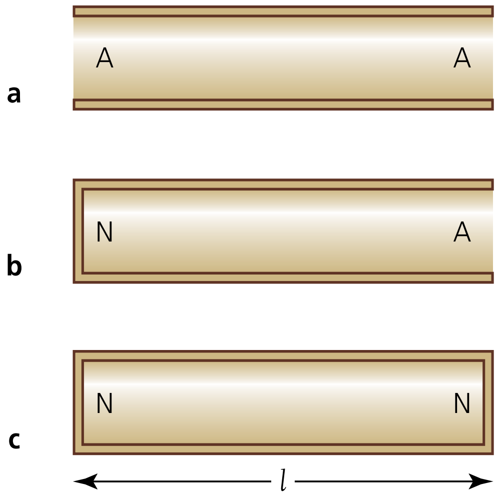


Figure 15.21 Nodes and antinodes at the ends of open and closed pipes

Air can be made to vibrate and produce standing longitudinal sound waves in various containers and tubes. The sound produced by blowing across the top of an empty bottle is an everyday example of this. Many musical instruments, such as a flute or a clarinet, use the same idea. For simplicity, we will only consider standing waves in **pipes** of uniform shape (sometimes called air **columns**).

As with strings, in order to understand what wavelengths and frequencies can be produced, we need to consider the length of the pipe and what happens at the end points (boundaries) of the wave, sometimes called the **boundary conditions**.

This is illustrated in Figure 15.21. In **a** the pipe is open at both ends, so it must have antinodes, A, at the ends, and at least one node in between. In **b** the pipe is open at one end (antinode) and closed at the other end (node, N). In **c** the pipe is closed at both ends, so it must have nodes at the ends, and at least one antinode in between.

Figure 15.22 shows the first three harmonics for a pipe open at both ends. The fundamental wavelength (twice the distance between adjacent nodes or antinodes) is $2l$. The fundamental frequency is therefore inversely proportional to the length. (Note that drawings of standing longitudinal waves can be confusing and the curved lines in the diagram are an indication of the amplitude of vibration. They should not be mistaken for transverse waves.)

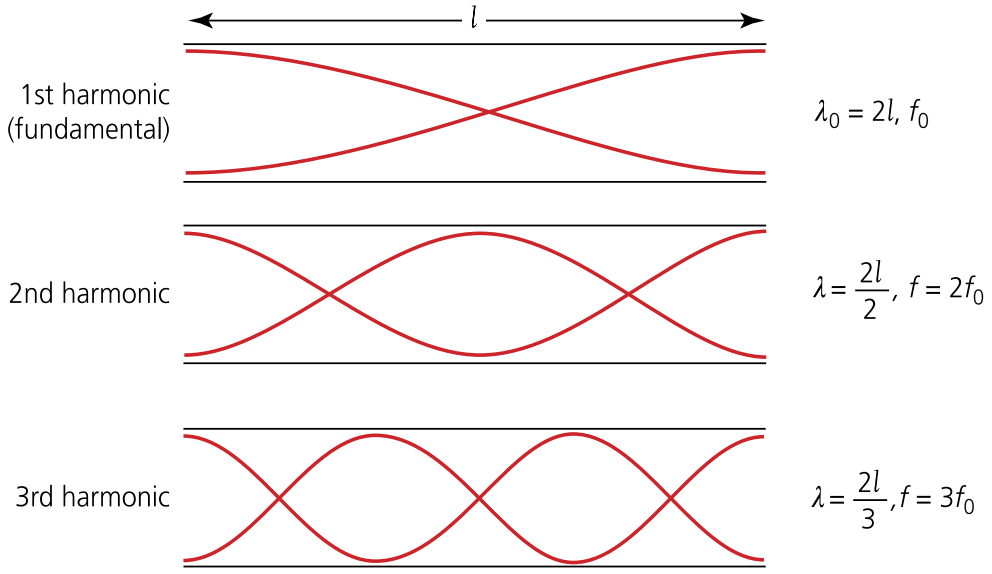


Figure 15.22 The first three harmonics in a pipe open at both ends

A pipe closed at both ends also has a fundamental wavelength of $2l$.

Figure 15.23 shows the first three possible harmonics for a pipe open at one end and closed at the other. Only odd harmonics are possible under these circumstances. The fundamental wavelength is $4l$.

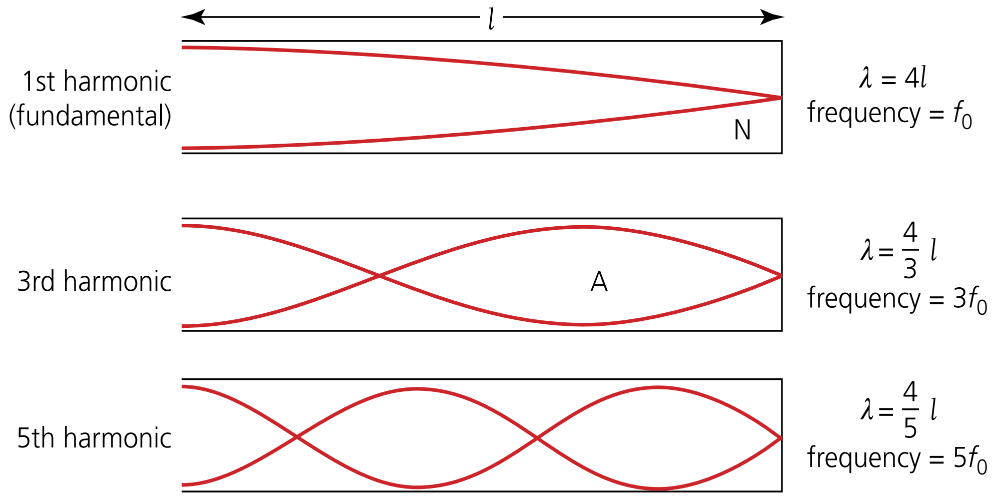


Figure 15.23 Harmonics in a pipe open at one end

Figure 15.24 shows a way of demonstrating standing waves with sound. A speaker is placed close to the open end of a long transparent pipe which is closed at the other end. Some powder is scattered all along the pipe and when the loudspeaker is turned on and the frequency carefully adjusted, the powder is seen to move into separate piles. This is because the powder tends to move from places where the vibrations are large (antinodes) to the nodes, where there are no vibrations.

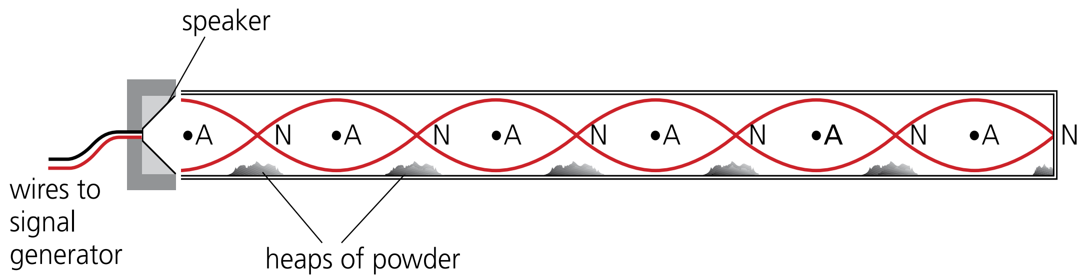


Figure 15.24 Demonstrating a standing wave with sound

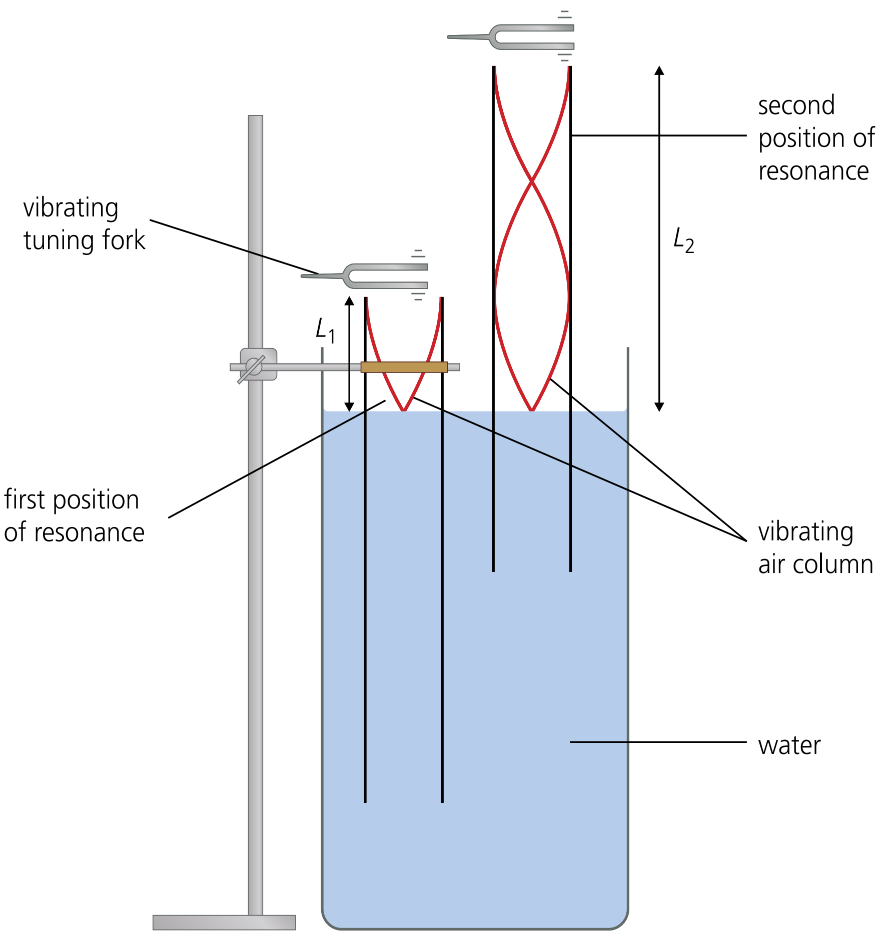


Figure 15.25 Demonstrating resonance with a tuning fork

The experiment shown in Figure 15.24 is a further demonstration of *resonance* (discussed in Chapter 4) because the applied frequency must be the same as one of the pipe's harmonic frequencies for energy to be transferred to move the powder.

Another way of demonstrating resonance in a pipe is by using a **tuning fork** of a known frequency, as shown in Figure 15.25. A vibrating tuning fork is held above the open end of a pipe. The pipe is open at the top and closed at the bottom by the level of water. The length of the pipe above the water is slowly increased until a louder sound is heard. This will be the first position of resonance. If the length of the pipe above water is increased again, then further positions of resonance may be found. Resonance will only occur when the length of the pipe above water is such that one of its harmonic frequencies is the same as the frequency of the tuning fork. Measurements can be made during this demonstration that will enable a value for the speed of sound to be determined.

Worked example

2 a If the tuning fork in Figure 15.25 had a frequency of 659 Hz, calculate the length L_1. (Assume the speed of sound in air is 340 m s^{-1}.)
b How far will the pipe need to be raised to obtain the next position of resonance?

a $\lambda = \frac{v}{f} = \frac{340}{659} = 0.516\,\text{m}$

This wavelength will be four times the length of the tube (see Figure 15.23).

So, $L_1 = \frac{0.516}{4} = 0.129\,\text{m}$

In reality this is only an approximate answer because antinodes do not occur exactly at the open ends of tubes. (If a more accurate answer is needed in a calculation, it is possible to use an 'end correction' which is related to the diameter of the tube.)

b Refer to Figure 15.25. In the first position of resonance the pipe contains one-quarter of a wavelength. In the second position the pipe contains three-quarters of a wavelength. Therefore, it must be raised half a wavelength, or 0.258 m.

Summary of differences between standing waves and travelling waves

A.2.4 **Compare** standing waves and travelling waves.

Table 15.1 Comparison of standing waves and travelling waves

	Standing waves	**Travelling waves**
Wave pattern	Stationary/standing	Progressive/travelling
Energy transfer	No energy is transferred	Energy is transferred through the medium
Amplitude (assuming no energy dissipation)	Amplitude at any one place is constant but it varies with position between nodes. Maximum amplitude at antinodes; zero amplitude at nodes	All oscillations have the same amplitude
Phase	All oscillations between adjacent nodes are in phase	Oscillations one wavelength apart are in phase. Oscillations between are not in phase
Frequency	All oscillations have the same frequency	All oscillations have the same frequency
Wavelength	Twice the distance between adjacent nodes	Shortest distance between points in phase

A.2.5 **Solve** problems involving standing waves.

12 A first harmonic is seen on a string of length 123 cm at a frequency of 23.8 Hz.
- **a** What is the wave speed?
- **b** What is the frequency of the third harmonic?
- **c** What is the wavelength of the fifth harmonic?

13 **a** What is the phase difference between two points on a standing wave which are:
- **i** one wavelength apart
- **ii** half a wavelength apart?

b The distance between adjacent nodes of the third harmonic on a stretched string is 18.0 cm, with a frequency of 76.4 Hz. Sketch the waveform of this harmonic.
c On the same drawing add the waveforms of the fundamental mode and the fourth harmonic.
d Calculate the wavelength and frequency of the fifth harmonic.
e What was the wave speed?

14 Explain why observing standing waves on a stretched string using a mechanical vibrator can be considered as a demonstration of resonance.

15 A certain guitar string has a fundamental frequency of 262 Hz.
- **a** If the tension in the string is increased, suggest what happens to the speed of waves along it.
- **b** If the string is adjusted so that the speed increases, explain what will happen to the fundamental frequency.
- **c** Suggest why a wave will travel more slowly along a thicker string of the same material, under the same tension.
- **d** Explain why thicker strings of the same material, same length and same tension produce notes of lower frequency.

16 Draw the first three harmonics for a pipe which is closed at both ends.

17 A and B are two similar pipes of the same length. A is closed at one end, but B is open at both ends. If the fundamental frequency of A is 180 Hz, what is the fundamental frequency of B?

18 If the frequency used in the demonstration shown in Figure 15.24 was 6.75 kHz and the piles of powder were 2.5 cm apart, what was the speed of the sound waves?

19 What length must an organ pipe (open at one end, see Figure 15.26) have if it is to produce a note of fundamental frequency 90 Hz?
(Use speed of sound = 340 m s^{-1})

20 An organ pipe open at both ends has a second harmonic of frequency 228 Hz.
- **a** What is its length?
- **b** What is the frequency of the third harmonic?
- **c** What is the wavelength of the fourth harmonic?
- **d** Suggest one advantage of using organ pipes that are closed at one end, rather than open at both ends.

Figure 15.26 On this church organ the wooden pipes are open at one end and the metal pipes behind are open at both ends

Additional Perspectives

Microwave ovens

The frequency of the microwaves which are used to cook food is chosen so that it is absorbed by water and other polarized molecules in the food. Such molecules are positively charged at one end and negative at the other end. They respond to the oscillating electromagnetic field of the microwave by gaining kinetic energy because of their increased vibrations, which means that the food gets hotter. Most microwave ovens operate at a frequency of 2.45 GHz. This frequency of radiation is penetrating, which means that the food its not just heated on the outside.

To ensure that the microwaves do not pass out of the oven into the surroundings, the walls, floor and ceiling are metallic, although the door may have a metallic mesh (with holes) that will reflect microwaves, but allow light, with its much smaller wavelength, to pass through, so that the contents of the oven can be seen from outside.

The microwaves reflect off the oven walls, which means that the walls do not absorb energy, as they do in other types of oven. The 'trapping' of the microwaves is the main reason why microwave ovens cook food quickly and efficiently. However, reflected microwaves can combine to produce various kinds of standing wave in the oven and this can result in the food being cooked unevenly. To reduce this effect the microwaves can be 'stirred' by a rotating deflector as they enter the cooking space, or the food can be rotated on a turntable.

Questions

1 Calculate the wavelength of microwaves used for cooking.
2 Design an experiment which would investigate if there were significant nodes and antinodes in a microwave oven. How far apart would you expect any nodes to be?

A3 The Doppler effect

A.3.1 **Describe** what is meant by the Doppler effect.
A.3.2 **Explain** the Doppler effect by reference to wavefront diagrams for moving-detector and moving-source situations.

When we hear a sound we can usually assume that the frequency (pitch) that is heard by our ears is the same as the frequency that was emitted by the source. But if the source of the sound is moving towards us (or away from us) we will hear a sound with a different frequency. This is usually only noticeable if the movement is fast; the most common example is the sound heard from a car or train that moves quickly past us.

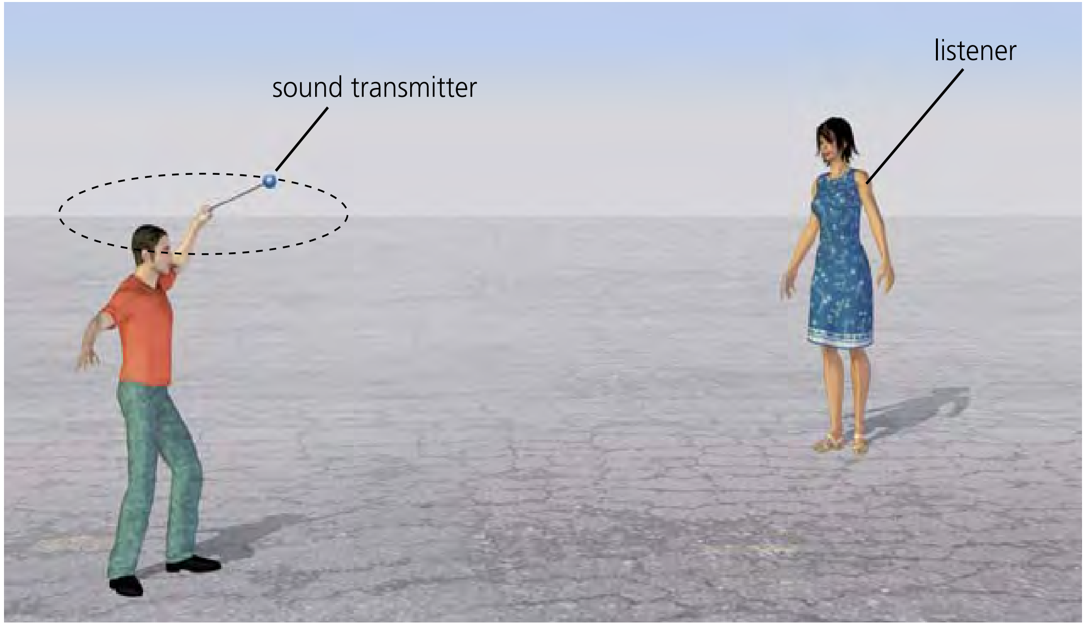


Figure 15.27 Demonstrating the Doppler effect with sound waves

This change of frequency that is detected when there is relative motion between a source and a receiver of waves is called the **Doppler effect**. (The Doppler effect is named after the Austrian physicist, Christian Doppler, who first proposed it in 1842.) The Doppler effect may occur with any kind of wave.

Figure 15.27 shows a way in which the Doppler effect with sound can be demonstrated. A small source of sound (of a single frequency) is spun around in a circle. When the source is moving towards the listener a higher frequency is heard; when it is moving away, a lower frequency is heard.

The easiest way to explain the Doppler effect is by considering wavefronts (Figure 15.28). Figure 15.28a shows the common situation in which a stationary source, S, emits waves which travel towards a stationary detector, D. Figure 15.28b shows a detector moving towards a stationary source and Figure 15.28c shows a source moving directly towards a stationary detector. Similar diagrams can be drawn to represent the situations where the source and detector are moving apart.

The detector in **b** will meet more wavefronts in a given time than if it remained in the same place, so that the received frequency, f', is greater than the emitted frequency, f. In **c** the distance between the wavefronts (the wavelength λ) between the source and the observer is reduced, which again means that the received frequency will be greater than the emitted frequency. (Frequency $= v/\lambda$ and the wave speed, v, is constant. The speed of sound through air does not vary with the motion of the source or observer.)

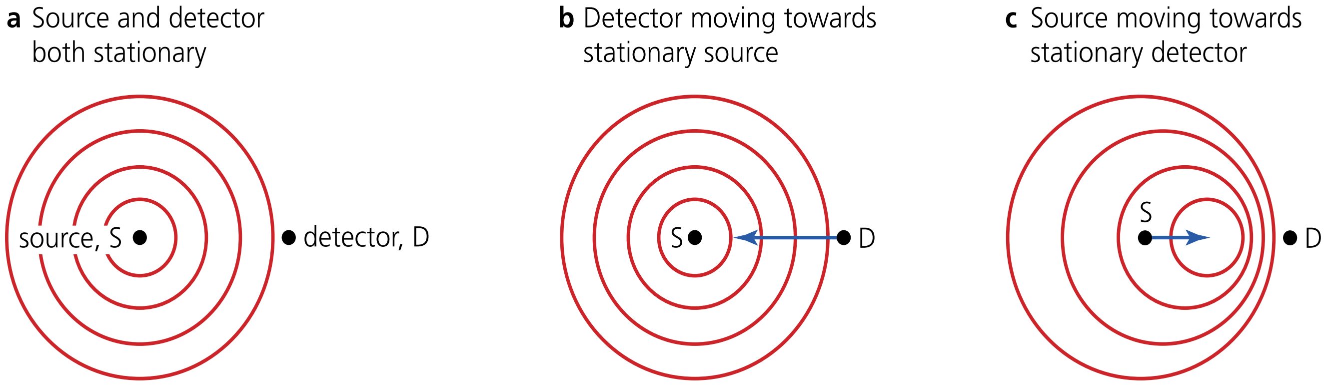


Figure 15.28 Wavefront diagrams to demonstrate the Doppler effect

Equations for the Doppler effect

A.3.3 **Apply** the Doppler effect equations for sound.

Figure 15.29a shows waves of frequency, f, and wavelength, λ, travelling at a speed, v, between a stationary source, S, and a stationary observer, O. (The term *observer* can be used with any kind of waves, not just light.) In the time, t, that it takes the first wavefront emitted from the source to reach the observer, the wave has travelled a distance vt. The number of waves between the source and observer is ft. The wavelength, λ, equals the total distance divided by the number of waves $= vt/ft = v/f$, as we would expect.

Figure 15.29b represents exactly the same waves emitted in the same time from a source moving towards a stationary observer with a speed u_s. In time, t, the source has moved from S_1 to S_2. The number of waves is the same as in **a**, but because the source has moved forwards a distance, $u_s t$, the waves between the source and the observer are now compressed within a length, $(vt - u_s t)$.

This means that the observed (received) wavelength, λ', equals the total distance divided by the number of waves:

$$\lambda' = \frac{vt - u_s t}{ft} = \frac{v - u_s}{f}$$

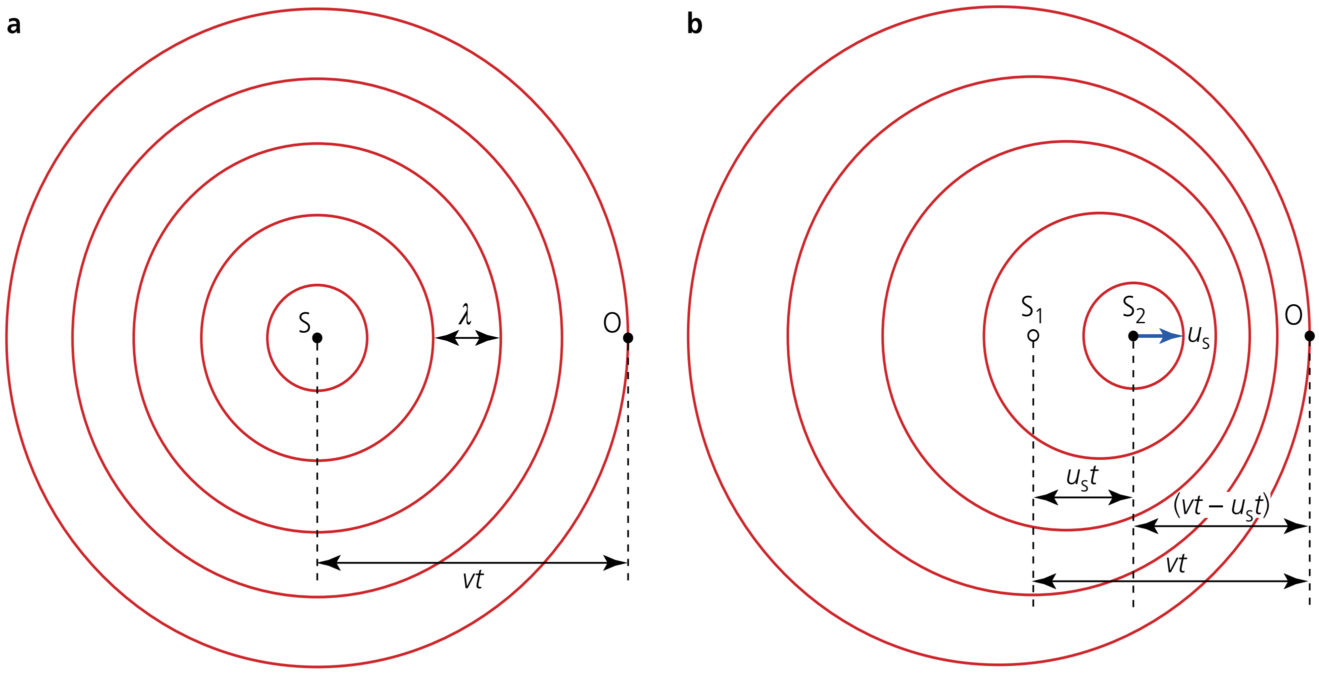


Figure 15.29a Waves between a stationary source and a stationary observer **b** Waves between a moving source and a stationary observer

The observed (received) frequency, f', is given by:

$$f' = \frac{v}{\lambda} = \frac{vf}{v - u_s}$$

If the source is moving away from the observer, the equation becomes

$$f' = \frac{vf}{v + u_s}$$

In general, we can write:

$$f' = \left(\frac{v}{v \pm u_s}\right)$$

This is the equation for the Doppler effect from a *moving source* detected by a stationary observer and it is given in the IB *Physics data booklet*.

In a similar situation, the equation for the frequency detected by a *moving observer* from a stationary source is:

$$f' = f\left(\frac{v \pm u_o}{v}\right)$$

This equation is also given in the IB *Physics data booklet*.

If the source of the sound and the observer are getting closer together, along a line directly between them, a higher frequency sound (than that emitted) will be detected by the observer and that frequency will be constant, although it will increase in intensity. Similarly, if the source and the observer are moving apart, the observed sound will have a lower, constant frequency (than that emitted) and it will decrease in intensity. The frequency must change quickly at the moment the source and detector move past each other. (If the motion is not directly between the source and the observer, or both the source and the observer are moving, the principles are the same, but the mathematics is more complicated and it is not included in this course.)

Worked examples

3 a A source of sound emitting a frequency of 480 Hz is moving directly towards a stationary observer at $50\,\text{m s}^{-1}$. If it is a hot day and the speed of sound is $350\,\text{m s}^{-1}$, what frequency is received?
b What frequency would be heard on a cold day when the speed of sound was $330\,\text{m s}^{-1}$?
c Explain why the speed of sound is less on a colder day.

a $f' = \dfrac{vf}{(v - u_s)}$

$$f' = \frac{(350 \times 480)}{(350 - 50)} = 560\,\text{Hz}$$

b $f' = \dfrac{(330 \times 480)}{(330 - 50)} = 566\,\text{Hz}$

c Sound is transferred though air by moving air molecules. On a colder day the molecules will have a lower average speed.

4 What frequency will be received by an observer moving at $24\,\text{m s}^{-1}$ directly away from a stationary source of sound waves of frequency 980 Hz? (Take the speed of sound to be $342\,\text{m s}^{-1}$.)

$$f' = \frac{f(v \pm u_o)}{v}$$

$$f' = \frac{980 \times (342 - 24)}{342} = 910\,\text{Hz}$$

A.3.4 Solve problems on the Doppler effect for sound.

21 An observer receives sound of frequency 436 Hz from a train moving at $18\,\text{m s}^{-1}$ directly towards him. What was the emitted frequency? (Take the speed of sound to be $342\,\text{m s}^{-1}$.)

22 An observer is moving directly towards a source of sound emitted at 256.0 Hz, with a speed of $26.0\,\text{m s}^{-1}$. If the received sound has a frequency of 275.7 Hz, what was the speed of sound?

23 A car emitting a sound of 190 Hz is moving directly away from an observer who detects a sound of frequency 174 Hz. What was the speed of the car? (Take the speed of sound to be $342\,\text{m s}^{-1}$.)

24 A train is moving at constant speed along a track as shown in Figure 15.30 and is emitting a sound of constant frequency.
a Suggest how the sound heard by an observer at point P will change as the train moves from A to B.
b Describe the sound heard by someone sitting on the train during the same time.

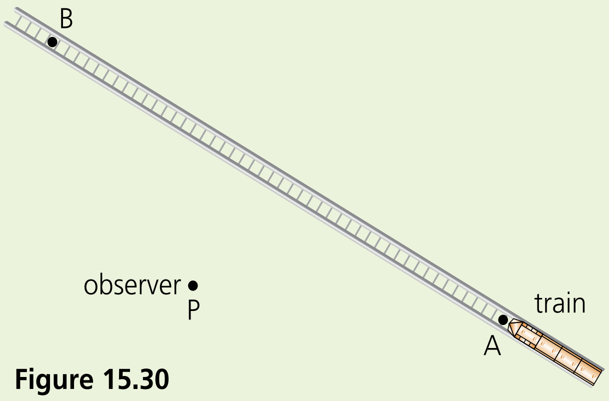


Figure 15.30

25 The source of sound shown in Figure 15.27 is rotating at 4.2 revolutions per second in a circle of radius 1.3 m. If the emitted sound has a frequency of 287 Hz, what is the difference between the highest and lowest frequencies which are heard? (Take the speed of sound to be $340\,\text{m s}^{-1}$.)

The Doppler effect with electromagnetic waves

A.3.5 **Solve** problems on the Doppler effect for electromagnetic waves using the approximation $\Delta f = \frac{v}{c} f$.

The Doppler effect also occurs with electromagnetic waves, but the situation is made more complicated because the received speed of electromagnetic waves is unaffected by the speed of the observer. The equations given in the previous section cannot be used with electromagnetic waves.

However, the following equation for the change (shift) in frequency, Δf, can be used if the relative speed between source and observer, v, is very much less than the speed of the electromagnetic waves, c, ($v << c$).

$$\Delta f = \frac{v}{c} f$$

This equation is given in the IB *Physics data booklet*.

Since the speed of electromagnetic waves is so high ($c = 3.00 \times 10^8\,\mathrm{m\,s^{-1}}$ in vacuum or air) this equation can nearly always be used with accuracy.

Worked example

5 A plane travelling at a speed of $250\,\mathrm{m\,s^{-1}}$ transmits a radio signal at a frequency of 130 MHz. What change of frequency will be detected by the airport it is travelling towards?

$$\Delta f = \frac{v}{c} f = \frac{250}{(3.00 \times 10^8)} \times (1.3 \times 10^8) = 108\,\mathrm{Hz}$$

The airport will receive a frequency 108 Hz higher than 1.3×10^8 Hz. This is a very small increase, requiring good quality electronic circuits to be detected.

Clearly the shift in frequencies of electromagnetic waves is not something we will observe in everyday life. It becomes more significant for very fast moving objects like stars.

Using the Doppler effect to measure speed

A.3.6 **Outline** an example in which the Doppler effect is used to measure speed.

In order to measure the (average) velocity of a moving object we usually observe the position of the object on two occasions between known time intervals. One way of doing this, especially if the object is inaccessible, like a plane for example, is to send waves towards it and then detect the waves as they are reflected back to the transmitter. From this data, the direction from the transmitter to the plane can be calculated. The distance between the transmitter and the plane can be calculated from the time delay between the sent and the received signals. If this process is repeated, the velocity of the plane can be determined. In this example of a simple **radar** system, microwaves are used.

The Doppler effect, however, provides a better way of determining the speed of a moving object, like a plane, using the same kind of waves (see Figure 15.30). Some animals, like bats (see Figure 15.31) and dolphins, use the same principles but with sound or ultrasonic waves.

Figure 15.30 Air traffic control uses the Doppler effect

If waves of a known speed, v, and frequency, f, are directed towards a moving object and reflected back, the object will effectively be acting like a moving source of waves; the Doppler equations can be used to determine the speed of the object if the received frequency, f', can be measured.

The measurement of the speeds of cars, as well as planes and other vehicles, is a common example of the use of the Doppler effect. In many countries the police reflect electromagnetic radiation off cars in order to determine their speeds. Microwaves and infrared radiation are commonly used for this purpose. The speed of athletes or balls in sports can also be determined using the Doppler effect. The measurement of the rate of blood flow in an artery is another interesting example, which is shown in Figure 15.32.

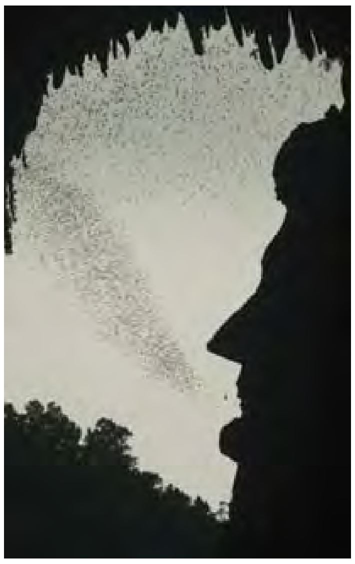


Figure 15.31 These bats in Malaysia use the Doppler effect to navigate

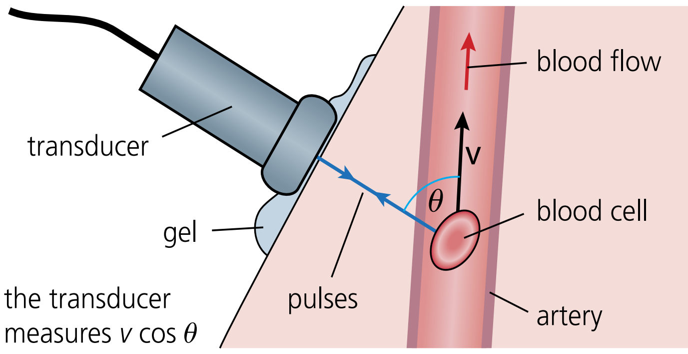


Figure 15.32 Measuring blood flow rate using the Doppler effect

Pulses of ultrasonic waves are sent into the body from the **transducer** and are reflected back from blood cells flowing in an artery. The received waves have a different frequency because of the Doppler effect and the change of frequency can be used to calculate the speed of the blood. This information can be used by doctors to diagnose many medical problems. Because the waves usually cannot be directed along the line of blood flow, the calculated speed will be the component ($v \cos\theta$).

There are also some very important applications of the Doppler effect used in astronomy. The decrease in frequency of radiation received from distant galaxies is known as the 'red shift', and it provides very important information about the nature and age of the universe.

26 **a** What change of frequency will be received back from a car moving directly away at 135 km h^{-1} if the radiation used in the 'speed gun' has a frequency of 24 GHz?
b Suggest reasons why ultrasonic waves are not usually used in speed guns.

27 **a** An airport radar system using microwaves of frequency 98.2 MHz sends out a pulse of waves that is reflected off a plane which is flying directly away. If the reflected signal was received back at the airport 8.43×10^{-5} s later, at a frequency 85.2 Hz lower than was emitted, what was the speed of the plane and how far away was it?
b Suggest how it might be possible for a plane to avoid being detected by radar.

28 In the television broadcast of some sports, for example, baseball, tennis and cricket, viewers can see a replay of the trajectory (path) of a moving ball. Use the Internet to find out how this is done. (Is the Doppler effect used?)

29 A star emits radiation of frequency 1.42×10^9 Hz. When received on Earth the frequency is 1.38×10^9 Hz. What is the speed of the star? Is it moving towards the Earth, or away from the Earth?

Additional Perspectives

Breaking the sound barrier

As an object, like a plane, flies faster and faster, the sound waves that it makes get closer and closer together in front of it. When a plane reaches the speed of sound, at about 1200 km h^{-1}, the waves superpose to create a 'shock wave'. This is shown in Figure 15.34.

When a plane reaches the speed of sound it is said to be travelling at 'Mach 1' (named after the Austrian physicist, Ernst Mach). Faster speeds are described as 'supersonic' and twice the speed of sound is called Mach 2, etc. As Figure 15.33 shows, the shock wave travels away from the side of the plane and may be heard on the ground as a 'sonic boom'. Similarly, 'bow waves' can often be seen spreading from the front (bow) of a boat because boats usually travel faster than the water waves they create. The energy transferred by such waves can cause a lot of damage to the land at the edges of rivers (river banks).

For many years some engineers doubted if the sound barrier could ever be broken. The first confirmed supersonic flight (with a pilot) was in 1947. Now it is common for military aircraft to travel faster than Mach 1, but Concorde and Tupolev 144 were the only supersonic passenger aircraft in regular service.

It is possible to use a whip to break the sound barrier. If the whip gets thinner towards its end then the speed of a wave along it can increase until the tip is travelling faster than sound (in air). The sound it produces is often described as a whip 'cracking'.

Figure 15.33 Plane breaking the sound barrier

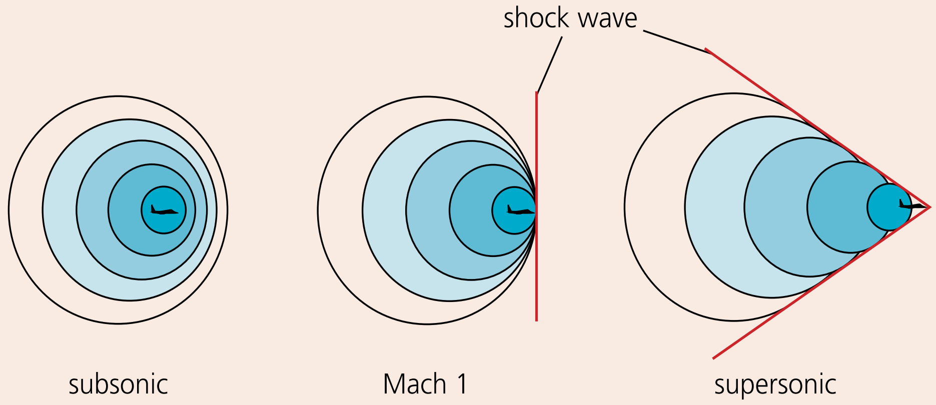


Figure 15.34 Creating a shock wave in air

Question

1 In a world in which time is so important for so many people, suggest reasons why there are no longer supersonic aircraft in commercial operation.

A4 Diffraction at a single slit

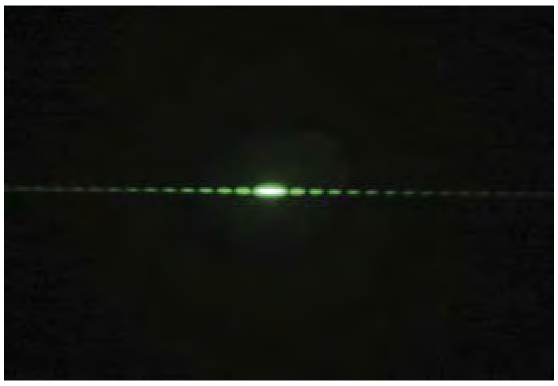


Figure 15.35 Diffraction pattern of monochromatic light passing through a narrow slit

Figure 15.35 shows a photograph of a diffraction pattern for light passing through a narrow vertical slit, which is the type of **aperture** (opening) usually used to produce simple diffraction patterns. The diffraction pattern seen on the screen is a series of bands (fringes) of light and dark. The central band is brighter and about twice the width of the others.

Figure 15.36 shows the diffraction pattern produced when light passes through a small circular aperture.

By comparing Figures 15.35 and 15.36, we can see that the shape of the pattern depends on the shape of the aperture. The light used to produce these patterns was **monochromatic**, which means that it has only one wavelength (and frequency). Light from most sources will consist of many different wavelengths and will not produce such a simple pattern.

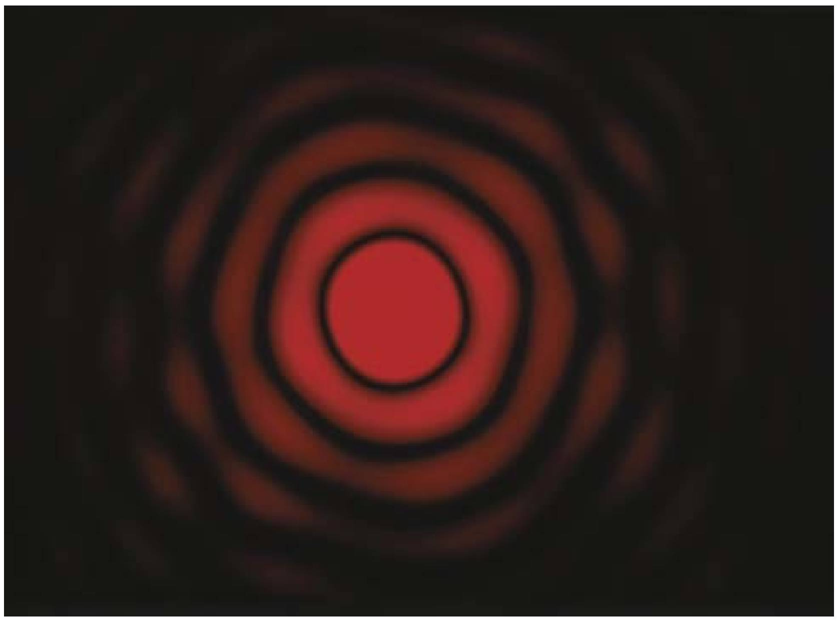


Figure 15.36 Monochromatic light diffracted by a very small circular aperture

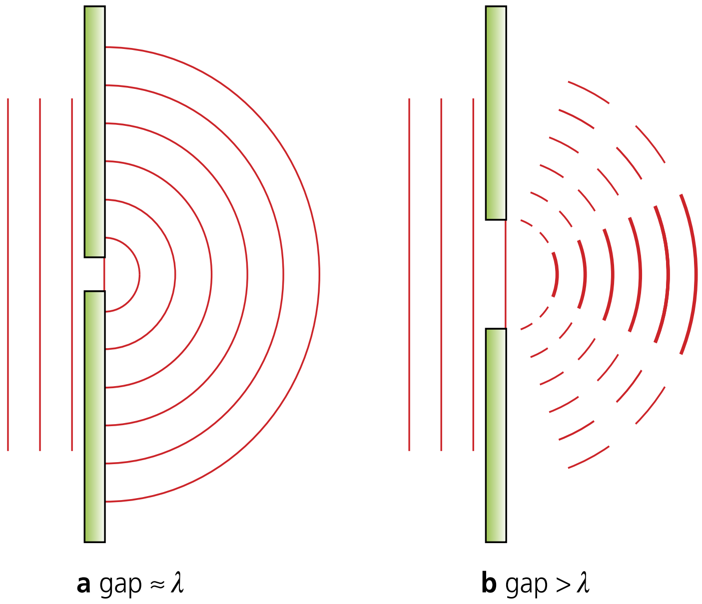


Figure 15.37 Diffraction of waves by different sized apertures

In order to produce these diffraction patterns for light, the waves must travel away from the aperture in some directions, but not in others. The simple diffraction theory covered in Chapter 4 does not explain this. Figure 15.37a shows the diffracted waves coming away from a gap of about the same width as the wavelength (as discussed in Chapter 4). The gap acts effectively like a point source, spreading waves forward equally in *all* directions. This *cannot* be used to help to explain the diffraction of light because the wavelengths of light are so much smaller than the width of even a very small hole. For example, a hole of diameter 0.5 mm is 1000 times greater than the wavelength of green light (approximately 5×10^{-7} m).

To produce a diffraction pattern for light, such as shown in Figure 15.35, we must develop a theory which explains why waves spread away from a gap that is larger than the wavelength, as shown in Figure 15.37b.

Explaining the diffraction pattern produced by a single slit

A.4.2 **Derive** the formula $\theta = \lambda/b$ for the position of the first minimum of the diffraction pattern produced at a single slit.

Diffraction patterns can be explained by considering that each point on a wavefront passing through the slit acts as a point source of **secondary waves**. (This idea was first suggested in 1678 by the Dutch physicist Christian Huygens, when he proposed that light waves propagate forwards because *all* the points on any wavefront act as sources of secondary waves.) What happens after the wave has passed through the slit depends on how these secondary waves **interfere** with each other.

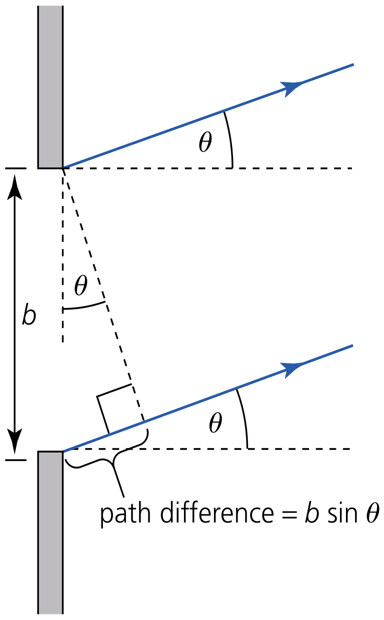


Figure 15.38 Path differences and interference

Figure 15.38 shows a direction, θ, in which secondary waves travel away from a slit of width, b. If θ is zero, all the secondary waves will interfere constructively in this direction (straight through the slit) because there is no **path difference** between them. (Of course, in theory, waves travelling parallel to each other in the same direction cannot meet and interfere, so we will assume that the waves' directions are *very nearly* parallel.)

Consider what happens for angles increasingly greater than zero. The path difference, as shown in Figure 15.38, equals $b \sin\theta$ and this increases as the angle θ increases. There will be an angle at which the waves from the two edges of the slit interfere constructively because the path difference has increased to become one wavelength, λ.

But if secondary waves from the edges of the slit interfere constructively, what about interference between all the other secondary waves? Consider Figure 15.39 on page 574 in which the slit has been divided into a number of point sources of secondary waves. (Ten points have been chosen, but it could be many more.)

If the angle, θ, is such that secondary waves from points 1 and 10 would interfere constructively because the path difference is one wavelength, then secondary waves from 1 and 6 must have a path difference of half a wavelength and interfere destructively. Similarly, waves from points 2 and 7, points 3 and 8, points 4 and 9 and points 5 and 10 must all interfere destructively. In this way waves from *all* points can be 'paired off' with

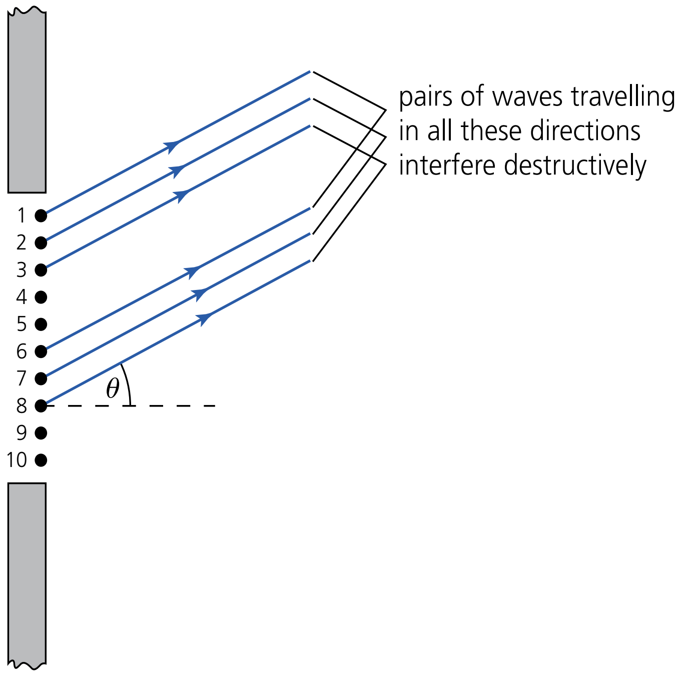


Figure 15.39 Secondary waves which will interfere destructively can be 'paired off'

others, so that the first *minimum* of the diffraction pattern occurs at such an angle that waves from the edges of the slit would otherwise interfere constructively.

The first minimum of the diffraction pattern occurs when the path difference between secondary waves from the edge of the slit is equal to one wavelength. That is, if $b\sin\theta = \lambda$.

For the diffraction of light, the angle θ is usually small and approximately equal to $\sin\theta$ if the angle is expressed in radians. (This was first explained in Chapter 4.)

The angle for the first minimum of a single slit diffraction pattern is $\theta = \dfrac{\lambda}{b}$.

This equation is given in the IB *Physics data booklet*.

Relative intensity–angle graph for single slit diffraction

A.4.1 **Sketch** the variation with angle of diffraction of the relative intensity of light diffracted at a single slit.

Similar reasoning to that discussed in the previous section can be used to show that further diffraction minima occur at angles $2\lambda/b$, $3\lambda/b$, $4\lambda/b$, etc. Figure 15.40 shows a graphical interpretation of these figures and approximately how it corresponds to a drawing of a single slit diffraction pattern.

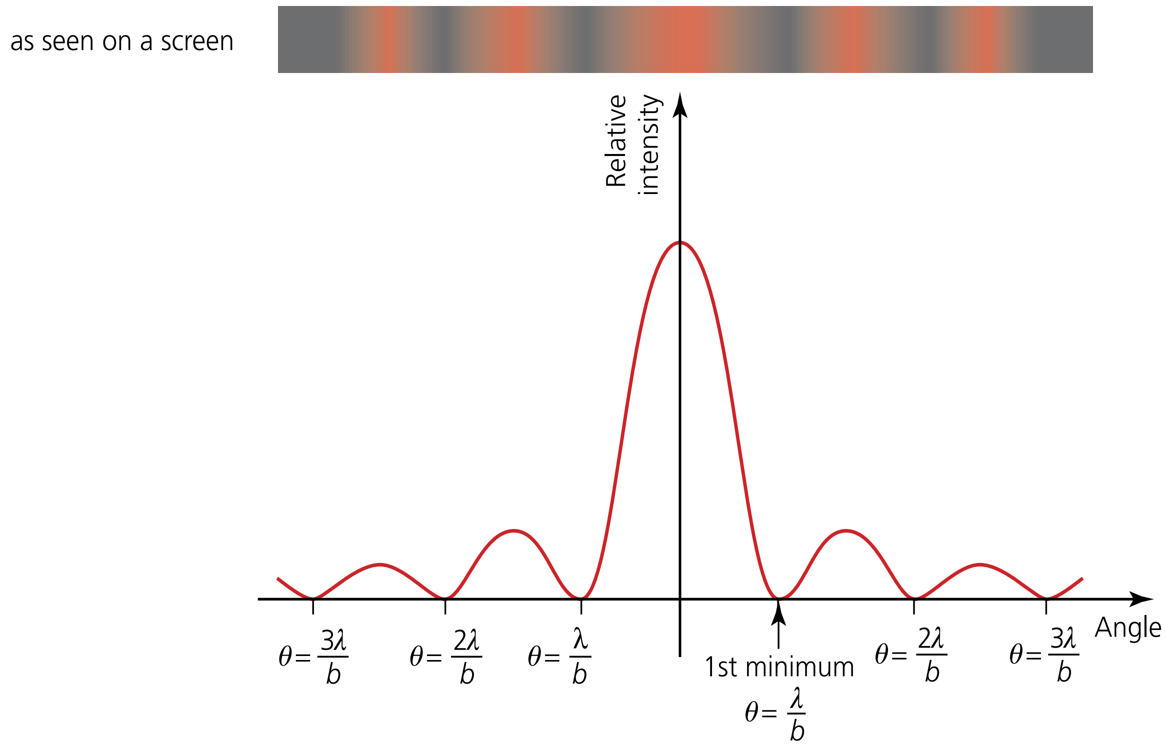


Figure 15.40 Variation of intensity with angle for single slit diffraction

Worked example

6 Monochromatic light of wavelength 663 nm is shone through a gap of width 0.0730 mm.

a At what angle is the first minimum of the diffraction pattern formed?

b If the pattern is observed on a screen which is 2.83 m from the slit, what is the width of the central maximum?

a $\theta = \lambda/b$

$$\theta = \frac{663 \times 10^{-9}}{7.30 \times 10^{-5}}$$

$$= 9.08 \times 10^{-3} \text{ radians}$$

b See Figure 15.41.

$$\theta \approx \sin\theta = \frac{\text{half width of central maximum}}{\text{slit to screen distance}}$$

half width of central maximum = $(9.08 \times 10^{-3}) \times 2.83 = 0.0257$ m
width of central maximum = $0.0257 \times 2 = 0.0514$ m

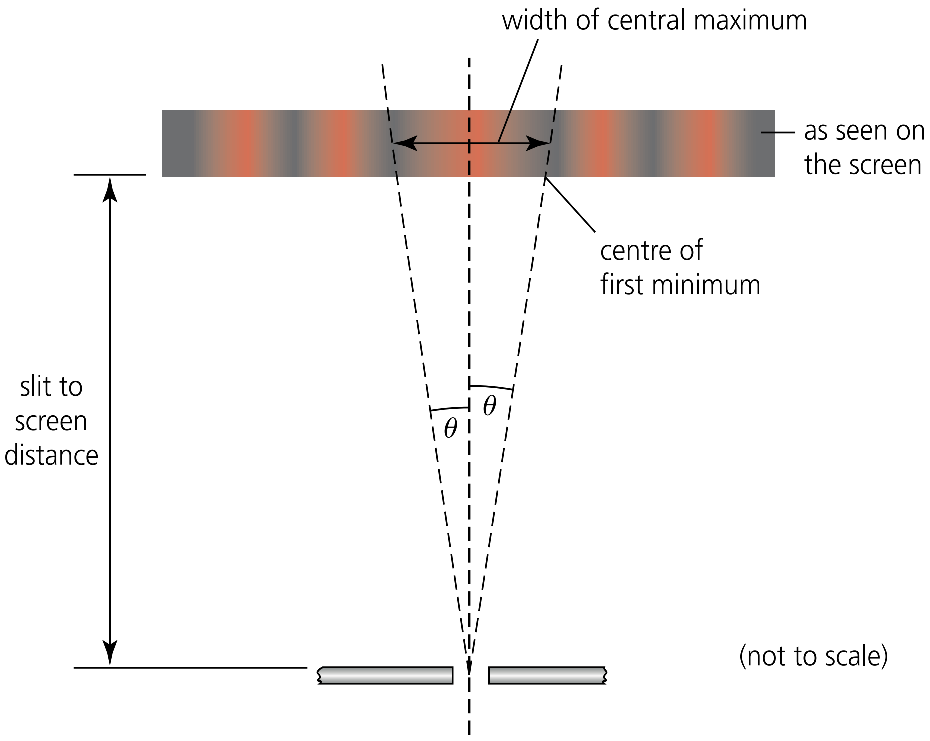


Figure 15.41

A.4.3 **Solve** problems involving single slit diffraction.

30 Electromagnetic radiation of wavelength 2.37×10^{-7} m passes through a narrow slit of width 4.70×10^{-5} m.
- **a** In what part of the electromagnetic spectrum is this radiation?
- **b** Suggest how it could be detected.
- **c** Calculate the angle of the first minimum of the diffraction pattern.

31 What is the wavelength of light that has a first diffraction minimum at an angle of 0.0038 radians when it passes through a slit of width 0.15 mm?

32 When light of wavelength 6.2×10^{-7} m was diffracted through a narrow slit, the central maximum had a width of 2.8 cm on a screen which was 1.92 m from the slit. What was the width of the slit?

33 **a** Sketch and label a relative intensity against angle graph for the diffraction of red light of wavelength 6.4×10^{-7} m through a slit of width 0.082 mm. Include at least five peaks of intensity.
- **b** Add to your graph a sketch to show how monochromatic blue light would be affected by the same slit.

The importance of diffraction

The diffraction of light is further evidence confirming the wave nature of light, and when it was discovered that electrons could be diffracted, their wave nature was also understood for the first time (see page 580 and Option B). Other electromagnetic radiations will be diffracted significantly by objects and gaps that have a size comparable to their wavelengths. For example, the diffraction of X-rays by atoms, ions and molecules can be used to provide detailed information about solids with regular structure. The diffraction of radio waves used in communication is also of considerable importance when choosing which wavelength to use in order to get information efficiently from the transmitter to the receiver.

Waves travelling from place to place can have their direction changed by diffraction, but diffraction can also be important when waves are emitted or received. They can then be considered to be passing through an aperture, and the ratio λ/b will be important when considering how the waves spread away from a source, or what direction they appear to come from when they are received.

Additional Perspectives

Diffraction by multiple apertures

We have discussed how light waves are affected by passing through a single slit, but diffraction effects can be greatly increased by the use of many identical slits arranged close together in a regular pattern. Such pieces of apparatus are called diffraction gratings. Figure 15.42 shows a diffraction grating, which will typically have 500 or more slits in every millimetre.

The use of multiple apertures greatly increases the intensity and the sharpness of diffraction patterns. Diffraction gratings are often used in a similar way to prisms for the analysis of light (discussed in Chapter 7).

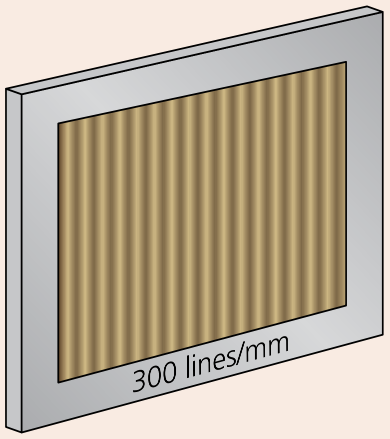


Figure 15.42 A diffraction grating

Question

1 What is the approximate ratio λ/b for each slit in a diffraction grating that has 500 slits per millimetre?

A5 Resolution

If you stand on a beach and look down at the sand, you probably will not be able to see the separate grains of sand. Similarly, if you look at a tree which is a long way away, you will not be able to see the separate leaves. In scientific terms, we say that you cannot **resolve** the detail. If you assume that a person has good eyesight, the ability of their eyes to resolve detail depends on the diffraction of light as it enters the eye through the **pupil** (aperture) and the separation of the light receptors on the **retina** of the eye. The larger the pupil, the less the diffraction and, so, the better the resolution. You can check this by looking at the world through a *very* small hole made in a piece of paper held in front of your eye (increasing the diffraction of light as it enters your eye).

When discussing resolution, in order to improve understanding, we usually simplify the situation by only considering waves of a single frequency coming from two point sources of light.

How wavelength and aperture width affect resolution

A.5.1 Sketch the variation with angle of diffraction of the relative intensity of light emitted by two point sources that has been diffracted at a single slit.

Figure 15.43 shows an eye looking at two distant identical point objects, O_1 and O_2; θ is called the **angular separation** of the objects. (This is sometimes called the angle **subtended** at the eye.)

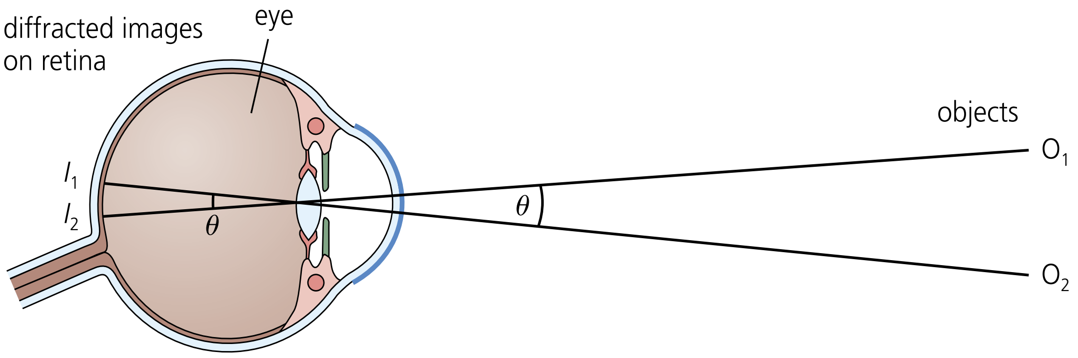


Figure 15.43 An eye receiving light from two separate objects

When the light from each of the objects enters the eye it will be diffracted and two single slit diffraction patterns (similar to that shown in Figure 15.40) will be received by the light receptors on the retina at the back of the eye. So, point objects do not form point images. (Note that the light rays are not shown being refracted by the eye in this example because they are passing through the optical centre of the lens.)

The ability of this eye to resolve two separate images depends on how much the two diffraction patterns overlap (assuming that the receptors in the retina are close enough together). Consider the diagrams in Figure 15.44, which show intensities which might be received by the eye. In **a** the two sources can easily be resolved because their diffraction patterns do not overlap. In **c** the sources are so close together that their diffraction patterns merge together and the eye cannot detect any fall in the resultant intensity between them; the images are not resolved.

Figure 15.44b represents the situation in which the sources can *just* be resolved because, although the images are close, there is a detectable fall in resultant intensity between them.

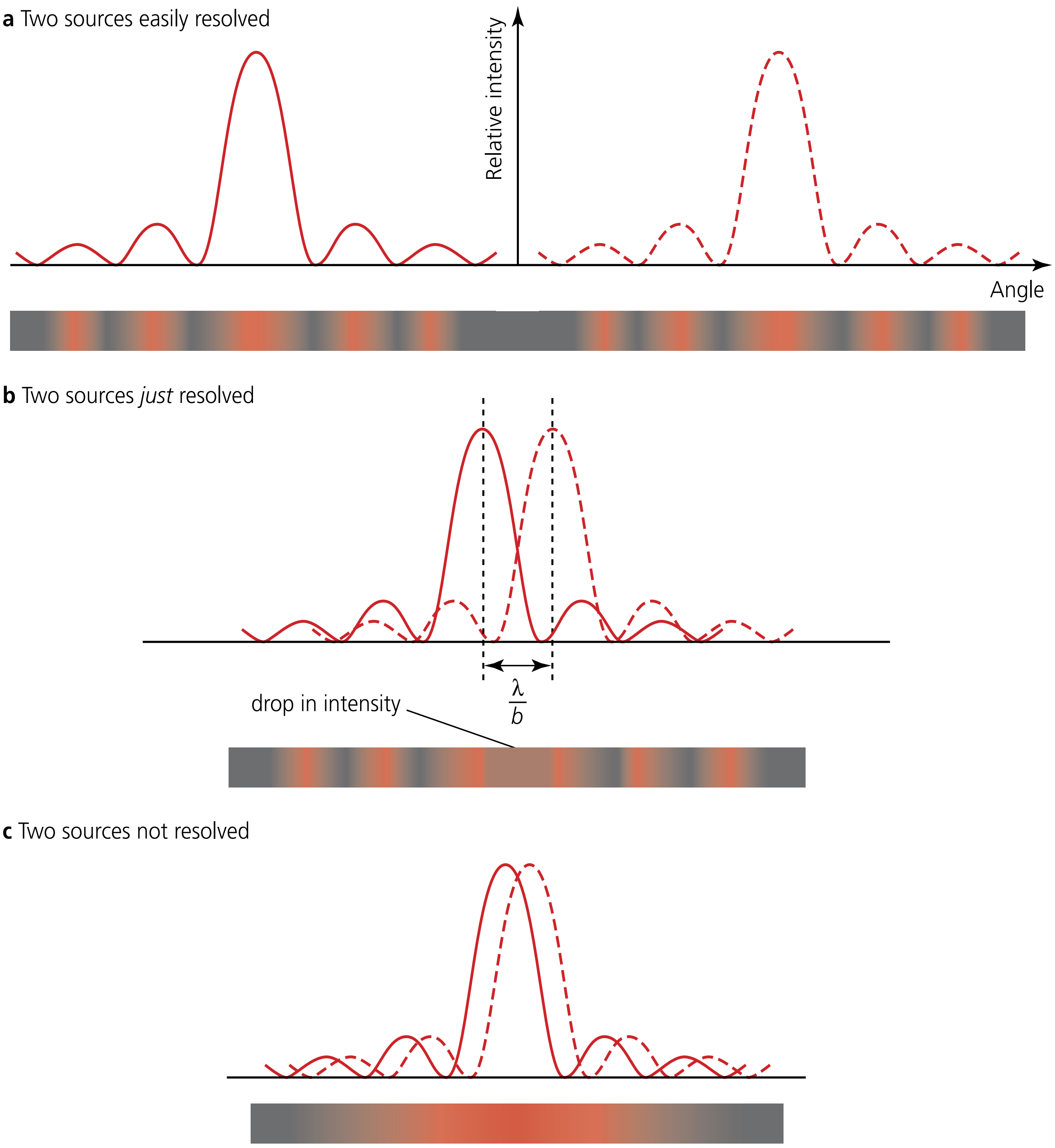


Figure 15.44 Intensities received by the eye as a result of two diffraction patterns

Rayleigh criterion

A.5.2 State the Rayleigh criterion for images of two sources to be just resolved.

Figure 15.45a shows two point sources viewed through circular apertures.

Rayleigh's criterion states that two point sources are *just* resolved if the first minimum of the diffraction pattern of one occurs at the same angle as the central maximum of the other (Figure 15.45b).

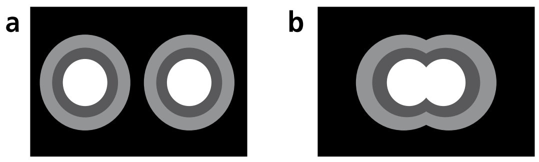


Figure 15.45 Images of two point sources observed through circular apertures that are **a** easily resolvable and **b** just resolvable

Rayleigh's criterion is a useful guide, not a law of physics, and there may be factors other than diffraction which can affect resolution. This can be expressed mathematically, as follows.

The images of two sources can *just* be resolved through a *narrow slit*, of width b, if they have an angular separation of $\theta = \frac{\lambda}{b}$.

If they subtend a larger angle they will be resolvable but if they subtend a smaller angle then they cannot be resolved.

When we consider eyes, telescopes and microscopes it is clear that light waves are usually received and detected through circular apertures, rather than slits. The resolution of a circular aperture will not be as good as with a slit of the same dimensions, so the criterion is adjusted, as follows.

The images of two sources can *just* be resolved through a *circular aperture* if they have an angular separation of $\theta = 1.22\frac{\lambda}{b}$.

This equation is given in the IB *Physics data booklet*.

Worked example

7 Two small point sources of light separated by 1.1 cm are placed 3.6 m away from an observer who has a pupil diameter of 1.9 mm. Can they be seen as separate when the average wavelength of the light used was 5.0×10^{-7} m?

They will be resolvable if their angular separation is greater than or equal to $1.22\lambda/b$.

$$1.22\frac{\lambda}{b} = \frac{1.22 \times 5.0 \times 10^{-7}}{1.9 \times 10^{-3}} = 3.2 \times 10^{-4} \text{ radians}$$

$$\text{Angular separation of sources} = \frac{1.1 \times 10^{-2}}{3.6} = 3.1 \times 10^{-3} \text{ radians}$$

The angular separation is much greater than $1.22\lambda/b$, so they are easily resolvable.

Significance of resolution

Radio telescopes

A.5.3 Describe the significance of resolution in the development of devices such as CDs and DVDs, the electron microscope and radio telescopes.

Stars emit many other kinds of electromagnetic radiation apart from light and infrared. Radio telescopes, like the one shown in Figure 15.46, detect wavelengths much longer than visible light and, therefore, would have very poor resolution if they were not made with big diameters.

Figure 15.46 The Jodrell Bank radio telescope in England, UK

Worked example

8 a The Jodrell Bank radio telescope in the UK has a diameter of 76 m. When used with radio waves of wavelength 21 cm, is it capable of resolving two stars which are 2.7×10^{11} m apart if they are both 1.23×10^{16} m from Earth?

b Compare the ability of this radio telescope to resolve with that of a typical human eye.

a $1.22\frac{\lambda}{b} = \frac{1.22 \times 21 \times 10^{-2}}{76} = 3.4 \times 10^{-3}$ radians

Angular separation of sources $= \frac{2.7 \times 10^{11}}{1.23 \times 10^{16}} = 2.2 \times 10^{-5}$ radians

The angular separation is much less than $1.22\lambda/b$, so they cannot be resolved – they will appear like a single star.

b Using the data from Worked example 7 as an example, a human eye can resolve two objects if their angular separation is about 3×10^{-4} radians or greater, but the radio telescope in this question requires the objects to be at least 3.4×10^{-3} radians apart. This angle is about ten times bigger. The eye has a much better resolution than a radio telescope mainly because it uses waves of a much smaller wavelength. The resolution of radio telescopes can be improved by making them even larger, but there are constructional limits to how big they can be made.

Additional Perspectives

More about radio telescopes

Despite their poor resolution (compared to the human eye), radio telescopes collect a lot of information about the universe that is not available from visible light radiation. Light waves from distant stars are affected when they pass through the Earth's atmosphere and this will decrease the possible resolution. This is why optical telescopes are often placed on mountain tops or on orbiting satellites. Radio telescopes do not have this problem.

Hydrogen is the most common element in the universe and its atoms emit electromagnetic waves with wavelength 21 cm, in the part of the electromagnetic spectrum known as radio waves. Many other similar wavelengths (from a few centimetres to many metres) are received on Earth from space and radio telescopes are designed to detect these waves.

In a dish telescope, such as that shown in Figure 15.46, the radio waves reflect off a parabolic reflector and are focused on a central receiver (like a reflecting optical telescope.) As we have seen, the resolution of a radio telescope will be disappointing unless its dish has a large aperture. Of course, having a large aperture also means that the telescope is able to detect fainter and more distant objects because more energy is received from them when using a larger dish.

There are many different designs of radio telescope, but the largest with a single dish is that built at Arecibo in Puerto Rico (diameter of 305 m), which uses a natural hollow in the ground to help provide support. However, a radio telescope is being built in China with a diameter of 500 m, which is due to be operational in 2013.

In radio *interferometry* signals received from individual telescopes, grouped together in a regular pattern (an **array**, as in Figure 15.47), are combined to produce a superposition (interference) pattern which has a much narrower spacing than the diffraction pattern of each individual telescope. This greatly improves the resolution of the system.

Figure 15.47 An array of linked radio telescopes

Question

1 Find out what you can about China's *FAST* telescope.

Electron microscopes

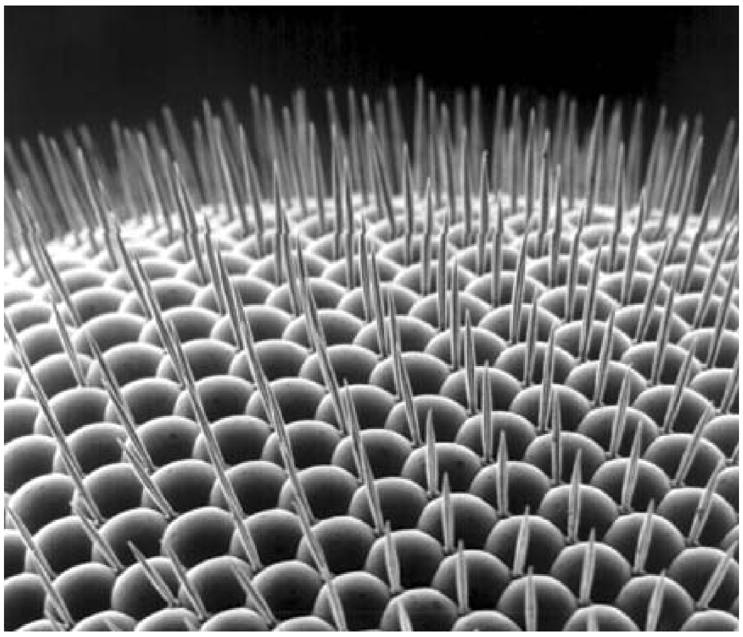


Figure 15.48 Image of an insect eye taken with an electron microscope

If a large enough voltage is used, electrons can be produced which have wavelengths as small as 5×10^{-12} m. This is about 10^5 smaller than visible light waves and makes electron waves ideal for examining very small objects with high magnification and resolution.

The electron waves are detected by a fluorescent screen or sensors for conversion to visible images. The size of atoms is much smaller than the wavelength of visible light, so it will never be possible to see an atom with an optical microscope, but blurry images of individual atoms can be obtained with electron microscopes.

Electronic displays

The number of picture elements (**pixels**) on an electronic display is usually given in a form such as 1366 × 768, meaning that there are 1366 pixels in each horizontal row and 768 in each vertical row, making a total of 1 049 088 pixels, or approximately one megapixel (1 MPx). If the screen size is 41 cm by 23 cm, then the centres of pixels are an average distance of 0.03 cm apart in their rows and columns. The spaces between the pixels are much smaller. (Confusingly, the number of pixels is often called the 'resolution' of the display.) Pixels, digital cameras and the storage of digital data are discussed in more detail in Option C.

When we look at a screen, we do not want to see individual pixels, so the angular separation as seen by a viewer must be smaller than $1.22\lambda/b$. Using 4.0×10^{-7} m as a value for the lowest wavelength produced by the screen, we can calculate a rough guide to the minimum distance at which a viewer (with a pupil of size of 2 mm) would need to be positioned so that they could not resolve individual pixels:

$$\frac{(1.22 \times 4.0 \times 10^{-7})}{(2 \times 10^{-3})} = \frac{\text{pixel separation}}{\text{minimum distance between viewer and screen}}$$

Using a pixel separation of 0.03 mm gives a minimum distance of about 10 cm.

Digital cameras

The quality of the lens and the sizes of the lens and aperture are the most important factors in determining the quality of the images produced by a camera. But, the distance between the pixels on the image sensor must be small enough to resolve the detail provided by the lens. The 'resolutions' of digital cameras are also described in terms of the numbers of (light-receiving) pixels. The settings used on the camera, the way in which the data is processed and the 'resolution' of the display used will all affect the amount of detail resolvable in the final image.

Optical storage of digital information

Information is stored using tiny bumps (called 'lands') on a plastic disc (CD or DVD). The information is 'read' by light from a laser that reflects off the reflective coating on the lands and/or the 'pits' between them.

The closer the lands (and pits) are located to each other, the greater the amount of information that can be stored on the disc. But, if the lands are *too* close together, the diffracting laser beam will not be able to resolve the difference between them. A laser with a shorter wavelength will diffract less and enable more data to be stored on the same sized disc.

On a typical CD (which stores 700 MB of information), the pits are about 8×10^{-6} m in length and they are read by light from a laser of wavelength 7.8×10^{-7} m. DVDs store information in the same way, but they have an improved data capacity; the pits and lands can be closer together because a laser light of shorter wavelength (6.5×10^{-7} m) is used. A single-layer DVD can store about seven times as much information as a CD.

The development of blue lasers, with their smaller wavelengths (4.05×10^{-7} m), has led to Blu-ray technology and greatly increased data storage capacity on discs of the same size. This enables the storage and playback of high definition (HD) video.

The storage of digital information is discussed at length in Option C.

A.5.4 Solve problems involving resolution.

34 a Why might you expect a camera with a lens with a larger diameter to produce better pictures?
b Suggest a reason why bigger lenses might produce poorer images.

35 Why do astronomers sometimes prefer to take photographs with blue filters?

36 The pupils in our eyes dilate (get bigger) when the light intensity decreases. Discuss whether this means that people can see better at night.

37 A car is driving towards an observer from a long way away at night. If the headlights are 1.8 m apart, estimate the maximum distance away at which the observer will see two distinct headlights. (Assume that the average wavelength of light is 5.2×10^{-7} m and the pupil diameter is 4.2 mm.)

38 A camera on a satellite orbiting at a height of 230 km above the Earth is required to take photographs which resolve objects that are 1.0 m apart. Assuming a wavelength of 5.5×10^{-7} m, what minimum diameter lens would be needed?

39 a A radio telescope with a dish of diameter 64 m is used to detect radiation of wavelength 1.4 m. How far away from the Earth are two stars separated by a distance of 3.8×10^{12} m which can just be resolved?
b What assumption(s) did you make?

40 Could an optical telescope with a (objective) lens of diameter 12 cm be used to read the writing on an advertising sign which is 5.4 km away if the letters are an average of 8.5 cm apart? Explain your answer.

41 The Moon is 3.8×10^{8} m from the Earth. Estimate the smallest distance between two features on the Moon's surface that can just be observed from the Earth by the human eye.

42 Use Figure 15.49 to measure the resolution of your own eyes.

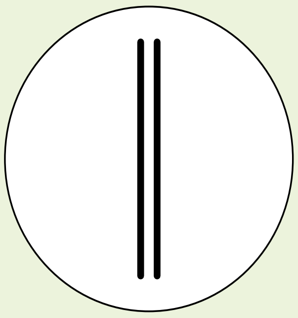


Figure 15.49

TOK Link: Differences between reality and our observations

Figure 15.50 shows a large number of individual pixels, and it may be difficult to decide if this is picture of anything or just dots. But, once we know what the picture is, it becomes instantly recognizable. This is a simple illustration of the fact that, when our brains interpret signals from our eyes (or ears, or other senses), how we interpret what we sense has a lot to do with our past experiences and our expectations.

Optical illusions like that shown in Figure 15.51 trick us because they are unusual experiences and our brain tries hard to make sense of them by interpreting them in terms of more familiar (probably three-dimensional) images. You tend to see what your brain thinks it should be seeing, rather that what is really there.

The expression 'seeing is believing' is far from true, because we cannot trust any of our senses to give us an objective interpretation of the world around us. A person sitting for a long time in a room at 16°C would probably describe the room as cold, but someone coming from outside, where the temperature was 10°C, would certainly think that the room was warm. If we listen to someone speaking a foreign language we may hear no sound that we recognize, but we would immediately recognize our name if it was mentioned.

Of course, this inherent uncertainty in our observations is a very good reason why modern science is based on quantitative measurements, which are much less open to misinterpretation. The earliest scientists (natural philosophers) were restricted by the limitations of qualitative observations because they did not do practical work.

Figure 15.50 A large number of individual dots can be used to create a picture

Figure 15.51 An optical illusion

Question

1 'If a tree falls in a forest and no one is around to hear it, does it make a sound?' is an old philosophical question which raises many issues, including the distinction between what something is and what it appears to be. In what ways do a tree and the sound that it makes exist, if nobody is there to observe them? Should we believe in things of which we have no direct experience? If we believe that a tree has actually fallen and disturbed the surrounding air, but sound is only the name we give to a sensation caused at the human ear, is there any sound in the forest?

A6 Polarization

Polarized waves

Basic wave properties were covered in Chapter 4, but one interesting property of (transverse) waves was not included – **polarization**. As a simple example, consider sending transverse waves along a rope. If your hand only oscillates vertically, the rope will only oscillate vertically and the wave can be described as **plane polarized** because it is only oscillating in one plane (vertical), as shown in Figure 15.52. If your hand only oscillates horizontally it will produce a wave polarized in the horizontal plane.

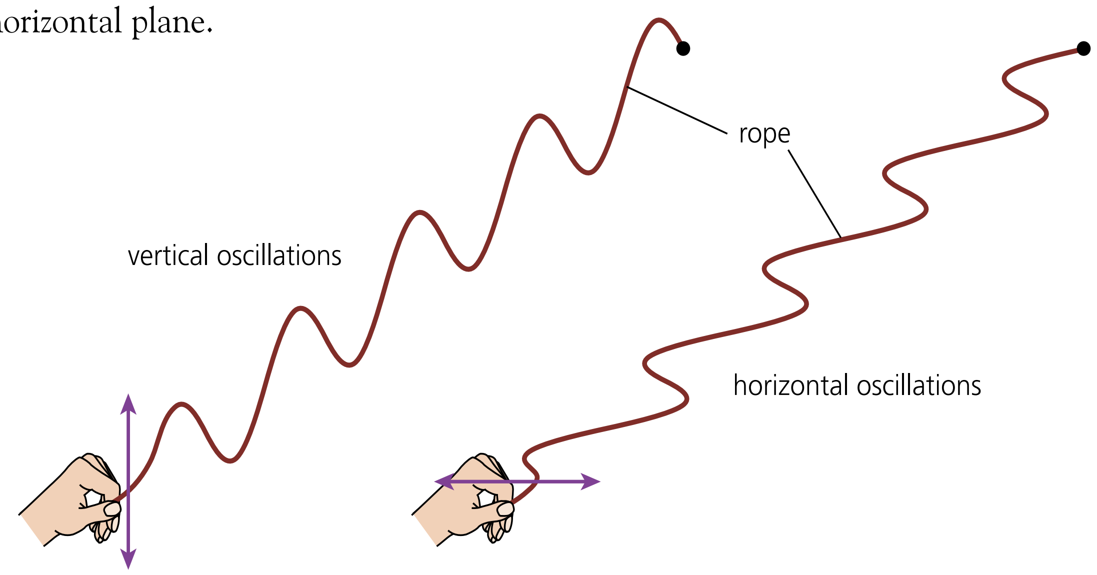


Figure 15.52 Transverse waves on a rope

A transverse wave is (plane) polarized if all the oscillations transferring the wave's energy are in the same plane (called the **plane of polarization**). The plane of polarization must be perpendicular to the direction of wave travel, so it is impossible for longitudinal waves, like sound, to be polarized.

Polarized light and other electromagnetic waves

A.6.1 Describe what is meant by polarized light.

In order to explain the polarization of light, we first need to understand the basic nature of electromagnetic waves. All electromagnetic radiations, including light, are transverse waves of oscillating electric and magnetic fields at right angles to each other. This is shown in Figure 15.53.

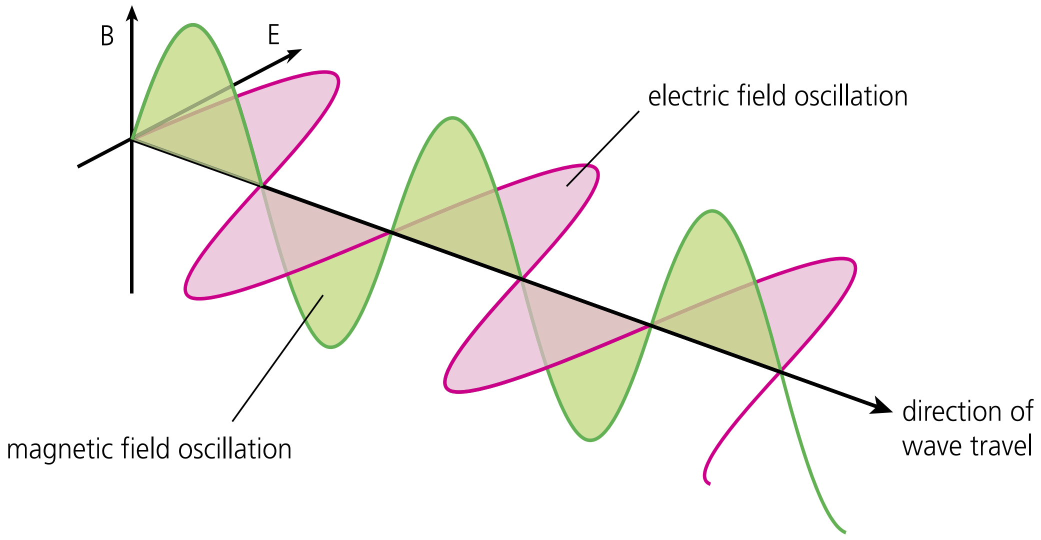


Figure 15.53 Nature of electromagnetic waves

In Figure 15.53 the waves are moving to the right, the electric field oscillations are in the horizontal plane and the magnetic field oscillations are in the vertical plane, but the oscillations could be in any plane which is perpendicular to the direction of wave travel (as long as the electric fields and magnetic fields are perpendicular to each other).

Electromagnetic waves, including light, are usually released during random, unpredictable processes, so we would expect them usually to oscillate in random directions and not be polarized, as shown in Figure 15.54a.

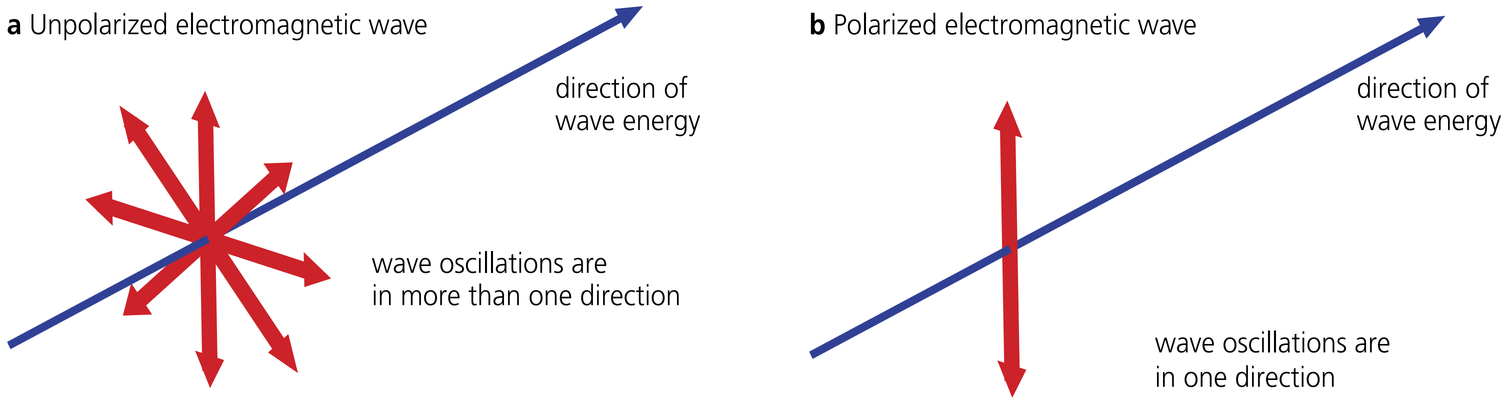


Figure 15.54 **a** unpolarized electromagnetic radiation **b** polarized electromagnetic radiation

Electromagnetic waves, such as light, are said to be (plane) polarized if all the electric field oscillations (or the magnetic field oscillations) are only in one plane, as shown by Figure 15.54b.

Electromagnetic waves which are produced and transmitted by currents oscillating in aerials (for example radio waves and microwaves) will be polarized, with their electric field oscillations parallel to the transmitting aerial.

For example, Figure 15.55 shows the transmission and reception of microwaves. In Figure 15.55a a strong signal is received because the transmitting and receiving aerials are aligned, but in Figure 15.55b no signal is received because the receiving aerial has been rotated through 90°.

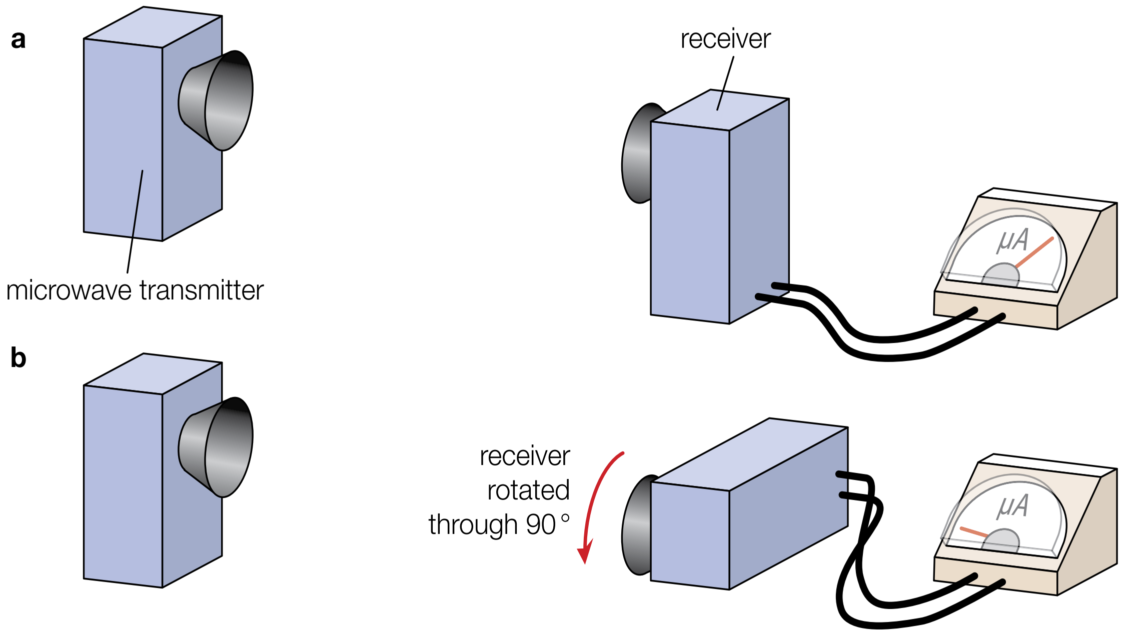


Figure 15.55 The receiver must be aligned in the same plane as the transmitter to detect microwaves

Polarization by absorption

A.6.4 **Explain** the terms polarizer and analyser.

Normal, unpolarized light can be converted into polarized light by passing it through a special filter, called a **polarizer**, which absorbs oscillations in all planes except one.

Polarizing filters are made of long chain molecules mostly aligned in one direction. Components of the electric field parallel to the long molecules are absorbed; components of the electric field perpendicular to the molecules are transmitted. Because of this, we would expect the transmitted intensity to be about half of the incident intensity, as shown in Figure 15.56.

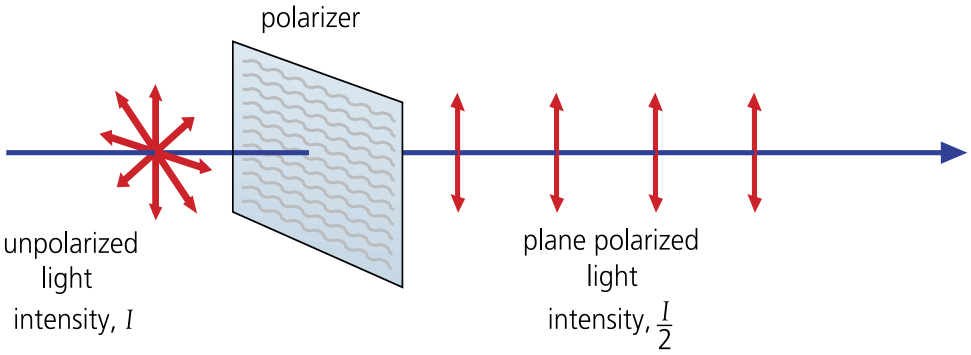


Figure 15.56 Polarizing light using a polarizer (polarizing filter)

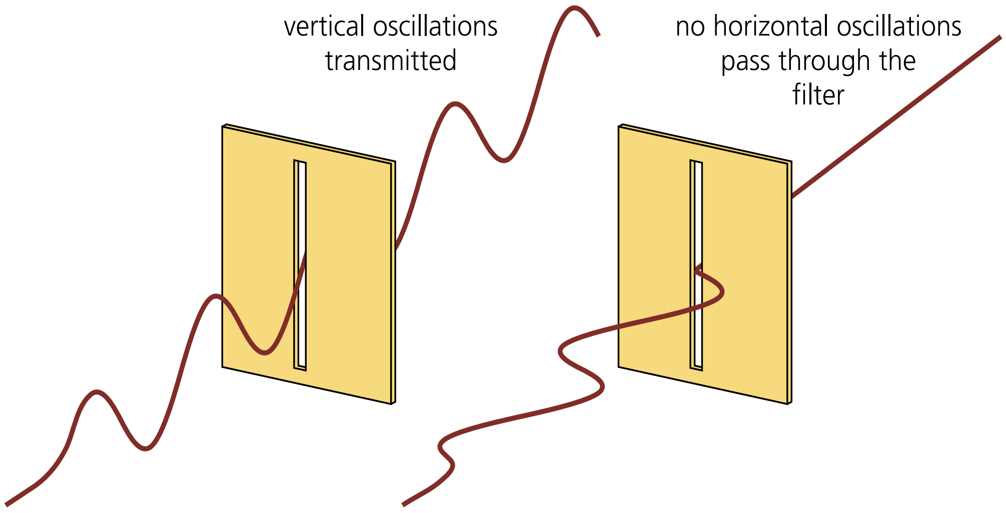


Figure 15.57 A vertical slit acts as a polarizing filter

In order to understand this, it may be helpful to consider how transverse waves made by oscillating a rope (as shown in Figure 15.52) would behave if they had to pass through a vertical slit; see Figure 15.57. Waves oscillating parallel to the slit would pass through, but others would be blocked. The slit acts as a polarizer.

What happens if polarized light is then incident on a second polarizing filter? This depends on how the filters are aligned. If the second filter (called the **analyser**) transmits waves in the same plane as the first (the polarizer), the waves will pass through unaffected, apart from a possible small decrease in intensity. The filters are said to be **in parallel**, as shown in Figure 15.58b. Figure 15.58a shows the situation in which the analyser only allows waves through in a plane which is at right angles to the plane transmitted by the polarizer. The filters are said to be '**crossed**' and no light will be transmitted by the analyser.

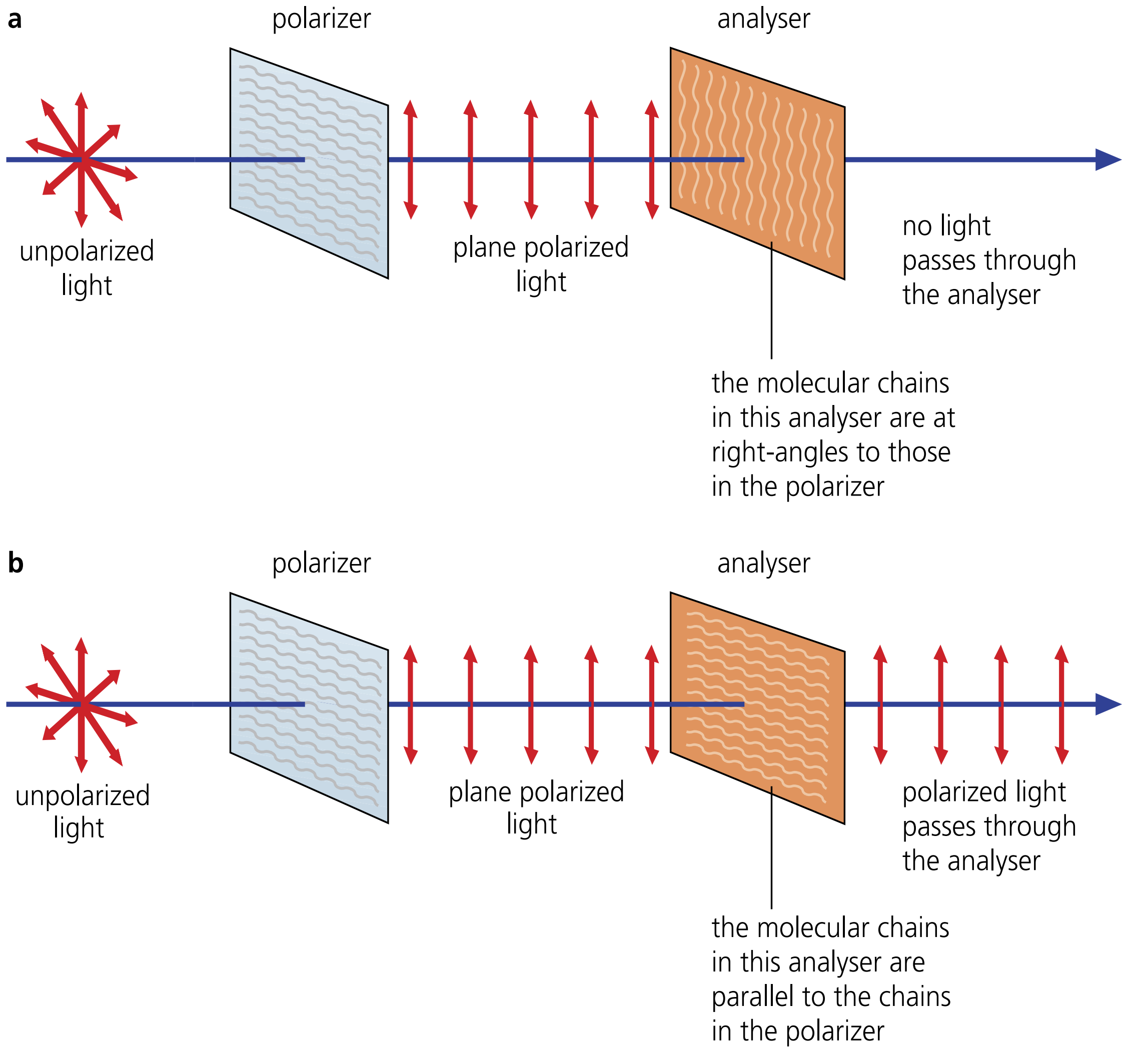


Figure 15.58 Polarizer and analyser: **a** crossed; **b** in parallel

The second filter is called an *analyser* because it can be rotated to analyse light to determine if it is polarized and, if so, in which direction.

If you rotate a polarizing filter in front of your eye, for example, as you look around at the light reflected from various objects, if the intensity changes then the light must be at least partly polarized. The most common type of transparent plastic used for polarizers and analysers is called **Polaroid®**.

Malus' law

A.6.5 Calculate the intensity of a transmitted beam of polarized light using Malus' law.

Figure 15.59 represents a direction (in the plane of the paper) for the electric field oscillations of polarized light which is travelling perpendicularly out of the page. There is an angle θ between the oscillations and the plane in which the polarizer (through which the light is passing) will transmit all of the waves.

If the amplitude of the oscillations incident on the analyser is A_0, then the component in the direction in which waves can be transmitted equals $A_0 \cos\theta$.

We know, from Chapter 4, that the intensity of waves is proportional to amplitude squared, $I \propto A_0^2$, so that the transmitted intensity is represented by the following equation:

$$I = I_0 \cos^2\theta$$

This equation is given in the IB *Physics data booklet.*

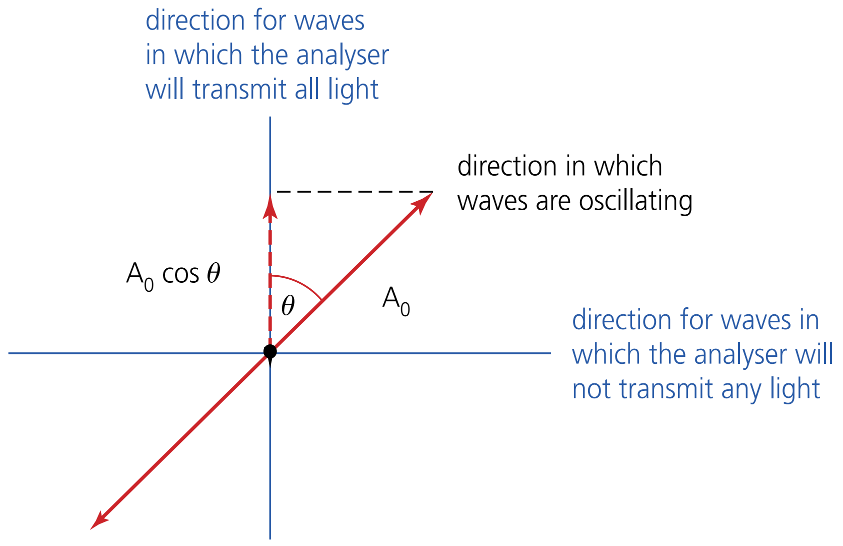


Figure 15.59 Angle between oscillations and the polarizer

Worked example

9 **a** If vertically polarized light falls on a polarizing filter (analyser) which is positioned so that its transmission direction (axis) is at 30° to the vertical, what percentage of light passes through the analyser?
b Repeat the calculation for an angle of 60°.
c What angle would allow 50% of the light to pass?
d Sketch a graph to show how the transmitted intensity would vary if the analyser was rotated through 360°.

a $I = I_0 \cos^2\theta$

$$\frac{I}{I_0} = \cos^2 30°$$

$$\frac{I}{I_0} = 0.75 \text{ or } 75\%$$

b $\frac{I}{I_0} = \cos^2 60° = 0.25 \text{ or } 25\%$

c $\frac{I}{I_0} = 0.50 = \cos^2\theta$

$\cos\theta = \sqrt{0.50} = 0.71$

$\theta = 45°$

d See Figure 15.60. This graph shows the variation with angle of intensity transmitted through a polarizing filter.

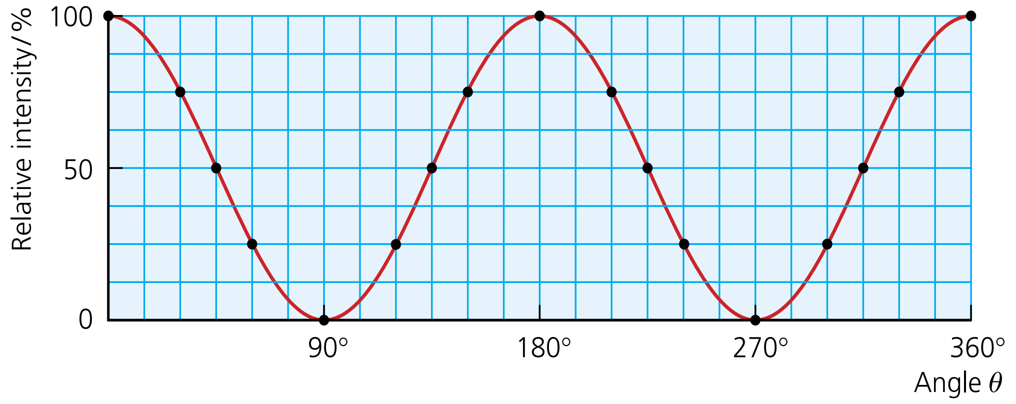


Figure 15.60

Polarization by reflection

A.6.2 **Describe** polarization by reflection.
A.6.3 **State** and **apply** Brewster's law.

When (unpolarized) light reflects off an insulator the waves may become polarized and the plane of polarization will be parallel to the reflecting surface.

Figure 15.61 The same scene with and without Polaroid®

The most common examples of polarization by reflection are the reflection of light from water and from glass. Such reflections are usually unwanted and the reflected light (sometimes called **glare**) entering the eyes can be reduced by wearing **polarizing (Polaroid®) sunglasses**, which also reduce the intensity of unpolarized light by half. Figure 15.61 shows an example. The fish under the water can be seen clearly when Polaroid sunglasses are used. The sunglasses greatly reduce the amount of light reflected off the water's surface entering the eye. But the reduction is not the same for all viewing angles because the amount of polarization depends on the angle of incidence (see below). Photographers may place a rotatable polarizing filter over the lens of their camera to reduce the intensity of reflected light.

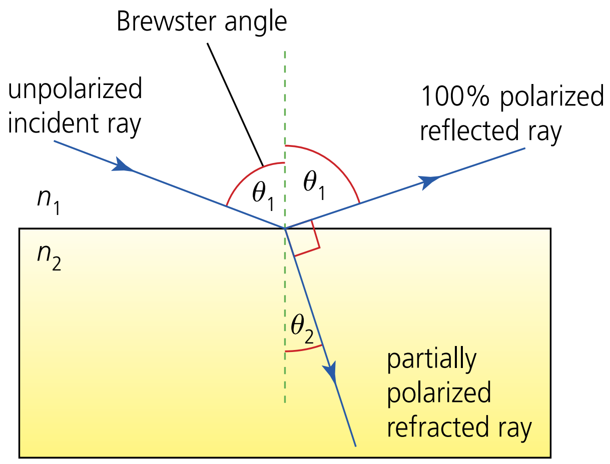


Figure 15.62 Brewster angle

For transparent materials (for example glass or water), at a particular angle, called the **Brewster angle**, all of the reflected light is polarized. This occurs when the refracted rays and the reflected rays are exactly at 90° to each other, as shown in Figure 15.62. (Remember that rays are lines showing the direction in which waves are travelling.) At the Brewster angle, incident light which is polarized in the right direction will be totally transmitted, without any reflection.

From Chapter 4, we know that Snell's law links refractive indices to angles of incidence and refraction as follows:

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1}$$

In Figure 15.62 the angle of incidence is θ_1 in a material of refractive index n_1. The wave is then refracted at an angle θ_2 in a material of refractive index n_2.

Looking at the angles in Figure 15.62 it is clear that:

$$\theta_1 + \theta_2 + 90° = 180°$$

So that,

$$\theta_2 = 90° - \theta_1$$

and

$$\sin\theta_2 = \cos\theta_1$$

Substituting this back into the Snell's law equation, we get:

$$\frac{n_2}{n_1} = \frac{\sin\theta_1}{\cos\theta_1} = \tan\theta_1$$

Usually material 1 will be air, so that $n_1 = 1$. Put into words, this tells us that the tangent of the Brewster angle, ϕ $(= \theta_1)$, is equal to the refractive index, n, of the material that the light is entering.

This is known as **Brewster's law**. It is given in the IB *Physics data booklet* in the form:

$$n = \tan\phi$$

Worked example

10 Calculate the Brewster angle for light passing from air into water (refractive index = 1.33).

$\tan\phi = n = 1.33$

$\phi = 53°$

A.6.10 Solve problems involving the polarization of light.

43 How could you quickly check if sunglasses were made with Polaroid®?

44 Plane polarized light passes through an analyser which has its transmission axis at 75° to the plane of polarization. What percentage of the incident light emerges?

45 When unpolarized light was passed through two polarizing filters only 20% of the incident light emerged. What was the angle between the transmission axes of the two filters?

46 Suggest why the blue light from the sky on a clear, sunny day is partly polarized.

47 **a** What is the Brewster angle for light passing into glass of refractive index 1.53?
b At what angle will light which is polarized parallel to the surface of water of refractive index 1.33 be totally refracted?
c What is the angle of refraction?

48 Use the Internet to find out about polarizing microscopes and their uses.

Applications of polarization

As we have seen, if two polarizing filters are 'crossed', the polarized light which passes through the polarizer cannot pass through the analyser, so no light is transmitted. However, if a transparent material is placed between the two filters it may *rotate the plane of polarization*, allowing some light to be transmitted through the analyser. Figure 15.63 shows a solution of sugar being used between the polarizing filters. This effect has some interesting applications.

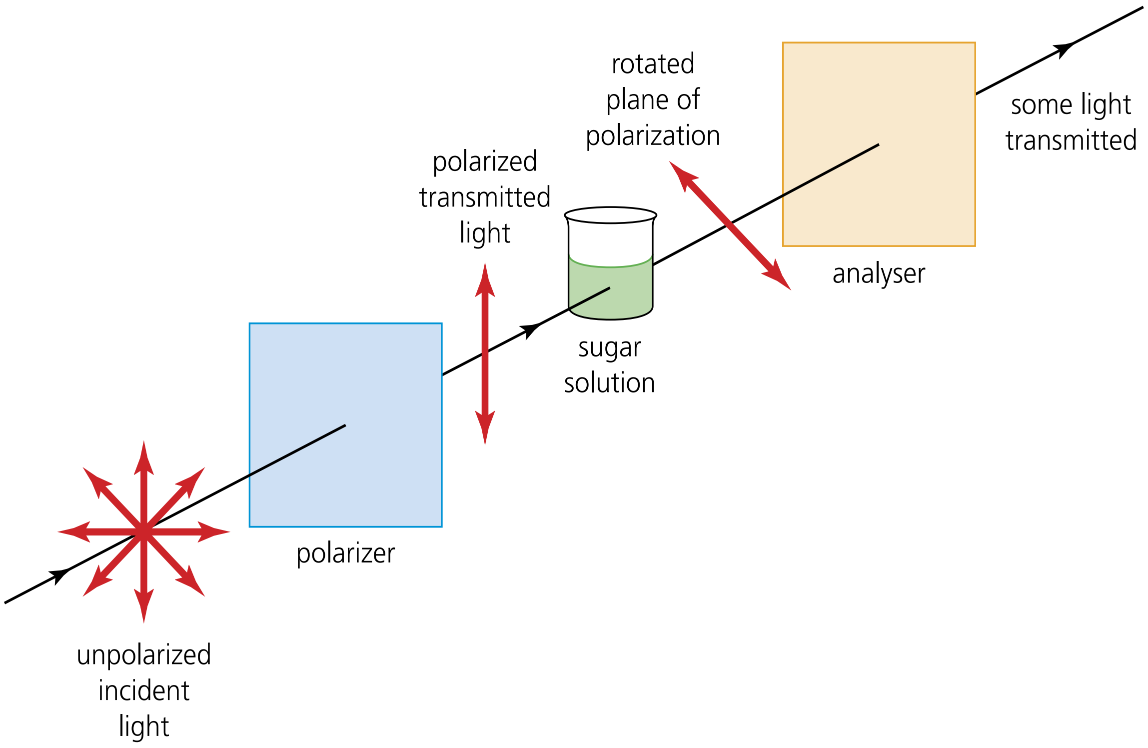


Figure 15.63 Rotating the plane of polarization by placing a transparent material, in this case sugar solution, between the polarizer and the crossed analyser, allowing some light to be transmitted through

Determining the concentration of solutions

A.6.6 **Describe** what is meant by an optically active substance.
A.6.7 **Describe** the use of polarization in the determination of the concentration of certain solutions.

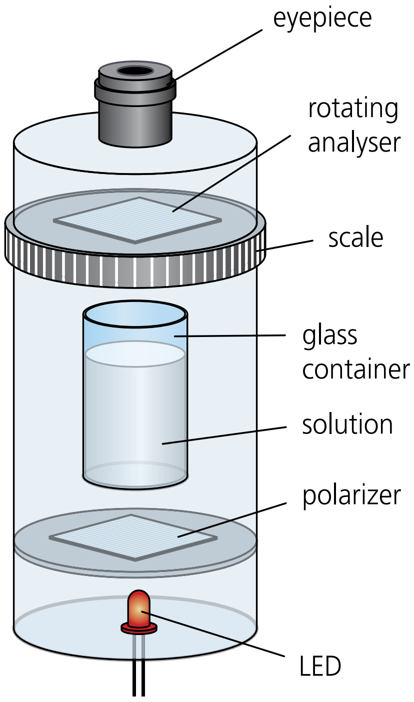


Figure 15.64 Measuring the concentration of a solution with a polarimeter

A substance which rotates the plane of polarization of light waves passing through it is said to be **optically active**.

The angle through which the plane is rotated depends on the substance, the length of the path through the substance, and the wavelength used (and the temperature of the solution). Solutions of some pure compounds are optically active and sugar solutions are a common example (as shown in Figure 15.63). The greater the concentration of the solution, the more the plane of polarization is rotated, so this can be used as a way of experimentally determining the concentration of certain solutions, as in the simple **polarimeter** shown in Figure 15.64.

Using the eyepiece in a polarimeter, the intensity of the light received from the LED at the bottom of the apparatus is observed as the analyser is rotated, and the angle at which no light is received is measured and compared to the result without the solution.

Stress analysis

A.6.8 **Outline** qualitatively how polarization may be used in stress analysis.

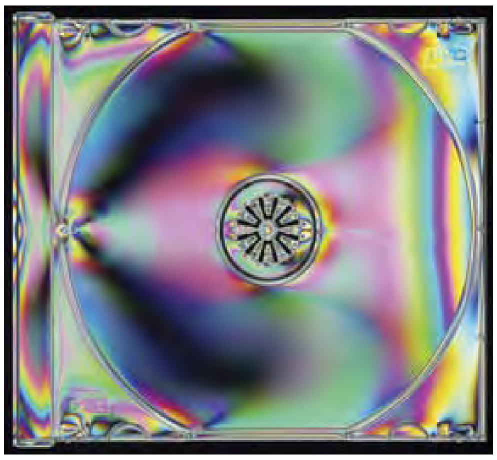


Figure 15.65 Stress concentration in a DVD case seen with polarized light

Some plastics and glasses become optically active and rotate the plane of polarization of light when they are stressed. This can be useful for engineers who can analyse possible concentrations of stress in a structure before it is made.

By building a (scale) model in a transparent material and placing it between polarizing filters, engineers can observe its behaviour under the action of various forces. An example of stress concentration is shown in Figure 15.65. The greater the stress, the greater is the rotation of the plane of polarization. Different colours are seen if white light is used because the effect depends on the wavelength of the light used.

Liquid-crystal displays (LCDs)

A.6.9 **Outline** qualitatively the action of liquid-crystal displays (LCDs).

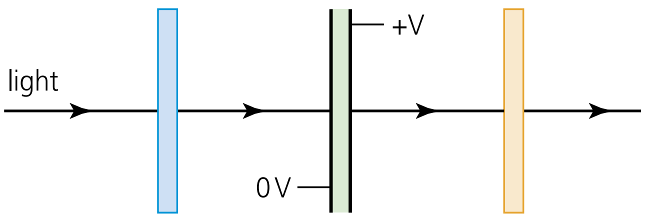


Figure 15.66 Arrangement of parts in a liquid-crystal display

Figure 15.67 A seven segment liquid-crystal display

A liquid crystal is a state of matter that has physical properties between those of a liquid and a solid (crystal). Most interestingly, the ability of certain kinds of liquid crystal to rotate the plane of polarization of light can be changed by applying a small potential difference across them, so that their molecules twist in the electric field. Figure 15.66 represents a simplified arrangement. If there is no potential difference (p.d.) across the liquid crystal no light is transmitted out of the analyser.

When a p.d. is applied to the liquid crystal, its molecules change orientation to align with the electric field and the plane of polarization rotates so that some, or all, of the light is now transmitted. The amount of rotation of the plane of polarization and the amount of light transmitted depend on the size of the p.d.

In simple displays (for example those used on many calculators, digital clocks and watches) light entering through the front of the display passes through the liquid crystals and is then reflected back to the viewer. Each segment of the display will then appear dark or light, depending on whether a p.d. has been applied to the liquid crystal (see Figure 15.67).

Using the same principles, large numbers of liquid crystals are used to make the tiny picture elements (pixels) in many computer and mobile phone displays, and televisions. Colours are created by using filters and the light is provided by a fluorescent lamp behind the display.

A.6.10 **Solve** problems involving the polarization of light.

49 Light of intensity $0.82\,W\,m^{-2}$ is incident on two crossed polarizing filters.
- **a** What is the intensity entering the second filter?
- **b** What intensity emerges from the second filter?
- **c** If a sugar solution placed between the filters rotates the plane of polarization by 28°, what intensity then emerges from the second filter?

50 Use the Internet to research and compare the main advantages and disadvantages of using liquid crystal and LED displays.

Additional Perspectives

3-D cinema

Our eyes and brain see objects in three dimensions (3-D) because, when our two eyes look at the same object, they each receive a slightly different image (because our eyes are not located at exactly the same place). This is known as stereoscopic vision. Our brain merges the two images to give the impression of three dimensions or 'depth'. But when we look at a two-dimensional image in a book or on a screen, both eyes receive essentially the same image.

Figure 15.68 Using polarizing glasses to watch a 3-D movie

If we want to create a 3-D image from a flat screen, we need to provide a different image for each eye and the use of polarized light makes this possible. (Some earlier, and less effective systems used different coloured filters.) In the simplest modern systems one camera is used to take images which are projected in the cinema as vertically polarized waves. At the same time a second camera, located close by, takes images that are later projected as horizontally polarized waves. Sometimes a second image can be generated by a computer program (rather than by a second camera) to give a 3-D effect.

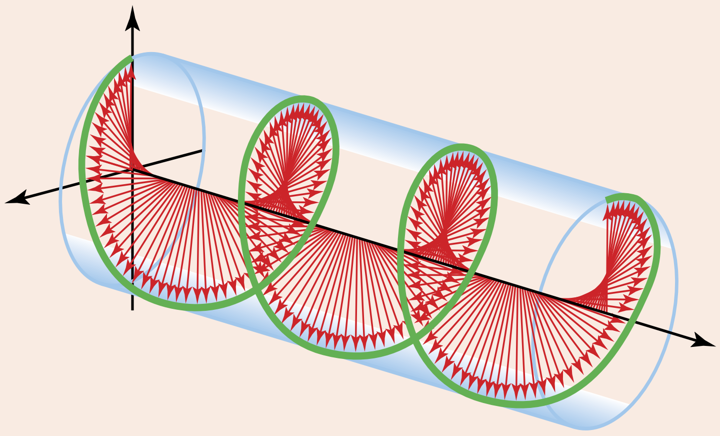


Figure 15.69 Circularly polarized waves

To make sure that each eye receives a different set of images, the viewer wears polarizing glasses, allowing the vertically polarized light into one eye and the horizontally polarized light into the other eye. One problem with using plane polarized waves is that viewers need to keep their heads level, but this can be overcome by using circularly polarized waves in which the direction of the electric field oscillations continually rotates in circles, as in Figure 15.69. A single projector sends images alternating between clockwise and anti-clockwise polarization to the screen.

Question

1 Use the Internet to investigate the latest developments in 3-D television techniques.

SUMMARY OF KNOWLEDGE

A1 The eye and sight

- The structure of the human eye is made up of the cornea, lens, retina, iris, pupil, ciliary muscles, aqueous humour and vitreous humour.
- Ciliary muscles change the shape of the lens in the eye so that light from objects at different distances away is focused correctly on the retina. This process is called accommodation.
- The nearest distance from the eye at which an object can be seen in clear focus (without undue straining) is called the near point. The furthest distance at which objects can be seen in focus is called the far point. The eye muscles are most relaxed when looking at distant objects. The range of normal vision is from 25 cm to infinity.
- It is important that objects at different distances from the eye can still be in acceptable (if not perfect) focus. The range of distances at which focus is acceptable is called depth of vision.
- There are two different kinds of light-sensitive cell in the retina: rods and cones. The cones are concentrated in the middle of the retina (fovea) and they are used under normal and bright lighting conditions. They produce detailed, coloured images. This is called photopic vision.
- The rods are more spread out across the retina (except at the centre) and they produce images without colour and detail. This is called scotopic vision. Rods are useful for detecting peripheral movement and, because they are more sensitive to light, they are useful under poor lighting conditions.
- There are three different types of cone cell, each responding most strongly in a different part of the visible spectrum. When the eye looks at a point on an object of a particular colour, the brain receives three different signals and interprets them together as a single colour.
- Spectral response graphs can be drawn for rods and the three different types of cone cells.
- If one of the types of cone cell is not functioning correctly then the brain will not be able to see some colours normally. Two types not functioning would result in complete colour blindness.
- Red, green and blue light are described as the primary colours. When they are combined in different proportions, all other colours can be produced. This is called colour addition. When combined in equal proportions they produce white light.
- Equal proportions of two primary colours produce the three secondary colours: cyan, magenta and yellow.
- Coloured light can be produced by passing white light through filters and subtracting colours. For example, a yellow filter subtracts mainly blue light.
- Our visual perception of drawings, interior design and buildings is greatly influenced by the effects of light and dark, shadows and colour. Our brains often interpret the images that we see in terms of what we expect to see based on previous experiences.

A2 Standing (stationary) waves

- Standing waves may be set up when waves of the same frequency and amplitude, travelling in opposite directions, combine together. The resultant at any place can be found (at any time) by using the principle of superposition.
- Travelling (progressive) waves move away from their source and transfer energy, but a standing wave pattern stays in the same position and energy is not transferred. There are places (called nodes) along a standing wave where the amplitude is always zero. Between nodes the amplitude of the oscillations varies, but all the oscillations are in phase. Places of maximum amplitude are called antinodes. The wavelength is twice the distance between adjacent nodes (or antinodes).
- It is usually possible for a particular system to vibrate with standing waves of different frequencies: these are called different modes of vibration. The lowest frequency mode is called the fundamental mode, or the first harmonic. The frequency of the fundamental (and other modes) will be dependent on the speed of the wave, the length of the system and whether the ends are fixed ($v = f\lambda$).
- There are various modes of vibration for standing transverse waves on strings and longitudinal sound waves in pipes (open and closed).
- Standing waves are often created by an external source of continuous vibration. If the external frequency is the same as a natural frequency of the system, resonance will occur and the amplitude of the standing wave may increase.

A3 The Doppler effect

- When there is relative motion between the source and the receiver of a wave, the frequency received will not be the same as the frequency emitted. This is known as the Doppler effect. The most common example is the sound heard from a moving vehicle.
- The Doppler effect can be explained by drawing the wavefronts from a moving source.
- The IB *Physics data booklet* provides the equations for the frequency received from a moving source and by a moving observer.
- For electromagnetic waves the change in frequency caused by the Doppler effect can be calculated from $\Delta f = (v/c)f$. This is an approximation which can be used only if the relative speed between source and detector is much less than the speed of electromagnetic radiation (which is true for most calculations).
- The Doppler effect can be used to determine speeds: if waves of a known speed and frequency are directed at a moving object, the change of frequency of the reflected waves can be used to calculate the speed of the object. This is used with ultrasonic waves to measure blood flow rates and with microwaves to measure plane and car speeds.

A4 Diffraction

- Because light has a very small wavelength, the diffraction of light only becomes important for small apertures. It is most easily understood by considering the diffraction of monochromatic light passing through a narrow slit. It is possible, by considering the path differences, to derive the formula $\theta = \lambda/b$ for the position of the first minimum of the diffraction pattern produced at a single slit.
- Graphs can be drawn to show the variation of relative intensity with angle for light diffracted by a single slit.

A5 Resolution

- The angular separation of two objects is equal to the angle they subtend at the observer.
- If we can see two objects that have a small angular separation as being separate, we say that they can be resolved. The ability of our eyes to resolve detail depends on our eyesight (obviously) and the diffraction of light entering the eyes.

- Consider the simplest example: if we look at two close and identical point sources of monochromatic light, then our eyes will receive two identical, overlapping diffraction patterns. Rayleigh's criterion states that the two images are just resolvable if the first diffraction minimum of one occurs at the same angle as the central maximum of the other.
- For light passing through a narrow slit, we can say that the images of two sources can be resolved if they have an angular separation of $\theta = \lambda/b$ or greater. For a circular aperture the resolution is poorer and the criterion becomes $\theta = 1.22(\lambda/b)$.
- Diffraction patterns of images can be sketched and interpreted to show images that are well resolved, just resolved and not resolved.
- To see objects in detail with high resolution, the wavelength used should be as small as possible and the aperture as large as possible. The resolution of optical instruments like telescopes, microscopes and cameras is limited by the wavelength of light, but they can be improved by using larger apertures. Radio telescopes need to be large because of the size of the radio waves that they detect. Electron microscopes achieve high resolution because of the small wavelength of electrons.
- Data is stored on CDs and DVDs using lands and pits on the discs. The closer together the lands and pits, the more data that can be stored on a given disc, but their separation is limited by the diffraction of the laser light used to read the data.

A6 Polarization

- A transverse wave is (plane) polarized if the all the oscillations transferring the wave's energy are in the same plane. Longitudinal waves cannot be polarized.
- Electromagnetic radiations are transverse waves of oscillating electric and magnetic fields at right angles to each other.
- Artificial radio waves and microwaves are polarized because of the controlled way in which they are produced. Other electromagnetic radiations, including light, are usually unpolarized.
- Light can be polarized by passing it through a polarizing filter (called a polarizer) and its intensity will then be halved. If the polarized light is then passed through a second polarizing filter (called an analyser), the intensity of the transmitted light will be zero if the filters are 'crossed', and remains unaltered if the filters are aligned. Malus' law ($I = I_0 \cos^2\theta$) can be used to calculate the transmitted intensity. A sketch graph provides a simple way to show how intensity varies with angle.
- An optically active substance rotates the plane of polarization of light passing through it. This effect can be used in stress analysis and in the determination of the concentration of certain solutions.
- Light which is reflected from the surface of an insulator (like glass or water) may become polarized. This can be checked by observing the reflection through a rotating polarizing filter. Polarized sunglasses are used to reduce the amount of reflected light entering the eyes.
- For light entering transparent materials at a particular angle, called the Brewster angle, ϕ, all of the reflected light is polarized. This occurs when the refracted rays and the reflected rays are at 90° to each other. Brewster's law relates this angle to the refractive index, n, of the material: $n = \tan\phi$.
- The plane of polarization of a liquid crystal can be rotated by applying a p.d. across it. A liquid-crystal display (LCD) contains individual pixels, each of which is connected to a separate p.d., so that the amount of light passing through it can be controlled.

Examination questions – a selection

All of the IB questions and IB style questions from Papers 1 and 2 which are to be found at the end of Chapter 11 are suitable for the revision of Option A, although the actual option examination paper (Paper 3) does *not* contain any multiple choice type questions.

Paper 3 IB questions and IB style questions

Q1 a Explain the terms
- **i** *accommodation* and
- **ii** *depth of vision*

as applied to the human eye.

b Describe, in simple terms, the changes that occur in the human eye in order to keep in sharp focus an object which is moving further away.

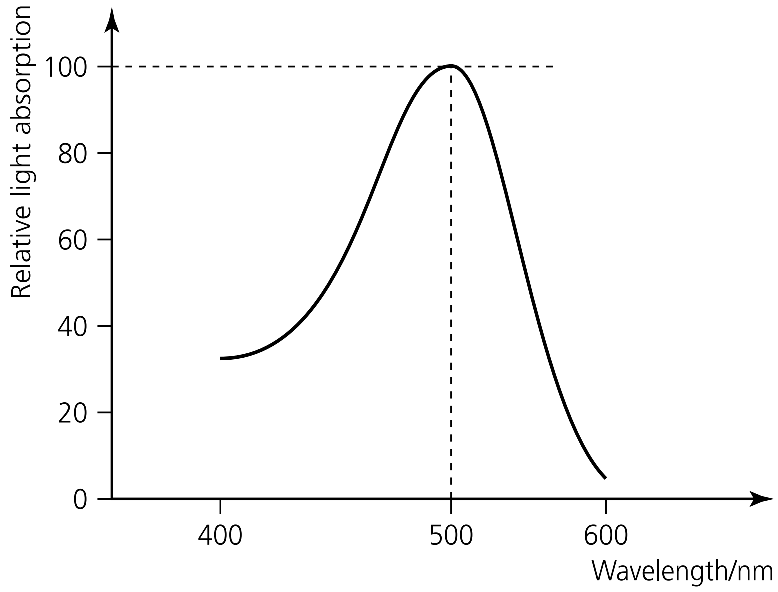
Q2 Red and blue light are shone onto a white surface in a darkened room.

a What colour will the surface appear to be?

b What assumption did you make?

c Use the idea of colour subtraction to suggest how an optical filter can transmit yellow light.

Q3 The graph shows the overall relative light absorption curve for the light-sensitive cells involved in scotopic vision. The relative light absorption is expressed as a percentage of the maximum.



a State the name of the cells involved in **scotopic** vision. [1]

b i On a copy of the graph, sketch a relative light absorption curve for a cell involved in **photopic** vision. [2]

ii State the colour to which the cell is most sensitive. [1]

c Outline how colour blindness may arise from defects in the retina's light-sensitive cells. [3]

Standard Level Paper 3, Specimen Paper 09, QA1

Option B

Chapter 16 Quantum physics and nuclear physics

STARTING POINTS

- According to the principle of conservation of energy, the total energy in an isolated system remains constant.
- Electric charge can be either positive or negative.
- Electric charge is measured in coulombs.
- Like charges repel; unlike charges attract each other.
- The variation with distance of the size of forces between charges follows an inverse square law.
- An electric field is a region of space where a charge experiences a force.
- Using wave theory, the energy of electromagnetic radiation is assumed to propagate in the form of continuous transverse waves.
- Using the 'particle' model electromagnetic radiation travels in the form of photons.
- The energy of a photon is determined by its frequency.
- Some large nuclei are unstable and decay to form more stable nuclei. Ionizing radiation is released during this process.
- The scattering of alpha particles from thin metal foils provides evidence for the nuclear model of the atom.

B1 Quantum physics

In 1900 it was believed that physics was almost fully understood, although there were a few knowledge 'gaps', such as the nature of atoms and molecules and the ways in which radiation interacted with matter. Many physicists believed that filling in these gaps was unlikely to involve any new theories. However, within a few years the 'small gaps' were seen to be fundamental, and a radically new theory was required. One important discovery was that energy only came in certain *discrete* (separate) amounts known as **quanta** (singular: **quantum**). The implications of this discovery were enormous and collectively they are known as *quantum physics*.

As quantum physics developed, some classical concepts had to be abandoned. There was no longer a clear distinction between particles and waves. The most fundamental change was the discovery that systems change in ways which cannot be predicted precisely; only the probability of events can be predicted.

The quantum nature of radiation

Photoelectric effect

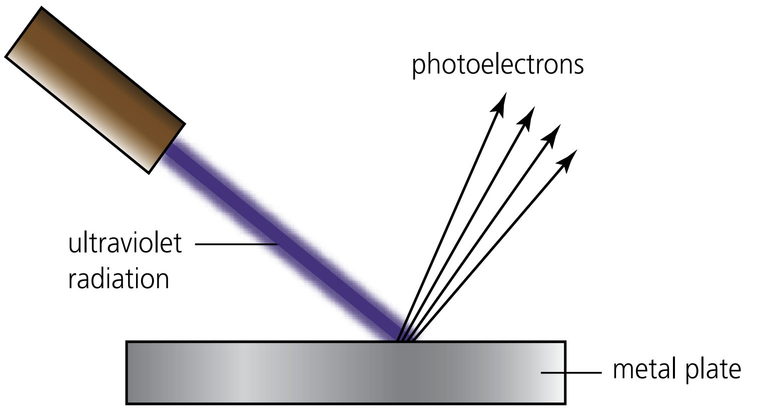


Figure 16.1 The photoelectric effect – a stream of photoelectrons is emitted from a metal surface illuminated with ultraviolet radiation

B.1.1 **Describe** the photoelectric effect.

When electromagnetic radiation is directed onto a clean surface of some metals, electrons are ejected. This is called the **photoelectric effect** (Figure 16.1) and the ejected electrons are known as **photoelectrons**. Under suitable circumstances the photoelectric effect can occur with visible light, X-rays and gamma rays, but it is most often demonstrated with ultraviolet radiation and zinc. Figure 16.2 on page 595 shows a typical arrangement.

Ultraviolet radiation is shone onto a zinc plate attached to a digital coulombmeter (which measures very small quantities

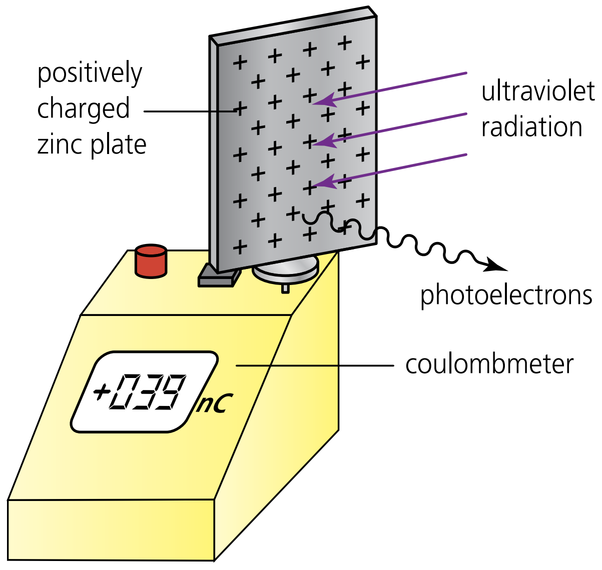


Figure 16.2 Demonstration of the photoelectric effect

of charge). The ultraviolet radiation causes the zinc plate to become positively charged because some negatively charged electrons on the (previously neutral) zinc plate have gained enough kinetic energy to escape from the surface.

Simple investigations of the photoelectric effect show a number of key features.

- If the intensity of the radiation is increased the charge on the plate increases more quickly because more photoelectrons are being released every second.
- There is no time delay between the radiation reaching the metal surface and the emission of photoelectrons. The release of photoelectrons from the surface is *instantaneous*.
- The photoelectric effect can only occur if the frequency of the radiation is above a certain *minimum* value. The lowest frequency for emission is called the **threshold frequency**, f_0. (Alternatively we could say that there is *maximum* wavelength above which the effect will not occur.) If the frequency used is below the threshold frequency, the effect will not occur, *even if the intensity of the radiation is greatly increased.* The threshold frequency of zinc, for example, is 1.04×10^{15} Hz, which is in the ultraviolet part of the spectrum. Visible light will not release photoelectrons from zinc (or other common metals).
- For a given incident frequency the photoelectric effect occurs with some metals, but not with others. This because different metals have different threshold frequencies.

Explaining the photoelectric effect

B.1.2 **Describe** the concept of the photon, and use it to **explain** the photoelectric effect.

If we use the wave theory of radiation to make predictions about the photoelectric effect, we would expect the following. (1) Radiation of *any* frequency will cause the photoelectric effect if the intensity is made high enough. (2) There may be a delay before the effect begins because it needs time for enough energy to be provided (like heating water until it boils).

These predictions are *wrong*, so an alternative theory is needed. Einstein realized that we cannot explain the photoelectric effect without first understanding the quantum nature of radiation.

Photons

The German physicist Max Planck was the first to propose (in 1900) that the energy transferred by electromagnetic radiation was in the form of a very large number of separate, individual amounts of energy (rather than continuous waves). These discrete 'packets' of energy are called quanta. Quanta are also commonly called **photons**. (The concept of photons was introduced in Chapter 7.)

This very important theory, developed further by Albert Einstein in the following years, essentially describes the nature of electromagnetic radiation as being *particles*, rather than *waves*. ('Wave–particle duality' is discussed later in this chapter.) Explaining the photoelectric effect played a central role in the early development of the photon theory.

The energy, E, carried by each photon (quantum) is given by the following relationship:

$$E = hf$$

This equation is in the IB *Physics data booklet*.

where f represents the frequency of the electromagnetic radiation and h represents a constant called **Planck's constant** (Chapter 7).

The value of Planck's constant is 6.63×10^{-34} J s. (A value for this *fundamental constant* is given in the IB *Physics data booklet*.)

Since $c = f\lambda$, this equation is also commonly expressed in term of wavelength as follows:

$$E = \frac{hc}{\lambda}$$

Photons travel at the speed of light and have zero rest mass. However, photons can transfer energy *and* momentum during interactions with subatomic particles, which indicates that photons have momentum themselves.

Using the wave theory of radiation, greater intensity is explained by waves of greater amplitude. Using a photon model of radiation, greater intensity is explained simply by having more photons (of the same energy).

We know that different parts of the electromagnetic spectrum have very different properties. This can be mostly explained by understanding that the amount of energy carried by photons can vary enormously. For example, a gamma ray photon has a frequency about one million times greater than a visible light photon. Therefore, gamma ray photons each transfer one million times as much energy as light photons, so when each individual photon is absorbed it can be dangerous to living organisms.

Within the visible spectrum, blue light has a higher frequency than red light and each of its photons carries more energy than photons of red light.

Worked examples

1 Determine the energy of a photon of light with a wavelength of 500 nm in joules and in electronvolts.

$$E = \frac{hc}{\lambda}$$

$$E = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{500 \times 10^{-9}} = 3.98 \times 10^{-19}\,\text{J}$$

$$E = \frac{3.98 \times 10^{-19}}{1.60 \times 10^{-19}} = 2.49\,\text{eV}$$

2 Electrical power of 60 watts is supplied to a lamp which radiates light of wavelength 590 nm. The lamp is 30% efficient. Determine the number of photons of visible light emitted per second by the lamp.

$$E = \frac{hc}{\lambda}$$

$$E = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{590 \times 10^{-9}} = 3.37 \times 10^{-19}\,\text{J}$$

$$\text{Number of photons emitted per second} = \frac{\text{useful power}}{\text{energy}} = \frac{0.30 \times 60}{3.37 \times 10^{-19}} = 5.3 \times 10^{19}$$

1 Calculate the number of photons emitted every second from a mobile phone operating at a frequency of 850 MHz and at a radiated power of 780 mW.

2 What is the approximate ratio of the energy of a photon of blue light to the energy of a photon of red light?

3 A detector of very low intensity light receives a total of 3.32×10^{-17} J from light of wavelength 600 nm. Calculate the number of photons received by the detector.

4 **a** Calculate an approximate value for the energy of an X-ray photon in joules and electronvolts.
b Suggest a reason why exposure to X-rays of low intensity for a short time is dangerous, but relatively high continuous intensities of visible light cause us no harm.

Einstein model

The Einstein model explains the photoelectric effect using the concept of photons. When a photon in the incident radiation interacts with an electron in the metal surface, it transfers *all* of its energy to that electron. It should be stressed that a *single* photon can only interact with a *single* electron and this transfer of energy is *instantaneous*; there is no need to wait for a build-up of energy. If a photoelectric effect is occurring, increasing the intensity of the radiation only increases the number of photoelectrons, not their energies.

Einstein realized that some of the energy carried by the photon (and then given to the electron) was used to overcome the attractive forces that normally keep an electron within the metal surface. The remaining energy is transferred to the kinetic energy of the newly released (photo) electron. Using the principle of conservation of energy, we can write:

energy carried by photon = work done in removing the electron from the surface + kinetic energy of photo electron

But, the energy required to remove different electrons is *not* always the same. It will vary with the position of the electron with respect to the surface. Electrons closer to the surface will require less energy to remove them. However, there is a well-defined *minimum* amount of energy needed to remove an electron, and this is called the **work function** of the metal surface. The symbol ϕ is used for work function. Different metals have different values for their work function. For example, the work function of a clean zinc surface is 4.3 eV. This means that *at least* 4.3 eV ($= 6.9 \times 10^{-19}$ J) of work has to be done to remove an electron from zinc.

To understand the photoelectric effect we need to compare the photon's energy, hf, to the work function, ϕ, of the metal:

- $hf < \phi$ If an incident photon has less energy than the work function of the metal, the photoelectric effect cannot occur. This means that the same radiation may cause the photoelectric effect with one metal, but not with another (which has a different work function).
- $hf_0 = \phi$ At the *threshold frequency*, f_0, the incident photon has exactly the same energy as the work function of the metal. We may assume that the photoelectric effect occurs, but the released photon will have zero kinetic energy.
- $hf > \phi$ If an incident photon has more energy than the work function of the metal, the photoelectric effect occurs and a photoelectron will be released. Photoelectrons produced by different photons (of the same frequency) will have a range of different kinetic energies because different amounts of work will have been done to release them.

It is important to consider the situation in which the *minimum* amount of work is done to remove an electron (equal to the work function):

energy carried by photon = work function + *maximum* kinetic energy of photoelectron

Or in symbols:

$$hf = \phi + E_{max}$$

This equation is often called Einstein's photoelectric equation, and it is given in the IB *Physics data booklet*. Since $hf_0 = \phi$, we can also write this as:

$$hf = hf_0 + E_{max}$$

Figure 16.3 shows a graphical representation of how the maximum kinetic energy of the emitted photons varies with the frequency of the incident photons. The equation of the line is $E_{max} = hf - \phi$.

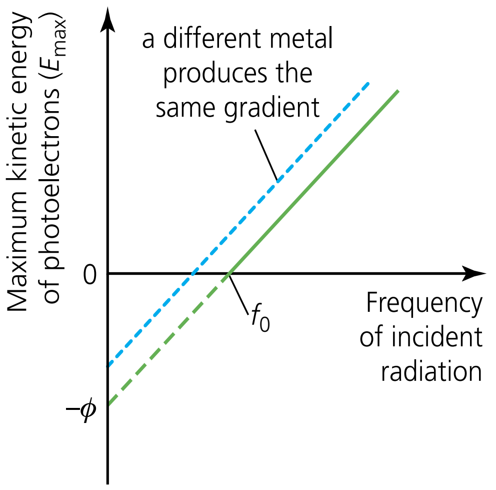


Figure 16.3 Theoretical variation of maximum kinetic energy of photoelectrons with incident frequency (for two different metals)

We can take the following measurements from this graph:

- The gradient of the line is equal to Planck's constant, h. (Compare the equation of the line to $y = mx + c$.) Clearly the gradient is the same for all circumstances because it does not depend on the frequency or the metal used.
- The intercept on the frequency axis gives us the value of the threshold frequency, f_0.
- A value for the work function can be determined from:
  i when $E_{max} = 0$, $\phi = hf_0$; or
  ii when $f = 0$, $\phi = -E_{max}$

Worked example

3 Radiation of wavelength 5.59×10^{-8} m was incident on a metal surface which had a work function of 2.71 eV.

a What was the frequency of the radiation?
b How much energy is carried by one photon of the radiation?
c What is the value of the work function expressed in joules?
d Did the photoelectric effect occur under these circumstances?
e What was the maximum kinetic energy of the photoelectrons?
f What is the threshold frequency for this metal?
g Sketch a fully labelled graph to show how the maximum kinetic energy of the photoelectrons would change if the frequency of the incident radiation was varied.

a $f = \frac{c}{\lambda}$

$f = \frac{3.00 \times 10^8}{5.59 \times 10^{-8}}$

$f = 5.37 \times 10^{15}$ Hz

b $E = hf$

$E = (6.63 \times 10^{-34}) \times (5.37 \times 10^{15})$

$E = 3.56 \times 10^{-18}$ J

c $2.71 \times (1.60 \times 10^{-19}) = 4.34 \times 10^{-19}$ J

d Yes, because the energy of each photon is greater than the work function.

e $E_{max} = hf - \phi$

$E_{max} = (3.56 \times 10^{-18}) - (4.34 \times 10^{-19}) = 3.13 \times 10^{-18}$ J

f $\phi = hf_0$

$f_0 = \frac{\phi}{h}$

$f_0 = \frac{4.34 \times 10^{-19}}{6.63 \times 10^{-34}} = 6.54 \times 10^{14}$ Hz

g The graph should be similar to Figure 16.3, with numerical values provided for the intercepts.

B.1.4 Solve problems involving the photoelectric effect.

5 Repeat Worked example 3 above, but for radiation of wavelength 6.11×10^{-7} m incident upon a metal with a work function of 2.21 eV. Omit part **e**.

6 a Explain how Einstein used the concept of photons to explain the photoelectric effect.
b Explain why a wave model of electromagnetic radiation is unable to explain the photoelectric effect.

7 The threshold frequency of a metal is 7.0×10^{14} Hz. Calculate the maximum kinetic energy of the electrons emitted when the frequency of the radiation incident on the metal is 1.0×10^{15} Hz.

8 a The longest wavelength that emits photoelectrons from potassium is 550 nm. Calculate the work function (in joules).
b What is the threshold wavelength for potassium? What is the name for this kind of radiation?
c Name one colour of visible light that will *not* produce the photoelectric effect with potassium.

9 When electromagnetic radiation of frequency 2.90×10^{15} Hz is incident upon a metal surface the emitted photoelectrons have a maximum kinetic energy of 9.70×10^{-19} J. Calculate the threshold frequency of the metal.

Experiments to test the Einstein model

B.1.3 **Describe** and **explain** an experiment to test the Einstein model.

Investigating stopping potentials

To test Einstein's equation (model) for the photoelectric effect, it is necessary to measure the maximum kinetic energy of the photoelectrons emitted under a variety of different circumstances. In order to do this, the kinetic energy must be transferred to another (measurable) form of energy.

The kinetic energy of the photoelectrons can be transferred to electric potential energy if they are repelled by a negative voltage (potential). This experiment was first performed by the American physicist Robert Millikan and a simplified version is shown in Figure 16.4.

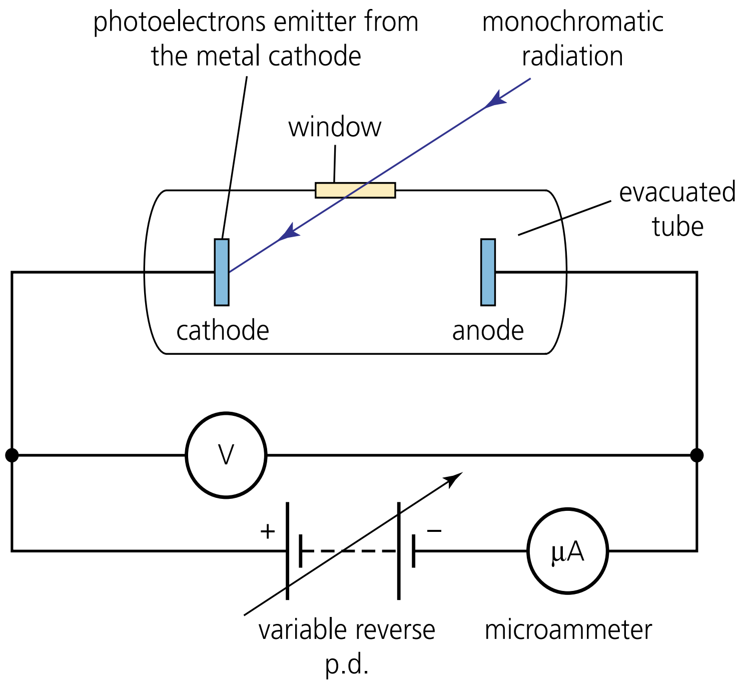


Figure 16.4 Experiment to test Einstein's model of photoelectricity

Ideally *monochromatic* radiations should used, but it is also possible to use a narrow range of frequencies, such as those obtained by using coloured filters with white light.

When radiation is incident upon a suitable emitting surface, photoelectrons will be released with a range of different energies (as explained in the previous section). Because it is emitting negative charge, this surface can be described as a **cathode**. Any photoelectrons which have enough kinetic energy will be able to move across the tube and reach the other electrode, the **anode**. The tube is *evacuated* (the air is removed to create a vacuum), so that the electrons do not collide with air molecules during their movement across the tube.

The most important thing to note about this circuit is that the (variable) source of p.d. is connected the 'wrong way around', it is supplying a **reverse potential difference** across the tube. This means that there is a negative voltage (potential) on the anode which will *repel* the photoelectrons. Photoelectrons moving towards the anode will have their kinetic energy reduced as it is transferred to electric potential energy. (Measurements for positive potential differences can be made by reconnecting the battery around the 'correct' way.)

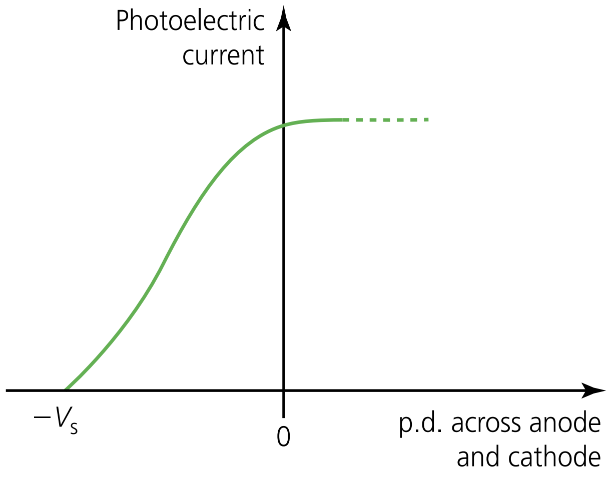


Figure 16.5 Increasing the reverse p.d. decreases the photoelectric current

Any flow of charge across the tube and around the circuit can be measured by a sensitive ammeter (microammeter or picoammeter). When the reverse potential on the anode is increased from zero, more and more photoelectrons will be prevented from reaching the anode and this will decrease the current. (Remember that the photoelectrons have a range of different energies.) Eventually the potential will be great enough to stop even the most energetic of photoelectrons, and the current will fall to zero (Figure 16.5)

The potential on the anode needed to just stop all photoelectrons reaching it is called the **stopping potential**, V_s.

Since, by definition, potential difference = energy transferred/charge,

$$\text{stopping potential, } V_s = \frac{\text{maximum kinetic energy of photoelectrons, } E_{max}}{\text{charge on electron, } e}$$

After measuring V_s we can use this equation to calculate values for the maximum kinetic energy of photoelectrons under a range of different circumstances:

$$E_{max} = eV_s$$

For convenience, it is common to quote all energies associated with the photoelectric effect in electronvolts (eV). In this case, the maximum kinetic energy of the photoelectrons is numerically equal to the stopping potential. That is, if the stopping potential is 3 V, then $E_{max} = 3\,eV$.

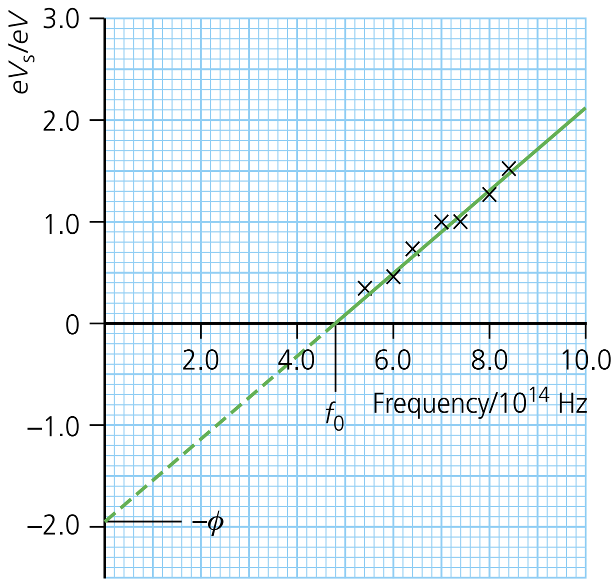


Figure 16.6 Experimental results showing variation of maximum potential energy (eV_s) of photoelectrons with incident frequency

Einstein's equation ($hf = \phi + E_{max}$) can now be re-written as:

$$hf = \phi + eV_s$$

Or, since $\phi = hf_0$,

$$hf = hf_0 + eV_s$$

This equation is given in the IB *Physics data booklet* (although V is used instead of V_s).

By *experimentally* determining the stopping potential for a range of different frequencies the *theory* shown previously in Figure 16.3 can now be confirmed by plotting a graph from actual data, as shown in Figure 16.6.

The threshold frequency, f_0, can be determined from the intercept on the frequency axis.

The work function, ϕ, can be calculated from $\phi = hf_0$, or from the intercept on the eV_s axis.

A value for Planck's constant, h, can be determined from the gradient.

B.1.4 Solve problems involving the photoelectric effect.

10 Calculate the maximum kinetic energy of photoelectrons emitted from a metal if the stopping potential was 2.4 eV. Give your answer in joules and in electronvolts.

11 Make a copy of Figure 16.5 and add lines to show the results that would be obtained with:
a the same radiation, but with a metal of greater work function
b the original metal and the same frequency of radiation, but using radiation with a greater intensity.

12 In an experiment using monochromatic radiation of frequency 7.93×10^{14} Hz with a metal that had a threshold frequency of 6.11×10^{14} Hz, it was found that the stopping potential was 0.775 V. Calculate a value for Plank's constant from these results.

Investigating photoelectric currents

Using the apparatus shown in Figure 16.4, it is also possible to investigate quantitatively the effects on the photoelectric current of changing the intensity, the frequency, and the metal used in the cathode.

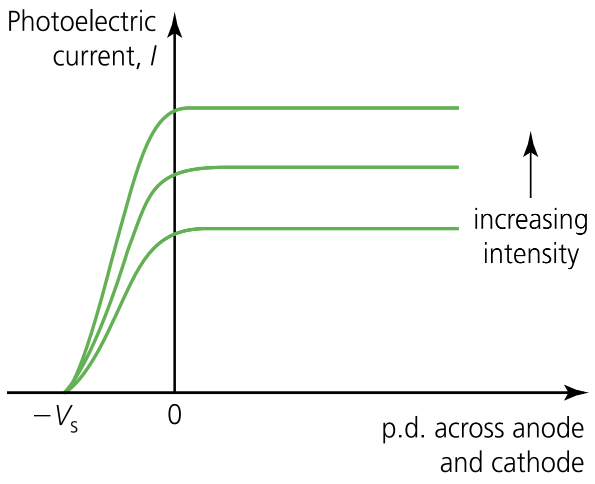


Figure 16.7 Variation of photoelectric current with p.d. for radiation of three different intensities (same frequency)

- **Intensity** – Figure 16.7 shows the photoelectric currents produced by monochromatic radiation of the same frequency at three different intensities.

  For positive potentials, the photoelectric currents remain constant because the photoelectrons are reaching the anode at the same rate as they are being produced at the cathode, and this does not depend on the size of the positive potential on the anode.

  Greater intensities (of the same frequency) produce greater photoelectric currents because there are more photons releasing more photoelectrons (of the same range of energies).

  Since the maximum kinetic energy of photons depends only on frequency and not intensity, all of these graphs have the same value for stopping potential, V_s.

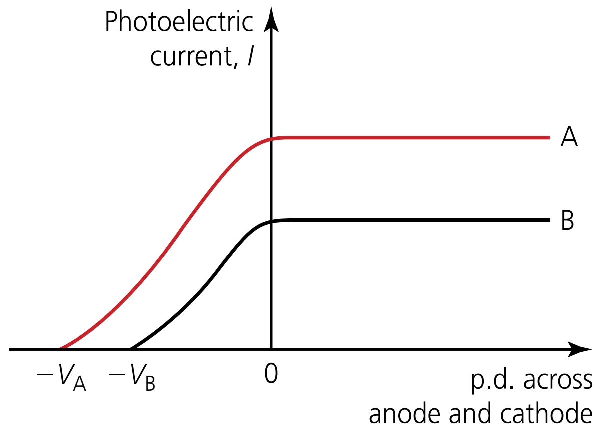


Figure 16.8 Variation of photoelectric current with p.d. for radiation of two different frequencies

- **Frequency** – Figure 16.8 shows the photoelectric currents produced by radiation from two monochromatic sources, of different frequencies, A and B, incident on the same metal.
  The individual photons in radiation A must have more energy (than B) and produce photoelectrons with a greater maximum kinetic energy. We know this because a greater reverse potential is needed to stop the more energetic photoelectrons produced by A.
  No conclusion can be drawn from the fact that the current for A has been drawn higher than for B, because the intensities of the two radiations are not known. In the unlikely circumstances that the two intensities were equal, the maximum current for B would have to be higher than for A because the radiation from B must have more photons, since each photon has less energy than in A.
- **Metal used in the cathode** – Experiments confirm that when different metals are tested using the same frequency, the photoelectric effect is observed with some metals, but not with others (for which the metal's work function is greater than the energy of the photons).

B.1.4 Solve problems involving the photoelectric effect.

13 Make a copy of Figure 16.5 and add the results that would be obtained using radiation of greater intensity (of the same frequency) incident on a metal which has a smaller work function.

14 Make a copy of Figure 16.8, line A only. Add to it a line showing the results that would be obtained with radiation of a higher frequency, but with same number of photons every second incident upon the metal.

The wave nature of matter

The de Broglie hypothesis and matter waves

B.1.5 Describe the de Broglie hypothesis and the concept of matter waves.

In 1924 the French physicist Louis de Broglie proposed that electrons, which were thought of as particles, might also have a wave-like character. He later generalized his hypothesis to suggest that *all* moving particles have a wave-like nature.

According to de Broglie, the wavelength, λ, of a moving particle is related to its momentum, p, by the equation:

$$\lambda = \frac{h}{p}$$

where h is Planck's constant. This is known as **de Broglie's equation**. Rearranging, this becomes:

$$p = \frac{h}{\lambda}$$

This equation is in the IB *Physics data booklet.*

Worked examples

4 Calculate the momentum of a moving particle which has a de Broglie wavelength of 200 pm (1 pm = 1×10^{-12} m).

$$p = \frac{h}{\lambda}$$

$$p = \frac{6.63 \times 10^{-34}}{2.00 \times 10^{-10}} = 3.32 \times 10^{-24}\,\text{kg m s}^{-1}$$

5 Calculate the de Broglie wavelength of a 1.00 g mass moving at a speed of 1.00 m per year.

$$v = \frac{1.00}{365 \times 24 \times 60 \times 60} = 3.17 \times 10^{-8}\,\text{m s}^{-1}$$

$$p = \frac{h}{\lambda}; \quad \lambda = \frac{6.63 \times 10^{-34}}{1.00 \times 10^{-3} \times 3.17 \times 10^{-8}} = 2.09 \times 10^{-23}\,\text{m}$$

The de Broglie wavelength for moving objects of ordinary size is very small because of their large mass. For this reason, the so-called **matter waves** of objects that are large enough to be seen cannot be detected. However, sub-atomic particles, like electrons, are so small that wave-like properties can be detected experimentally.

Matter waves, like electromagnetic waves, can travel in a vacuum. Unlike electromagnetic waves, matter waves are not produced by accelerated charges. Matter waves are probability waves (see page 611). The greater the amplitude of the wave, the greater the probability that the particle exists at a particular point.

The wavelength of de Broglie waves can also be related to the kinetic energy of the particle:

$$E_K = \frac{1}{2}mv^2$$

$$mE_K = \frac{1}{2}m^2v^2$$

$$m^2v^2 = 2mE_K$$

$$mv = \sqrt{2mE_K}$$

But mv represents momentum, p, so:

$$p = \sqrt{2mE_K}$$

Substituting this value of p into the de Broglie relationship, we obtain the following expression:

$$\lambda = \frac{h}{\sqrt{2mE_K}}$$

If a charged particle carrying a charge of q coulombs is accelerated from rest by applying a potential difference of V volts, then the kinetic energy of the particle is given by:

$$E_K = qV$$

Substituting qV in the equation for wavelength gives:

$$\lambda = \frac{h}{\sqrt{2mqV}}$$

The de Broglie hypothesis is the basis for the Schrödinger model of the hydrogen atom (see page 609). The electron of the hydrogen atom forms a standing wave around the central proton (nucleus).

15 Explain the statement, 'matter and radiation have dual character'.

16 **a** Derive the relationship between the wavelength of the de Broglie wave and the kinetic energy of the particle.
b If you double the kinetic energy of an electron (at speeds well below the speed of light), how does its de Broglie wavelength change?
c What happens to the wavelength if you double the speed of the electron?

17 Construct a spreadsheet that converts values of mass and velocity into momentum and wavelength (via the use of the de Broglie equation).

The Davisson–Germer experiment

B.1.6 Outline an experiment to verify the de Broglie hypothesis.

We have seen that de Broglie's hypothesis suggests that electrons have wave-like properties. They should therefore be able to undergo interference and diffraction.

Experimental confirmation of de Broglie's hypothesis and the wave nature of electrons was first obtained in 1927 when Clinton Davisson and Lester Germer showed that an electron beam

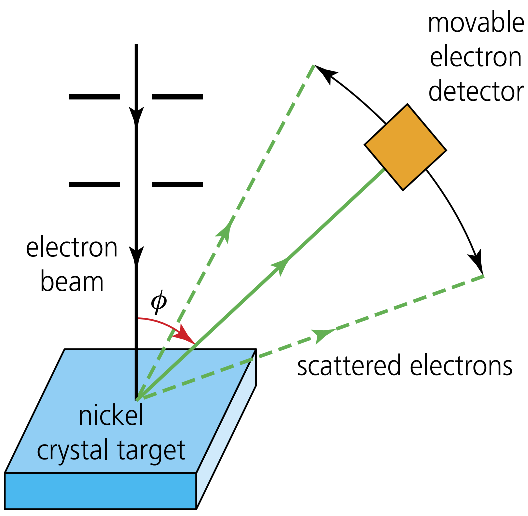


Figure 16.9 The principle behind the Davisson and Germer experiment

could be diffracted by a metal crystal (Figure 16.9). They used a beam of electrons fired at a nickel crystal target and recorded the electrons scattered at different angles. They found that the intensity of the scattered electrons varied with the angle and also depended on their speed (which could be altered by changing the accelerating potential difference).

Work had been done on metal crystals using X-rays, so the spacing between layers of atoms in the nickel crystal was known. The angle at which the maximum scattered intensity was recorded agreed with the angle predicted by constructive interference of waves by layers of atoms in the surface of the metal crystal. Accurate measurements of this diffraction pattern demonstrated that de Broglie's equation was correct.

The diffraction of electrons can be demonstrated in the school laboratory using a high voltage evacuated tube. A beam of electrons is accelerated in an electron gun to a high potential and then directed onto a very thin sheet of graphite (carbon atoms) (Figure 16.10). The atomic spacing of carbon atoms in the graphite is in the same order as the predicted wavelength of the electron (when travelling at very high velocity).

The electrons diffract (Chapter 11) from the carbon atoms and form a pattern on the screen of a series of concentric rings (Figure 16.10). This observed diffraction of electrons is strong experimental evidence for the wave nature of electrons. The characteristic pattern is due to the regular spacing of the carbon atoms in the different layers of graphite. A very thin sample of graphite has to be used to obtain sharp diffraction rings. If the carbon sample is too thick, the electrons lose kinetic energy and their wavelength becomes longer, smearing out the diffraction pattern.

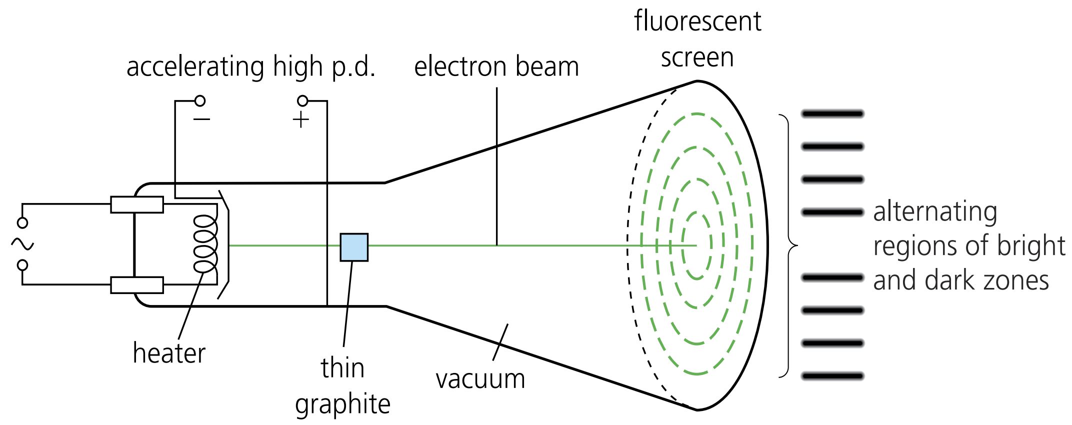


Figure 16.10 An electron diffraction apparatus

18 Describe the Davisson–Germer experiment and explain its importance.

19 Find out about neutron diffraction and its use in establishing the structure of substances, including proteins.

TOK Link: Nature of light

The photoelectric effect clearly demonstrates that light can behave as a stream of particles. However, its ability to undergo diffraction clearly demonstrates its wave-like nature. Is light a wave or particle? Whether a particle or wave model is useful depends on the nature of the experiment, but neither model is a complete description of light behaviour. Physicists often use the wave model when light is travelling through a medium, but use a particle model when light interacts with matter. Neither model is consistent with all observations. This dual nature of light is termed **wave–particle duality**.

Question

1 Try and explain the concept of wave–particle duality to a student who does not study physics. Try and think of some analogies to help convey the concept.

Calculations involving matter waves

B.1.7 **Solve** problems involving matter waves.

Worked examples

6 Calculate the de Broglie wavelength of an electron travelling with a speed of $1.0 \times 10^7\,\text{m s}^{-1}$. (Planck's constant = 6.63×10^{-34} J s; electron mass, $m_e = 9.110 \times 10^{-31}$ kg.)

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34}}{(9.110 \times 10^{-31}) \times (1.0 \times 10^7)} = 7.3 \times 10^{-11}\,\text{m}$$

7 Estimate the de Broglie wavelength of a ball of mass 0.058 kg moving with a velocity of $10^2\,\text{m s}^{-1}$.

$p = mv$; $p = 0.058 \times 10^2 = 5.8\,\text{kg m s}^{-1}$

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34}}{5.8} \approx 10^{-34}\,\text{m}$$

To exhibit wave-like properties the ball would need to interact with an object with dimensions of the order of 10^{-34} m (over one million million million times smaller than the nucleus of an atom). Hence, the ball does not exhibit any measurable or detectable wave-like properties.

8 Calculate the de Broglie wavelength of an electron which has been accelerated through a potential difference of 500 V.

The speed of the electrons can be calculated, using the law of conservation of energy:
loss of electrical potential energy = gain in kinetic energy

$$Ve = \frac{1}{2}mv^2$$

$$v = \sqrt{\frac{2Ve}{m}} = \sqrt{\frac{2 \times 500 \times 1.60 \times 10^{-19}}{9.110 \times 10^{-31}}} = 1.32 \times 10^7\,\text{m s}^{-1}$$

$$p = mv = 1.2 \times 10^{-23}\,\text{kg m s}^{-1}$$

$$\lambda = \frac{h}{p} = 5.5 \times 10^{-11}\,\text{m}$$

9 Calculate the de Broglie wavelength of a lithium nucleus $\left({}^7_3\text{Li}\right)$ which has been accelerated through a potential difference of 5.00 MV. (The mass of a lithium-7 nucleus is 1.165×10^{-26} kg.)

$$\frac{1}{2}mv^2 = 3e \times V$$

$$v^2 = \frac{6 \times (1.60 \times 10^{-19}) \times (5.00 \times 10^6)}{(1.165 \times 10^{-26})} = 4.12 \times 10^{14}$$

$$v = 2.03 \times 10^7\,\text{m s}^{-1}$$

$$\lambda = \frac{6.63 \times 10^{-34}}{(1.165 \times 10^{-26}) \times (2.03 \times 10^7)} = 2.81 \times 10^{-15}\,\text{m}$$

20 Calculate the wavelength (in metres) of an electron moving with a velocity of $2.05 \times 10^7\,\text{m s}^{-1}$.

21 A neutron has a de Broglie wavelength of 800 pm. Calculate the velocity of the neutron. The mass of the neutron is 1.675×10^{-27} kg.

22 The mass of an electron is 9.110×10^{-31} kg. If its kinetic energy is 3.0×10^{-25} J, calculate its de Broglie wavelength in metres.

23 Over what potential difference do you have to accelerate electrons for them to have a wavelength of 1.2×10^{-10} m?

24 **a** Which is associated with a de Broglie wavelength of longer wavelength – a proton or an electron travelling at the same velocity? Explain your answer.

b The following (non-relativistic particles) all have the same kinetic energy. Rank them in order of their de Broglie wavelengths, greatest first: electron, alpha particle, neutron and gold nucleus.

25 Explain why a moving airplane has no detectable wave properties.

Atomic spectra and atomic energy states

B.1.8 **Outline** a laboratory procedure for producing and observing atomic spectra.

Emission spectra

Figure 16.11 The flame test for potassium ions from potassium chloride

When a sample of an element is excited by heating (the flame test – Figure 16.11) or by passing an electric current through its gas (at low pressure), the atoms of the element emit electromagnetic radiations of definite frequencies. These characteristic frequencies form the **emission spectrum** of the element. Since the radiations in the spectrum are emitted due to energy changes taking place in the atoms, this spectrum is known as an atomic emission spectrum. An atomic emission spectrum for an element consists of a unique series of coloured lines on a black background.

The instrument used for obtaining and analysing an emission spectrum is known as a **spectrometer**. A simple spectrometer (Figure 16.12) works by passing a beam of light from excited gaseous atoms through a lens via a slit to a second lens and then to a prism (or diffraction grating). This separates the **spectral lines**, which can be observed and measured using an eyepiece (or travelling microscope).

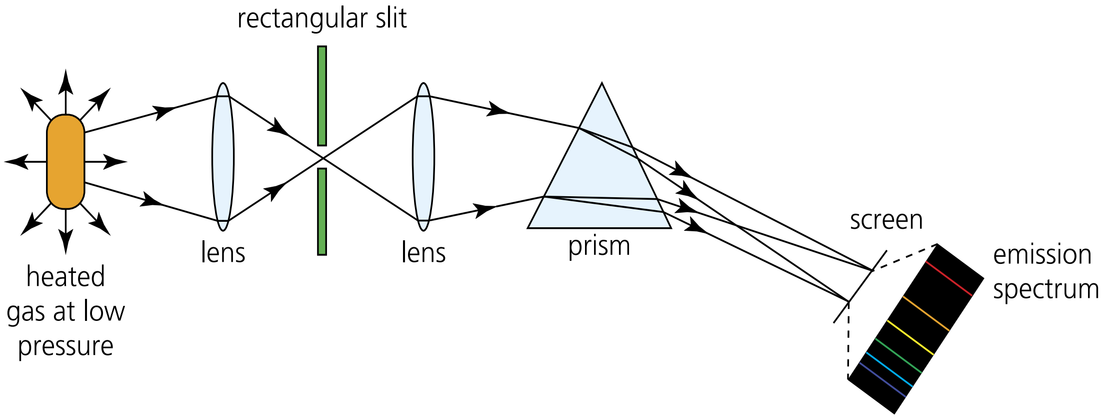


Figure 16.12 Schematic diagram showing the production of an atomic emission spectrum

Different elements have different **energy levels** for their electrons. In a gas excited by heating or by being subjected to electric discharge, electrons in the excited atoms 'fall back' to lower energy levels. They do this by emitting photons of specific frequencies seen in the **line spectrum**. This provides useful information for physicists to identify elements and study their atomic structures.

Absorption spectra

Cool gaseous atoms (and molecules) absorb electromagnetic radiation as well as emitting it. When white light is shone through a sample of a gaseous element (at low pressure and low temperature), the light that emerges has the element's characteristic wavelengths (or frequencies) missing (Figure 16.13). There are dark lines or bands in the discontinuous spectrum.

For each element, the position of these dark bands is the same as the position of the lines in the emission spectrum. An element's characteristic lines or frequencies is called its **absorption spectrum.**

In the formation of an absorption spectrum, the wavelengths that are absorbed are re-emitted, but in all directions, so the original beam of light has a very much reduced (but not zero) intensity of light at these wavelengths.

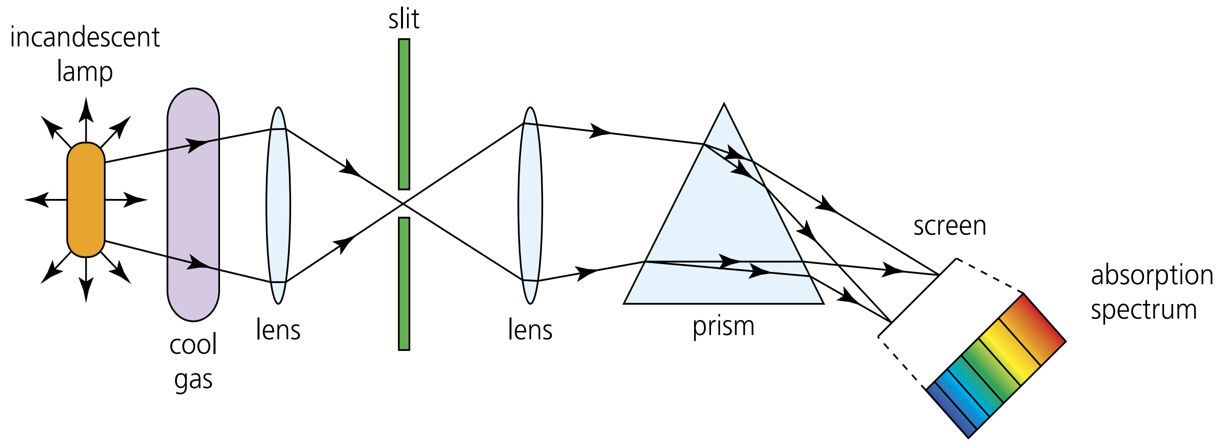


Figure 16.13 Schematic diagram showing the production of an atomic absorption spectrum

26 a You are given samples of argon, neon and helium gases. All these elements are colourless gases at room temperature. Outline laboratory procedures for producing and observing emission and absorption atomic spectra for these gases.
b Use the Internet to find the emission and absorption spectra of the three gases. Sketch the visible regions of their emission and absorption spectra.

Quantized energy in atoms

B.1.9 **Explain** how atomic spectra provide evidence for the quantization of energy in atoms.

We have seen that atomic emission and absorption spectra have discrete frequencies corresponding to specific energies. The Bohr model of the hydrogen atom (Chapter 7) can be used to explain how these spectral lines arise as a result of quantized electron energies.

In Bohr's model, electrons can only exist in certain energy levels. Electrons in these energy levels behave like stationary waves fitted into the circumference of the orbit. Energy levels are labelled $n = 1, 2, 3 \ldots$, etc. This is shown in Figure 16.14 for $n = 4$ and $n = 8$. Electrons can only have these energies, but not energies in between.

A 'free' electron at $n = \infty$ has zero energy (the hydrogen atom has become an ion, and is said to be ionized). The energy level at $n = 1$ has the lowest energy: −13.6 eV for a hydrogen atom. A hydrogen atom with its electron in the $n = 1$ energy level is in the **ground state**. All energy values on this scale are negative, indicating a loss of energy as the electron moves closer to the nucleus (with decreasing values of n) (Figure 16.15).

For an electron to move from one energy level to another, energy must be absorbed or given out. Transitions between energy levels can be represented by vertical arrows drawn between the energy levels. (Figures 16.16a and 16.16b illustrate this for the important example of hydrogen.)

When a gas is heated, thermal energy raises some electrons to a higher energy level, resulting in the production of an **excited state**. An excited electron in an upper energy level will fall back to a lower energy level, usually after a very short time interval. The downward electron transition corresponds to the emission of a photon whose energy is the same as the energy *difference* between the energy levels. This is given by the relationship:

$$hf = E_2 - E_1$$

where E_2 is the energy of the higher energy level and E_1 is the energy of the lower energy level.

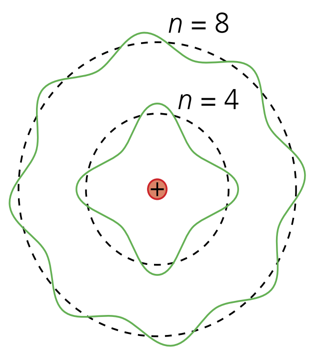


Figure 16.14 Electron waves in the fourth and eighth energy levels of the Bohr model of the hydrogen atom

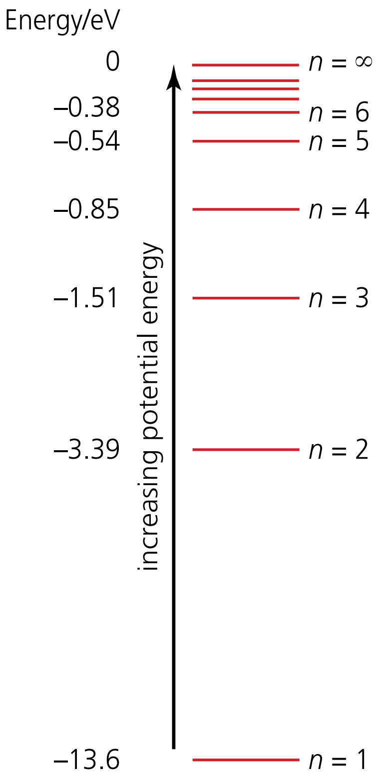


Figure 16.15 Energy levels for the Bohr model of the hydrogen atom

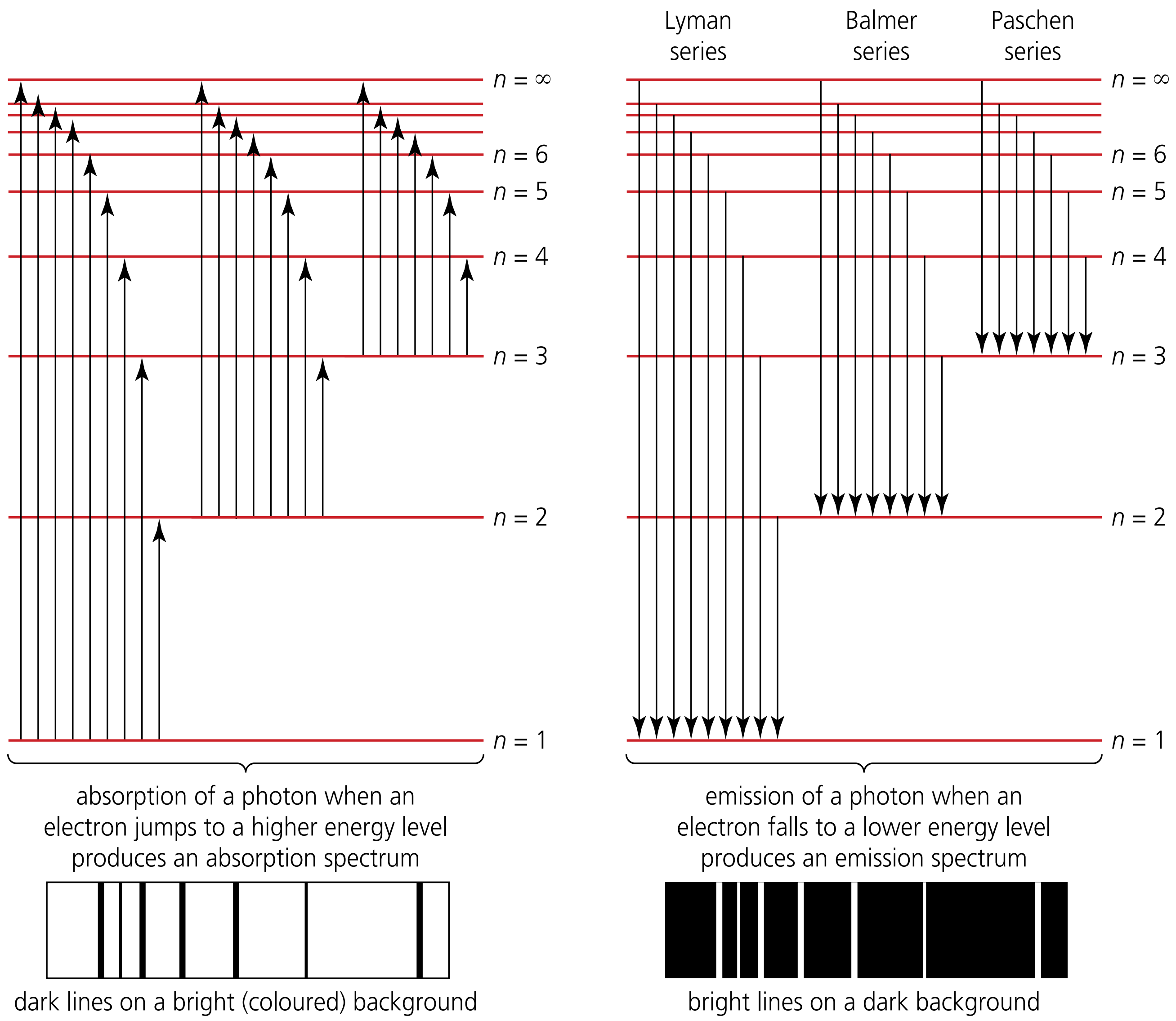


Figure 16.16 a Electron transitions from a lower to a higher energy state in a hydrogen atom **b** Electron transitions from a higher to a lower energy state in a hydrogen atom

In a similar way, photons can only be absorbed if their energy corresponds to the energy difference between two energy levels. So the lines in absorption and emission spectra have the same characteristic frequencies, as they correspond to the same energy differences.

Further away from the nucleus, the energy levels become closer and closer in energy. This arrangement of energy levels is shown in emission and absorption spectra, which show **convergence** of lines at high frequency.

Transitions between the higher energy levels, where the energy difference between levels is smaller, correspond to photons with less energy in the visible and infrared regions of the electromagnetic spectrum. A transition with a large energy difference corresponds to an X-ray photon.

27 Electron energies within atoms and molecules are quantized. Explain what this statement means.

28 **a** State which feature of emission and absorption spectra (of isolated gaseous atoms) provides strong experimental evidence for electron quantization.
b Account (in general terms) for the production of these features in absorption and emission spectra.

B.1.10 Calculate wavelengths of spectral lines from energy level differences and vice versa.

Calculations involving the interconversion between energy levels and wavelengths

The **Planck relationship**, $hf = E_2 - E_1$, and the wave equation, $v = f\lambda$, can be used to calculate the wavelengths (or frequencies) of spectral lines from energy level differences in atoms, or calculate energy level differences from wavelengths (or frequencies) in a spectrum.

Worked examples

10 Calculate the wavelength of the electromagnetic radiation emitted when the electron in a hydrogen atom makes a transition from the energy level at -0.54×10^{-18} J to the energy level at -2.18×10^{-18} J.
(Planck's constant $h = 6.63 \times 10^{-34}$ J s; speed of light in a vacuum, $c = 3.00 \times 10^8$ m s^{-1}.)

$$\Delta E = E_2 - E_1 = -0.54 \times 10^{-18} - (-2.18 \times 10^{-18}) = 1.64 \times 10^{-18}\,\text{J}$$

$$\lambda = \frac{hc}{\Delta E} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^8}{1.64 \times 10^{-18}} = 1.21 \times 10^{-7}\,\text{m} = 121\,\text{nm}$$

11 The shortest wavelength line in the atomic spectrum of hydrogen (in the Lyman series) has a wavelength of 9.17×10^{-8} m. Calculate the energy of the transition in joules.

$$\Delta E = \frac{hc}{\lambda} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^8}{9.17 \times 10^{-8}} = 2.17 \times 10^{-18}\,\text{J}$$

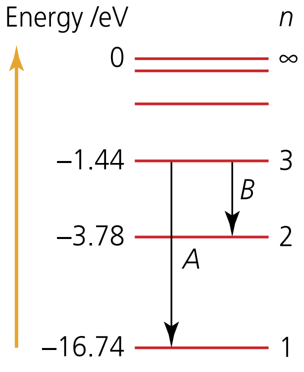


Figure 16.17

29 Figure 16.17 shows two transitions between energy levels of an atom.
a Calculate the energy change (in eV) in the electron transition labelled A.
b Calculate the wavelength of the photons (light) emitted by this transition.
c How would the wavelength for transition A compare with that for transition B? (No calculation is required.) Explain your answer.

30 Explain why the energy levels in the atom are given negative values.

The 'electron in a box' model

B.1.11 **Explain** the origin of atomic energy levels in terms of the 'electron in a box' model.

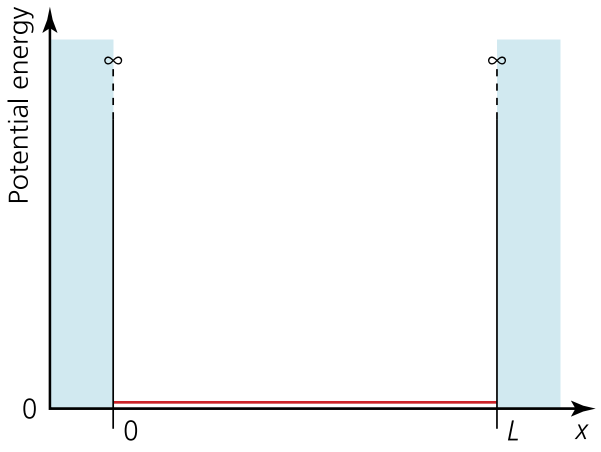


Figure 16.18 The 'electron in a box' model

The 'electron in a box' model uses the idea of the electron behaving as a wave to explain why the energy levels in atoms are quantized. In this model the electron is confined in a small region between two walls and can travel in a straight line in one dimension (along the x-axis) from $x = 0$ to $x = L$ (Figure 16.18). The potential energy of the electron inside the box is zero, but increases to infinity at the walls. The electron is trapped in a *potential well*.

If we think of the electron as behaving as a wave inside the 'box', then it must be subject to *boundary conditions*, like those applied to the waves of a violin string fixed at both ends. The waves are *standing waves* (Chapter 11) generated by the electron reflecting backwards and forwards from the sides of the 'box'. The standing waves have nodes (regions of no vibration) and antinodes (regions of maximum vibration).

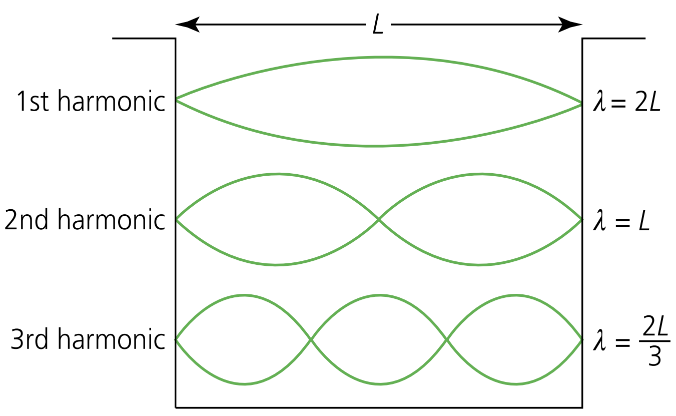


Figure 16.19 Standing electron waves within a box of length L

The walls of the box must always be nodes, so these standing waves must have an *integer* (whole) number of half wavelengths within the 'box'. This results in quantization of energy levels, just as the boundary conditions for a violin string produce particular harmonics. The wavelength of the electron wave is determined by the size of the 'box', the longest possible wavelength being $2L$ as shown in Figure 16.19 (a similar diagram is shown in Chapter 11).

The de Broglie wavelength of the electron of mass m_e is given by:

$$\lambda = \frac{h}{p} = \frac{h}{m_e v}$$

$$\text{so } v = \frac{h}{m_e \lambda}$$

The kinetic energy of the electron is $\frac{1}{2}m_e v^2$. Substituting for v in this expression gives the kinetic energy, E_K, as:

$$E_K = \frac{1}{2}m_e \times \frac{h^2}{m_e^2\lambda^2} = \frac{h^2}{2\lambda^2 m_e}$$

Figure 16.19 shows that the electron in the 'box' can have the following values of wavelength: $2L$, L, $\frac{2L}{3}$, $\frac{L}{2}$, etc. All possible standing electron waves are defined by the following equation:

$$\lambda = \frac{2L}{n} \quad \text{where } n \text{ represents an integer (1, 2, 3, 4, etc.)}$$

Substituting this expression for the wavelength into the equation for kinetic energy gives:

$$E_K = \frac{n^2h^2}{8m_eL^2}$$

This equation is in the IB *Physics data booklet.*

Because the potential energy is zero inside the box, the total energy is the kinetic energy. As the 'box' gets smaller, the wavelengths decrease and the energy increases.

The electron in a box model is not an accurate model of electrons within an atom, but as we have seen, it clearly explains the quantized (discrete) nature of the electron's energies when the electron is treated as a wave.

Worked examples

12 Calculate the ground state energy ($n = 1$) in eV (to the nearest integer) for an electron confined to a box of length 0.50×10^{-10} m.

$$E_K = \frac{n^2h^2}{8m_eL^2} \text{ where } n = 1$$

$$E_1 = \frac{(1)^2 \times (6.63 \times 10^{-34})^2}{8 \times (9.110 \times 10^{-31}) \times (0.50 \times 10^{-10})^2}$$

$$= 2.4 \times 10^{-17}\,\text{J} = 151\,\text{eV}$$

13 An electron confined to a box has a ground state energy of 20 eV. Calculate the width (in nm) of the box.

$$E_K = \frac{n^2h^2}{8m_eL^2} \text{ where } n = 1$$

$$20 \times 1.60 \times 10^{-19} = \frac{(1)^2 \times (6.63 \times 10^{-34})^2}{8 \times (9.110 \times 10^{-31})L^2}$$

$$L = 1.4 \times 10^{-10}\,\text{m} = 0.14\,\text{nm}$$

31 **a** Describe the 'electron in a box' model. Explain the origin of the atomic energy levels.
b What are the allowed wavelengths for an electron in this model?

32 Calculate the kinetic energies for an electron ($m_e = 9.110 \times 10^{-31}$ kg) in the first two energy levels ($n = 1$ and $n = 2$) in a one-dimensional box of length 1.0×10^{-10} m.

The Schrödinger model of the hydrogen atom

B.1.12 **Outline** the Schrödinger model of the hydrogen atom.

In 1926 the Austrian physicist Erwin Schrödinger proposed a quantum mechanical model describing the behaviour of the electron within a hydrogen atom. In Schrödinger's theory (wave equation) the electron in the atom is described by a **wavefunction**, $\Psi(x, t)$, in which the amplitude psi (Ψ) of the wave is a function of position x and time t.

According to quantum mechanics, a particle cannot have a precise path; instead, there is only a probability that it may be found at a particular place at any particular time (this is discussed in more detail in the next section on the Heisenberg uncertainty principle). A wavefunction may be viewed as a blurred version of the path of an electron (Figure 16.20). The darker the area of shading, the greater the probability of finding an electron there.

If the electrostatic forces that act on the electron are known then the **Schrödinger wave equation** can be solved to obtain $\Psi(x, t)$ (Figure 16.21), which allows the energy levels to be calculated.

Figure 16.20 Classical and quantum mechanical descriptions of a moving electron

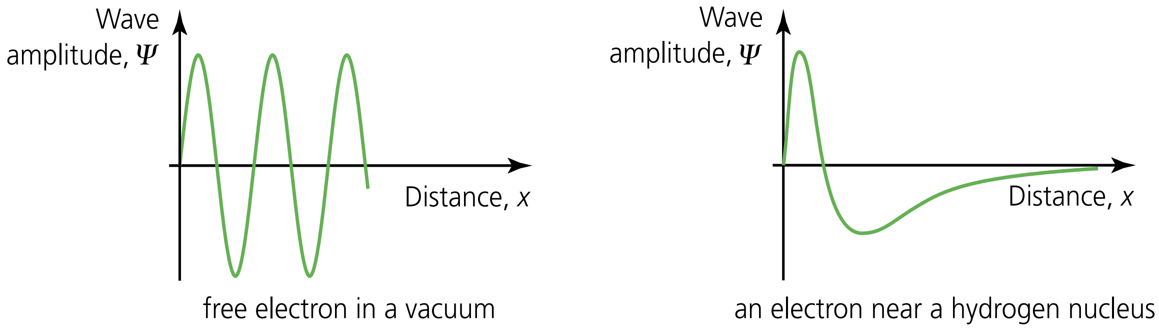


Figure 16.21 The wavefunctions, Ψ, of a free electron in a vacuum and an electron near a hydrogen nucleus (a proton)

A more useful quantity is $|\Psi(x, t)|^2$, the square of the absolute value of the amplitude. This is a measure of the electron density (**probability distribution**) in a region of space, and so gives us the probability that an electron will be found near a point x at time t. The wavefunctions of electrons in atoms are known as **orbitals** and often described as 'electron clouds'.

Solving the Schrödinger wave equation for the single electron in a hydrogen atom gives results similar to the simple 'electron in a box' model described in the previous section. It predicts that the total energy, E, of the electron in the hydrogen atom (the sum of the kinetic energy and potential energy) is given by the expression:

$$E \text{ (in eV)} = -\frac{13.6}{n^2} \quad \text{where } n \text{ is the energy level}$$

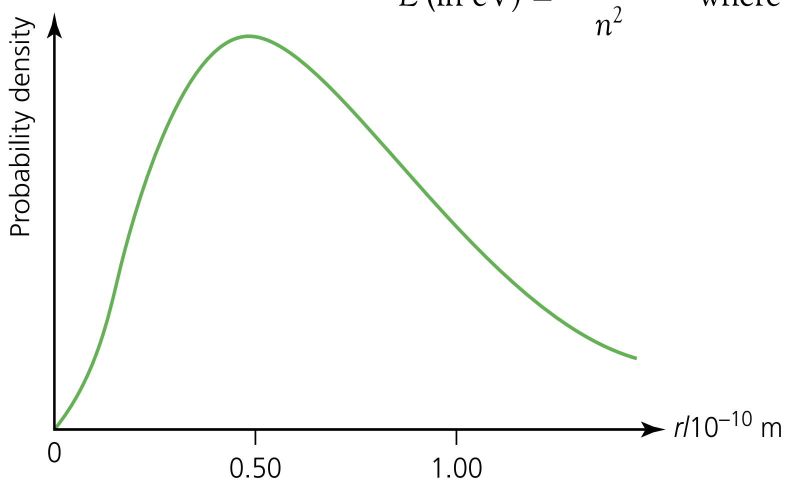


Figure 16.22 Probability distribution function for the variation with distance r from the nucleus for the energy level $n = 1$ (lowest energy, or ground state) of the hydrogen atom

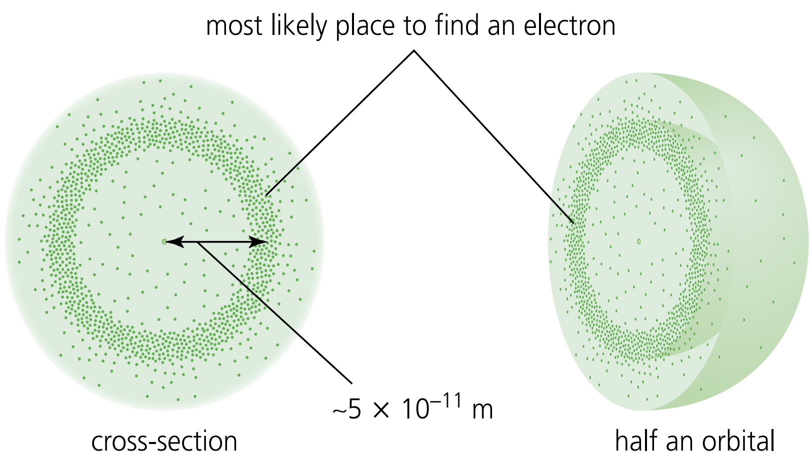


Figure 16.23 Electron cloud for the 1s orbital in the hydrogen atom

Schrödinger's theory therefore predicts that the electron in the hydrogen atom has quantized values of total energy. The electron will be found in one of the energy levels of the hydrogen atom depending on the value of the integer n. These energy levels are similar to those predicted by Bohr theory and electron transitions between energy levels are also predicted. The Schrödinger wave equation can also be applied to atoms other than hydrogen and solved approximately.

The graph in Figure 16.22 shows the *calculated* variation in *electron density* (proportional to $|\Psi|^2$) with radial distance r from the hydrogen nucleus for the lowest energy level $n = 1$ (ground state). The peak indicates the most probable distance of the electron from the nucleus, which is about 0.50×10^{-10} m. This is where the electron density is highest. The ground state orbital of a hydrogen atom is known as the 1s orbital, where 1 refers to the first energy level and s its shape.

The wavefunction exists in three dimensions and often the electron orbital is pictured as an 'electron cloud' (Figure 16.23) of varying electron density. The electron density distribution is like a photograph of the atom with a long exposure time. The electron is more likely to be in the positions where the electron density (represented by dots) is highest. For the first main energy level ($n = 1$) the orbital takes the form of a fuzzy sphere.

33 Summarize the main features of Schrödinger's quantum mechanical model of the hydrogen atom.

34 Calculate the wavelength of the photon emitted in the transition from $n = 2$ to $n = 1$.

35 Find out about the energies and shapes of p, d and f orbitals in atoms.

36 Describe how a Bohr orbit is different from a Schrödinger orbital.

The Heisenberg uncertainty principle

B.1.13 Outline the Heisenberg uncertainty principle with regard to position–momentum and time–energy.

Imagine in a 'thought experiment' recording the measurement of the speed and position of a single electron in a beam of electrons, as shown in Figure 16.24. The position of an electron in the beam can be found by aiming a photon from a light source at it. After interacting with the electron the photon may rebound into a photomultiplier tube, a device for detecting photons. However, when the photon interacts with the electron there will be a transfer of energy, making the precise measurement of the electron's speed impossible.

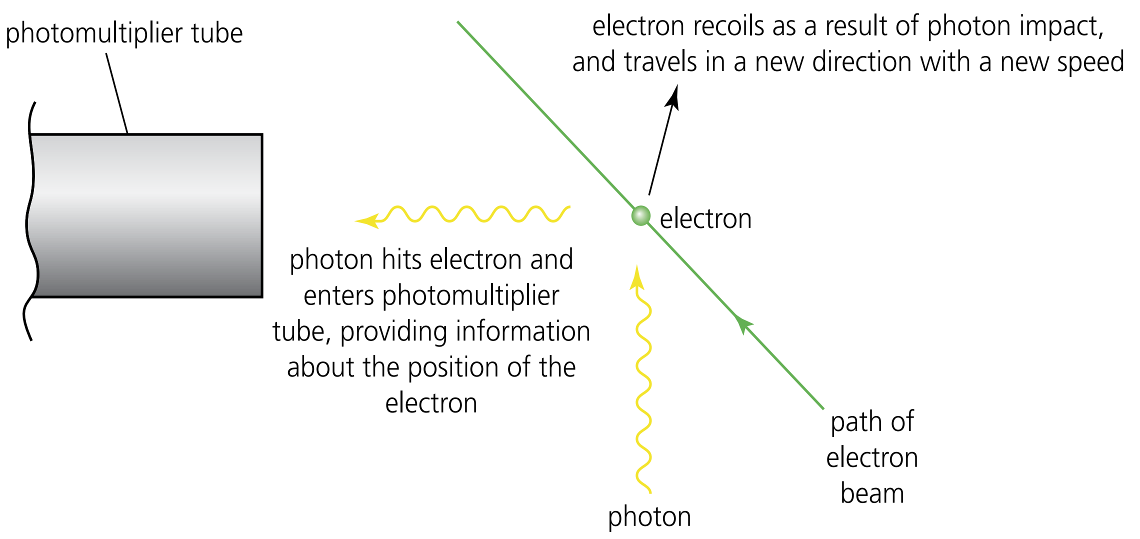


Figure 16.24 When the speed and position of an electron are measured (using photons) these quantities are altered, introducing uncertainties into the measurements

The German physicist Werner Heisenberg put these ideas into a mathematical relationship known as the Heisenberg uncertainty principle.

For the measurement of position the relationship can be written as:

$$\Delta x \Delta p \geq \frac{h}{4\pi}$$

This equation is in the IB *Physics data booklet.*

where Δx represents the uncertainty in the measurement of position, Δp represents the uncertainty in the measurement of momentum (the product of mass and velocity) and h represents Planck's constant.

The Heisenberg uncertainty principle implies that the more accurately we know the speed of the electron (the smaller Δp), the less we know about where it is (the larger Δx), and this uncertainty is significant. Because of this, momentum and position are linked variables and termed **conjugate quantities**. In particular, if one is made zero, the other has to be infinite.

This implies the following link between the de Broglie hypothesis (page 601) and the Heisenberg uncertainty principle: if a particle has a uniquely defined de Broglie wavelength, then its momentum will be known precisely, but there is no knowledge of its position. This can also be applied to the Schrödinger model of the hydrogen atom: if the wavelength of the electron's matter wave is well defined, then the position of the electron is unknown. The uncertainty principle is a general result that follows from wave–particle duality.

Measurements of time and energy are also linked variables and described by the energy–time uncertainty principle:

$$\Delta E \Delta t \geq \frac{h}{4\pi}$$

This equation is in the IB *Physics data booklet.*

where ΔE represents the uncertainty in the measurement of energy and Δt represents the uncertainty in the measurement of time.

Worked examples

14 An electron moves in a straight line with a constant speed ($1.30 \times 10^6\,\text{m s}^{-1}$). The speed can be measured to a precision of 0.10%. What is the maximum possible precision (minimum uncertainty) in its position if it is measured simultaneously?

$p = mv$

$p = (9.110 \times 10^{-31}) \times (1.30 \times 10^6) = 1.18 \times 10^{-24}\,\text{kg m s}^{-1}$

$\Delta p = 0.0010p = 1.18 \times 10^{-27}\,\text{kg m s}^{-1}$

$$\Delta x \geq \frac{h}{4\pi \times \Delta p} = \frac{6.63 \times 10^{-34}}{4 \times \pi \times 1.18 \times 10^{-27}}$$

$\Delta x_{min} = 4.45 \times 10^{-8}\,\text{m}$

Maximum possible precision of position (minimum uncertainty), $\Delta x = 44.5\,\text{nm}$.

15 The position of an electron is measured to the nearest $0.50 \times 10^{-10}\,\text{m}$. Calculate its minimum uncertainty in momentum.

$$\Delta p \geq \frac{h}{4\pi \times \Delta x} = \frac{6.63 \times 10^{-34}}{4 \times \pi \times 0.50 \times 10^{-10}}$$

$\Delta p_{min} = 1.1 \times 10^{-24}\,\text{kg m s}^{-1}$

16 Calculate the uncertainty in velocity for an electron in a $1.0 \times 10^{-10}\,\text{m}$ radius orbit in which the positional uncertainty is 1% of the radius.

$$\Delta p \geq \frac{h}{4\pi \times \Delta x} = \frac{6.63 \times 10^{-34}}{4 \times \pi \times 1.0 \times 10^{-12}} = 5.27 \times 10^{-23}\,\text{kg m s}^{-1}$$

$$\Delta v = \frac{\Delta p}{m} = \frac{5.27 \times 10^{-23}}{9.11 \times 10^{-31}} = 5.8 \times 10^{7}\,\text{m s}^{-1}$$

37 Find about Schrödinger's hypothetical 'cat in the box' experiment and explain its connection with quantum theory.

38 An electron is located within an atom to a distance of $0.1 \times 10^{-10}\,\text{m}$. What is the uncertainty involved in the measurement of its velocity?

39 A mass of 40.00 g is moving at a velocity of $45.00\,\text{m s}^{-1}$. If the velocity can be calculated with an accuracy of 2%, calculate the uncertainty in the position.

40 The lifetime of a free neutron is 15 minutes. How uncertain is its energy (in eV)?

41 Find out about quantum tunnelling.

TOK Link: Uncertainty principle

The idea of the detached observer of classical or Newtonian mechanics is false, since a completely isolated universe cannot be observed. An observer must always form part of an experiment – otherwise there is no experiment. The uncertainty principle also places a definite limit to the precision of measurements and hence human knowledge. The uncertainty principle is not due to any limitation of the measuring device (Chapter 1) but is a direct consequence of the dual nature of matter (wave–particle duality).

Question

1 Why does the philosophical concept of free will depend on the uncertainty principle?

B2 Nuclear physics

Investigating the nucleus

This section looks at ways of estimating the size and mass of the nucleus, and evidence for nuclear energy levels.

Rutherford scattering

B.2.1 **Explain** how the radii of nuclei may be estimated from charged particle scattering experiments.

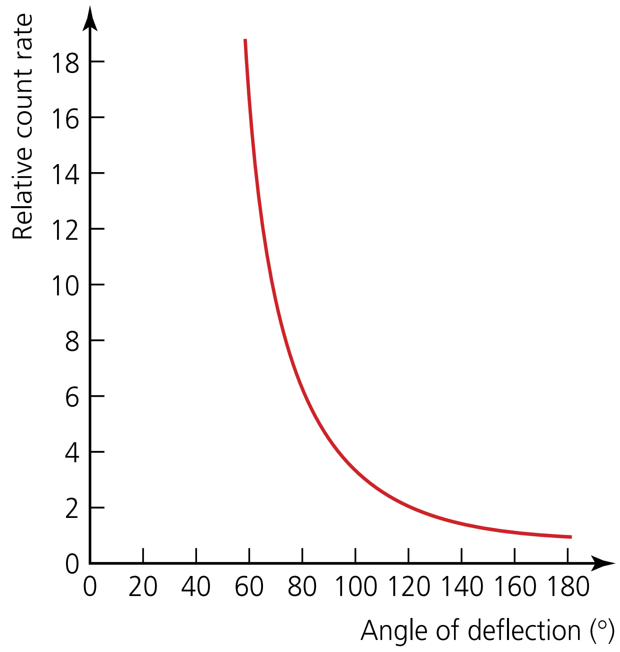


Figure 16.25 The rate of scattering of alpha particles at different angles

In Chapter 7 we looked at the experiment conducted by Geiger and Marsden, under the supervision of Ernest Rutherford, in which alpha particles were scattered by gold foil. The graph of Figure 16.25 shows the typical results obtained from this experiment. The most unexpected observation was that some alpha particles were scattered through large angles. In fact, about 1 in every 10 000 alpha particles incident on the gold foil was deflected by an angle greater than 90°.

In order for an alpha particle to be scattered straight back (with very little loss of kinetic energy), it must have 'collided' with a much larger mass (see elastic collisions in Chapter 2). Because most of the alpha particles are *not* deflected, the mass of each atom must be concentrated in a very small centre (the nucleus), such that most particles do not collide with it. Rutherford further realized that the *pattern* into which the alpha particles were scattered could only be explained by the action of large repulsions between the positive alpha particles and tiny positively charged nuclei. He proposed that each nucleus has a charge much larger than that of an alpha particle and that this is effectively concentrated at a *point* at the centre of the atom. The size of the forces between the charges is described by Coulomb's law ($F = \frac{kq_1q_2}{r^2}$, see Chapter 6), and another name for this effect is 'Coulomb scattering'.

Rutherford was able to estimate the radius of a gold nucleus by calculating how close an alpha particle got to the nucleus during its interaction. At its nearest distance the alpha particle moving directly towards the nucleus stops moving, so all its original kinetic energy is now stored as electrical potential energy. It is as if an 'invisible spring' is being squeezed between the alpha particle and gold nucleus as they come closer and closer. When the alpha particle stops moving all its initial kinetic energy (E_K) is stored in the 'spring' (Figure 16.26).

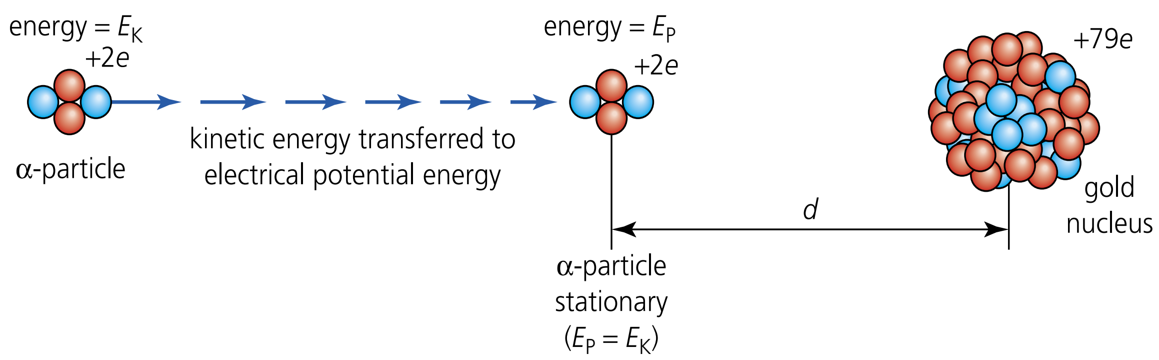


Figure 16.26 At closest approach in a head-on collision, the electrical potential energy stored in the electric field is equal to the initial kinetic energy of the alpha particle, $E_P = E_K$

The equation for electrical potential energy (E_P) depends on the separation r of the two charges:

$$E_P = \frac{q_1q_2}{4\pi\varepsilon_0 r}$$

This expression is one form of Coulomb's law (Chapter 6).

At the distance of closest approach the kinetic energy is transferred to electrical potential energy, so:

$$E_P = \frac{q_1 q_2}{4\pi\varepsilon_0 r} = E_K$$

If the kinetic energy of the alpha particle and the two charges (of the gold nucleus and the alpha particle) are known, then their separation, r, can be calculated:

$$r = \frac{q_1 q_2}{4\pi\varepsilon_0 E_K} = \frac{2e \times Ze}{4\pi\varepsilon_0 E_K}$$

where Z represents the atomic number of the gold nucleus.

If the target nucleus is gold ($Z = 79$) and the incident alpha particles have a kinetic energy of about 4 MeV (which is typical for naturally produced alpha particles), the distance of closest approach is:

$$r = \frac{2 \times 79 \times (1.6 \times 10^{-19})^2}{4\pi \times 8.9 \times 10^{-12} \times 4.0 \times 10^6 \times 1.6 \times 10^{-19}} \approx 10^{-14}\,\text{m}$$

This is the separation of the alpha particle and gold nuclei at closest approach and gives an *upper limit* for the sum of their radii. From this calculation it is assumed that the radius of a gold nucleus is of the order of 10^{-14} m. We cannot be sure that the alpha particle 'touches' the nucleus; a more energetic alpha particle might get closer still. More accurate measurements from the scattering of high energy electrons confirm that the gold nucleus has a radius of 6.9×10^{-15} m.

The behaviour of the alpha particle when it collides with the **potential hill** around the nucleus can be modelled with the apparatus shown in Figure 16.27. It is a gravitational analogue, with a ball bearing representing an alpha particle. The curved surface is constructed so that its height above the bench is proportional to $1/r$, where r represents the distance from its centre. The ball bearings are rolled down a ramp to give them a fixed amount of kinetic energy. As they move over the curved surface which represents the potential hill around the nucleus, they slow down and then accelerate again, coming out in a different direction.

Because there is a uniform gravitational field in the laboratory, the ball bearings gain gravitational potential energy in proportion to $1/r$. This means that the force on them varies as $1/r^2$. The Coulomb repulsion between a nucleus and an alpha particle varies in a similar way. The electrical potential energy is proportional to $1/r$ and the electrostatic repulsive force is proportional to $1/r^2$.

Figure 16.27 A model of the potential hill around a nucleus

42 Outline how the radii of metal nuclei may be estimated from Coulomb scattering experiments.

43 Calculate the velocity at which an alpha particle (mass of 6.64×10^{-27} kg) should travel towards the nucleus of a gold atom (charge $+79e$) so it gets within 2.7×10^{-14} m of it. Assume the gold nucleus remains stationary.

The mass spectrometer

B.2.2 Describe how the masses of nuclei may be determined using a Bainbridge mass spectrometer.

The first direct investigation of the mass of atoms (and molecules) was made possible by the development of the **mass spectrometer**. The first mass spectrometer was built in 1918 by William Aston, a student of J.J. Thomson. It provided direct evidence for the existence of isotopes (Chapter 7). A more accurate mass spectrometer was developed by William Bainbridge in 1932.

This device uses the interaction of charged ions with electric and magnetic fields to measure the relative masses of atoms and to find their relative abundances. Figure 16.28 shows how a simple Bainbridge mass spectrometer works. There is a vacuum inside the machine; no air is allowed to enter. This is to avoid collisions with particles in the air that would disrupt the flight of the ions produced inside the instrument.

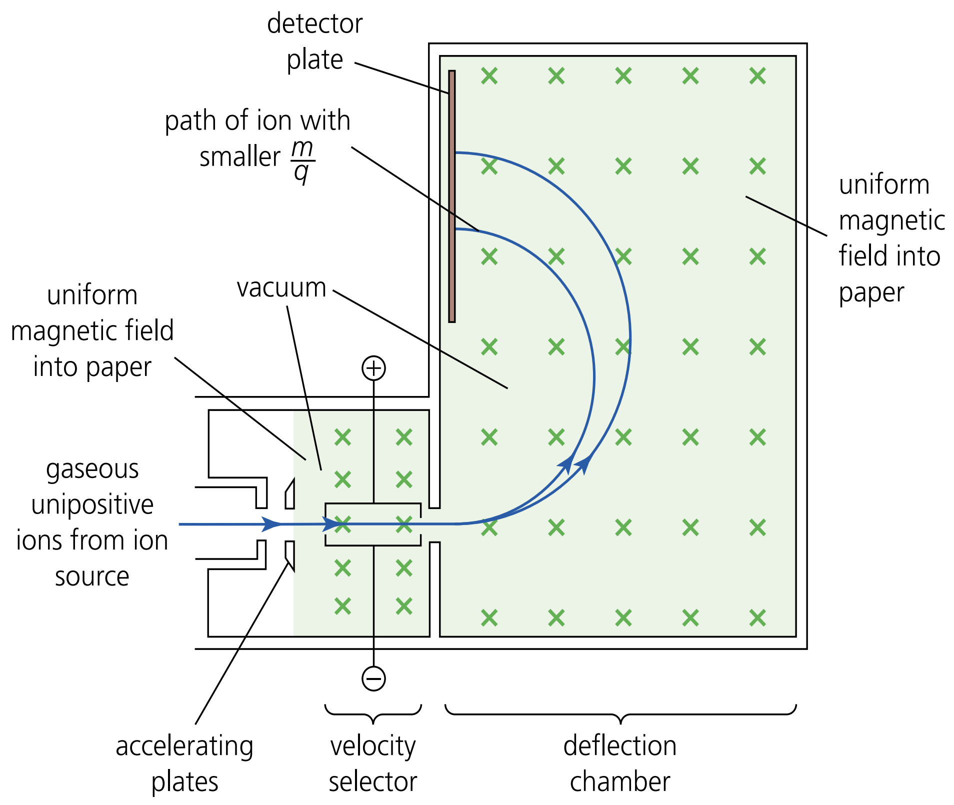


Figure 16.28 Essential features of a Bainbridge mass spectrometer

Positive ions are produced in the mass spectrometer by bombarding gaseous atoms or molecules (M in the following equation) with a stream of high speed electrons:

$$M(g) + e^- \rightarrow M^+(g) + e^- + e^-$$

The ions are accelerated (by a pair of positive plates) and then pass through a **velocity selector**. Here electric and magnetic fields are applied to the ions so that only a narrow beam of ions travelling at the same velocity continue in a straight line and can enter the next chamber. These ions then travel through a uniform magnetic field. This causes the ions to move in a circular path with radius, r, which depends on the mass to charge ratio. For ions with mass m and charge, q, travelling with velocity, v, through the magnetic field, B:

centripetal force = force due to magnetic field

$$\frac{mv^2}{r} = Bqv$$

The radius of the circle for the ion is therefore:

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

If the ions have the same charge, q (usually unipositive, i.e. with a single positive charge), and they are all selected to be travelling at the same velocity, v, then the radius of the circle of each ion's path will only depend on the mass of the ion. An ion with a larger relative mass will travel in a larger circle (for the same magnetic field strength).

A number of vertical lines will be obtained on the detector plate, each line corresponding to a different isotope of the same element. The position of a line on the plate will allow the radius, r, to be determined. As the magnetic field strength, B, the charge on the ion, e, and velocity of the ion, v, are all known, the mass of the ion, m, can be easily determined. The older type of Bainbridge mass spectrometer shown is actually a mass spectrograph, since the beam of ions is directed onto a photographic plate (Figure 16.29). The relative intensities of the lines allowed an estimate to be made of the relative amounts of isotopes.

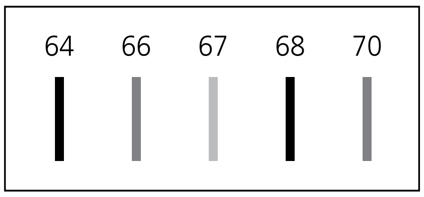


Figure 16.29 A mass spectrum obtained by a mass spectrograph

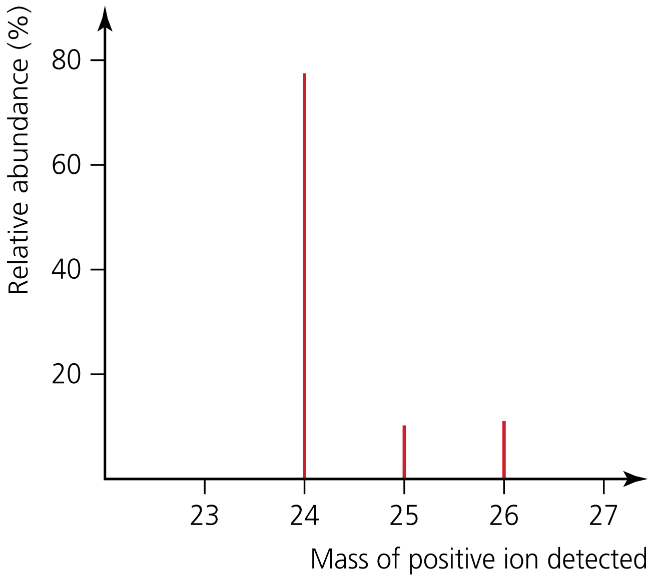


Figure 16.30 The mass spectrum of naturally occurring magnesium atoms showing the percentage abundances of three isotopes with their masses

Modern mass spectrometers neutralize the positive ions with electrons and count the number of ions directly, before amplifying the signal. The results are displayed on a computer screen in the form of a 'bar chart' (Figure 16.30).

Mass spectrometers are sensitive detectors of isotopes based on their masses. A number of satellites and spacecraft are equipped with mass spectrometers to allow them to identify the small numbers of particles intercepted in space. For example, the SOHO satellite uses a mass spectrometer to analyse the solar wind. They are also used in carbon dating (Chapter 7).

44 Describe the structure and operation of a Bainbridge mass spectrometer.

45 Find out how mass spectrometers on missions to Mars have been used to study the planet. What information have they provided to scientists?

46 Unipositive neon-21 ions, $^{21}Ne^+$, travelling with a velocity of $2.50 \times 10^5\,m\,s^{-1}$, enter a magnetic field of value 0.80 T, which deflects them into a circular path. Calculate the radius of the circular path.

Additional Perspectives

The velocity selector

The **velocity selector** uses an electric field, E, and a magnetic field, B, which are at right angles to each other, and also at right angles to the initial direction of motion of the positive ions passing through them. These fields both exert a force on the stream of positive ions. The two forces are arranged to act in opposite directions, as shown in Figure 16.31. The force on a particle of charge q due to the electric field, F_E, remains constant, but the force due to the magnetic field, F_B, varies according to the velocity, v, of the particle.

$$F_E = qE \qquad F_B = Bqv$$

If particles are to pass through the selector undeviated, then:

$$F_E = F_B$$

Therefore $qE = Bqv$, so:

$$v = \frac{E}{B}$$

The velocity that is selected can be varied by changing the ratio of the strength of the magnetic and electric fields.

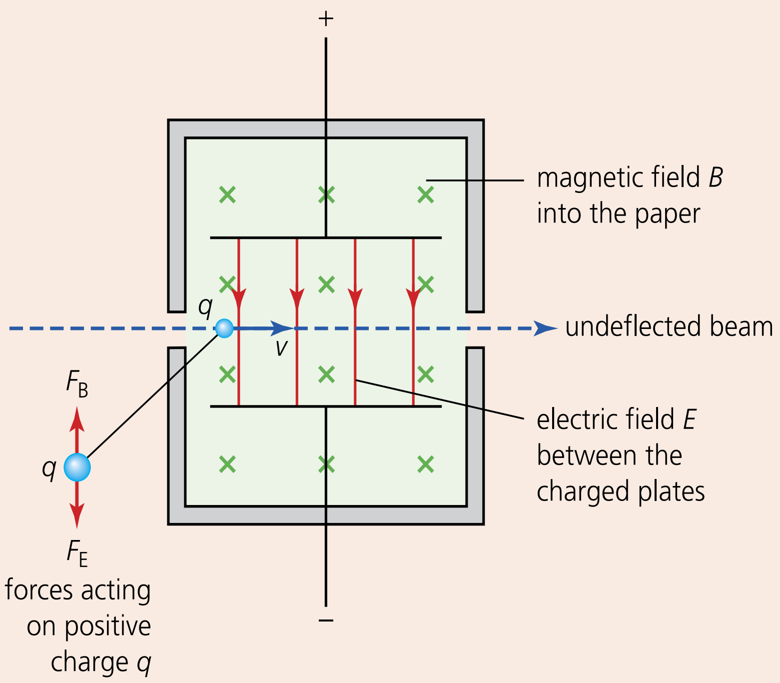


Figure 16.31 Velocity selector (of a mass spectrometer)

Question

1 It is required to select charged ions which have a speed of $3.5 \times 10^6\,m\,s^{-1}$. The electric field strength in the velocity selector is $4.5 \times 10^4\,V\,m^{-1}$. Calculate the magnetic field strength required in teslas.

Nuclear energy levels

B.2.3 Describe one piece of evidence for the existence of nuclear energy levels.

The nucleus is a quantum system and has discrete energy levels. The observation that the energies of gamma rays are discrete provides strong experimental evidence for the existence of nuclear energy levels. (This is in contrast to beta decays, in which the electron or positron has a continuous range of energies – see the next section on radioactive decay.)

Alpha decay

During alpha decay an alpha particle is ejected from the nucleus. The smaller alpha particle has the higher speed and the parent nucleus moves backwards (recoils) at a much lower speed. This is an example of Newton's third law of motion (Chapter 2). Energy is released during alpha decay. In a simple alpha decay, this energy is carried away as the kinetic energy of the alpha particle and the recoiling nucleus.

For a particular radioactive isotope decaying by simple alpha emission, the products of the decay are always the same for each nucleus – the daughter nuclide and an alpha particle. Plotting the proportion of alpha particles with a particular energy against alpha particle energy produces an energy spectrum. For the simple two-particle decay of radon-220, the graph shows that the energy of the alpha particles is discrete (Figure 16.32). Since these particles have the same mass in each decay, they always share the energy in the same way. Thus the energy involved in this decay process is also discrete.

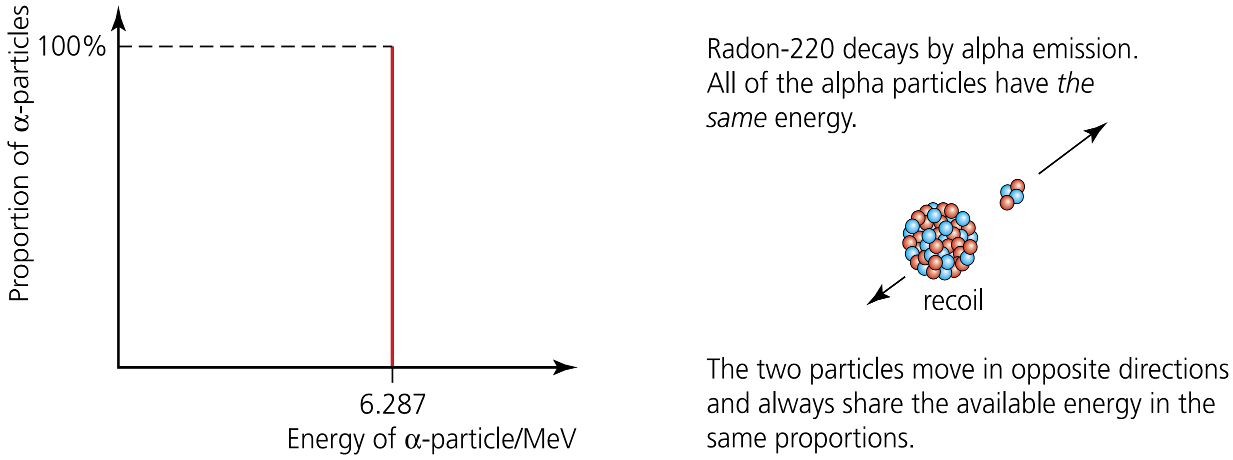


Figure 16.32 Energy spectrum for the alpha decay of radon-220 to plutonium-216

Gamma rays

Evidence for the existence of nuclear energy levels comes from studying gamma ray emissions from radioactive nuclei. These emissions do not change the numbers of protons and neutrons in the nucleus, but remove energy from the nucleus in the form of gamma rays (photons).

The gamma rays have energies that are discrete and distinctive for the nucleus. This suggests that the gamma rays are emitted as a result of **nuclear transitions** from an excited state with a higher energy level to a lower energy level, similar to electronic transitions in the atom (Chapter 7). For example, as shown in Figure 16.33, when carbon-15 decays by beta emission, some of the nitrogen-15 daughter nuclei are left in an excited state. These excited nuclei release gamma rays of a particular frequency, and hence energy, to reach the lowest energy level (the ground state).

The photons emitted by radioactive nuclei carry more energy, and therefore have higher frequencies, than photons emitted as a result of transitions involving electrons. This means that the energy changes involved in nuclear processes are much larger than those involved in transitions of electrons.

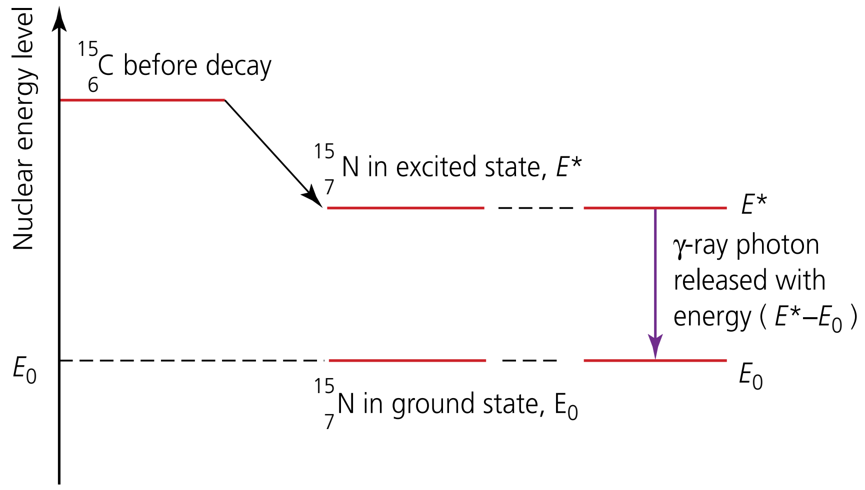


Figure 16.33 The gamma rays released when nitrogen-15 nuclei fall from an excited state to the ground state have a characteristic frequency

47 Figure 16.34 shows the lowest two energy levels for the nucleus of a heavy atom.
- **a** How do nuclear energies compare with electron energies?
- **b** Calculate the wavelength of the electromagnetic radiation emitted in a nuclear change involving a transition from the first excited state to the ground state.
- **c** To what part of the electromagnetic spectrum does this photon belong?

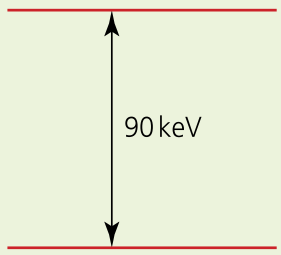


Figure 16.34

Radioactive decay

Beta energy spectra

B.2.4 Describe β^+ decay, including the existence of the neutrino.

Calculations using the masses of nuclei suggest that beta particles emitted from a radioactive source should all have the same energy, in a similar way to alpha particles. However, measurements show that although the *maximum* kinetic energy of the beta particles is characteristic of the beta source, the particles are emitted with kinetic energies that vary continuously, from zero to the maximum (Figure 16.35).

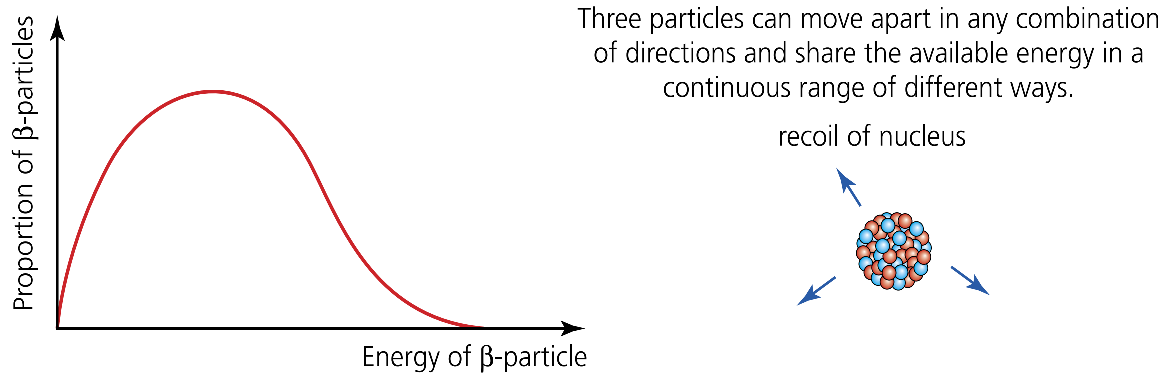


Figure 16.35 Typical energy spectrum for beta decay

This unexpected observation was first recorded in 1928 and suggested that energy and momentum might not be conserved. However, in 1933 the Austrian physicist Wolfgang Pauli suggested that another undetected particle was also involved in beta decay. This particle was hypothesized to be neutral (otherwise the electrical charge would not conserved), have little or no mass (otherwise the energy curve of the beta particles would be of a different shape), and be able to carry the kinetic energy not carried away by the nucleus or beta particle. This new particle was named the **neutrino** (and its antiparticle, the antineutrino). Because the neutrino interacts weakly with matter, it was not until 1955 that evidence for its existence was obtained.

There are two kinds of beta decay; (i) **beta-negative decay** (see Chapter 7) in which a neutron changes into a proton with the emission of an electron and an antineutrino; (ii) **beta-positive decay**, in which a proton changes into a neutron with the emission of a positron (positive electron) and a neutrino.

The full beta decay equations then become:

Beta-negative decay	*Beta-positive decay*
$n \rightarrow p + e^- + \bar{\nu}_e$	$p \rightarrow n + e^+ + \nu_e$

The neutrinos carry away the energy not carried away by the beta particle (positron or electron). It can be shared in any ratio, explaining the continuous spectrum of beta particle energies. The three emissions can vary in relative directions.

48 Describe beta decay (negative and positive) and beta energy spectra and explain why the neutrino was postulated to account for these spectra.

Additional Perspectives

Quarks

Protons and neutrons are not fundamental particles. They are composed of smaller particles called quarks. The six different types of quarks are known as *flavours*. The two most stable quarks are the *up* and *down* quarks. A proton is composed of two up quarks and one down quark; a neutron is composed of two down quarks and an up quark. Figure 16.36 summarizes what is happening inside protons and neutrons when they undergo beta decay (the production of neutrinos is ignored).

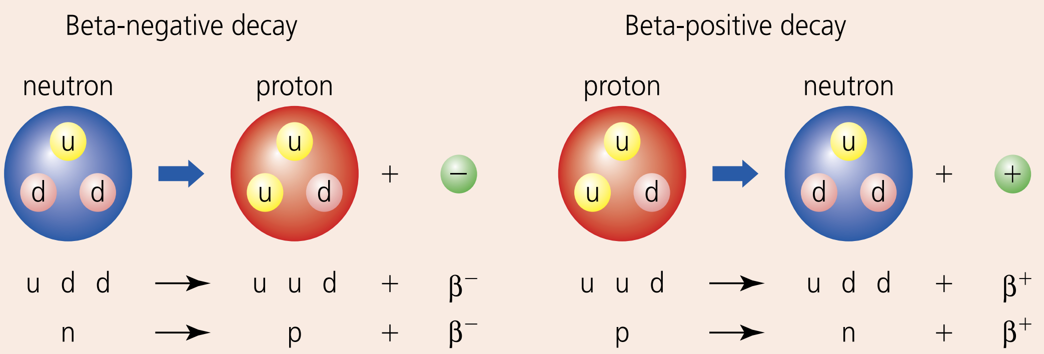


Figure 16.36 Beta decay can be explained by quarks changing their flavours. In beta-negative decay, a down quark in a neutron decays to form an up quark, and an electron is emitted. In beta-positive decay, an up quark in a proton decays to form a down quark, and a positron is emitted

Question

1 Find out why free quarks will never be observed.

Activity and decay constant

B.2.5 **State** the radioactive decay law as an exponential function and **define** the *decay constant*.

In the dice experiment in Chapter 7, increasing the number of dice thrown increases the number of sixes that appear (this experiment provides a simple analogy for radioactive decay). Similarly, if the decay of a radioactive material is investigated, for example, by using a Geiger–Müller tube, then it is found that the greater the number of radioactive nuclides in the sample, the greater the rate of decay.

This can be described mathematically by the following expression:

$$-\frac{\Delta N}{\Delta t} \propto N$$

where delta, Δ, represents 'change in' and N represents the number of undecayed nuclides in the sample.

Therefore $\frac{\Delta N}{\Delta t}$ represents the rate at which the number of nuclides in the sample is changing, and hence $-\frac{\Delta N}{\Delta t}$ represents the rate of decay.

Introducing a constant of proportionality, λ, we get:

$$-\frac{\Delta N}{\Delta t} = \lambda N$$

The constant λ (lambda) is known as the **decay constant**. It has units of reciprocal of time (s^{-1}).

The decay constant is defined as the probability per unit time that any particular nucleus will undergo decay.

$$\lambda = \frac{-\Delta N}{N\Delta t}$$

In Chapter 7 we introduced the **activity**, A, of a radioactive source as the number of nuclei decaying per second. Activity is the same as the rate of decay, therefore:

$$-\frac{\Delta N}{\Delta t} = A$$

and so

$$A = \lambda N$$

Both of these equations are given in the IB *Physics data booklet.*

Activity is measured in **becquerels** (Bq), where 1 becquerel is 1 decay per second. The becquerel is named after Henri Becquerel, who shared a Nobel Prize for Physics with Pierre and Marie Curie for their work in discovering radioactivity.

Worked examples

17 The activity of a radioactive sample is 2.5×10^5 Bq. The sample has a decay constant of $1.8 \times 10^{-16}\,s^{-1}$. Determine the number of undecayed nuclei remaining in the sample at that time.

$A = \lambda N$

$2.5 \times 10^{-5} = (1.8 \times 10^{-16}) \times N$

$N = \frac{2.5 \times 10^5}{1.8 \times 10^{-16}} = 1.4 \times 10^{21}$

18 A radioactive sample emits alpha particles at the rate of 2.1×10^{12} per second at the time when there are 5.0×10^{20} undecayed nuclei left in the sample. Determine the decay constant of the radioactive sample.

$\lambda = \frac{A}{N} = \frac{2.1 \times 10^{12}}{5.0 \times 10^{20}} = 4.2 \times 10^{-9}\,s^{-1}$

19 A sample of a radioactive nuclide initially contains 2.0×10^5 nuclei. Its decay constant is $0.40\,s^{-1}$. What is the initial activity?

$A = \lambda N$

$A = 0.40 \times 2.0 \times 10^5 = 8.0 \times 10^4\,s^{-1}$

The equation $-\frac{\Delta N}{\Delta t} = \lambda N$ is important because it relates a quantity that can be measured $\left(-\frac{\Delta N}{\Delta t}\text{, the rate of decay or the activity}\right)$ to a quantity which cannot in practice be determined (N, the number of undecayed nuclei).

The equation $\frac{\Delta N}{\Delta t} = -\lambda N$ has the following solution:

$$N = N_0 e^{-\lambda t}$$

This equation is in the IB *Physics data booklet.*

In this equation N_0 represents the initial number of undecayed nuclei in the sample, and N represents the number of undecayed nuclei after time t.

Since A is proportional to N, the equation can be expresed as

$$A = \lambda N = \lambda N_0 e^{-\lambda t}$$

This equation is in the IB *Physics data booklet.*

Worked example

20 A radioactive element decays $\frac{7}{8}$ of its original mass in 12 days. Calculate the fraction of radioactive mass that is left after 24 days.

After 12 days $\frac{N}{N_0} = \frac{1}{8}$

therefore $\frac{N}{N_0} = \frac{1}{8} = e^{-\lambda(12)}$

$8 = e^{12\lambda}$

$\lambda = \frac{\ln 8}{12} = 0.173$

After 24 days $\frac{N}{N_0} = e^{-0.173(24)} = 0.0157 = \frac{1}{64}$

49 State the radioactive decay law as an exponential function and define the decay constant.

50 A sample of radium contains 6.64×10^{23} atoms. It emits alpha particles and has a decay constant $1.36 \times 10^{-11}\,s^{-1}$. How many atoms are left after 100 years?

51 A radioactive nuclide has a decay constant of $0.0126\,s^{-1}$. Initially a sample of the nuclide contains 10 000 nuclei.
a What is the initial activity of the sample?
b How many nuclei remain undecayed after 200 s?

Additional Perspectives

Exponential graphs

The graph of N against t for the relationship $N = N_0 e^{-\lambda t}$ has the shape known as an exponential decay curve. Any relationship where the rate of change of a quantity is proportional to the quantity will give a graph with the same shape. As well as radioactive decay, many other physical effects, for example the discharge of a capacitor (Chapter 14), can be modelled to a good approximation by a negative exponential function. This is why **e** and logarithms to the base e (ln) are useful in physics.

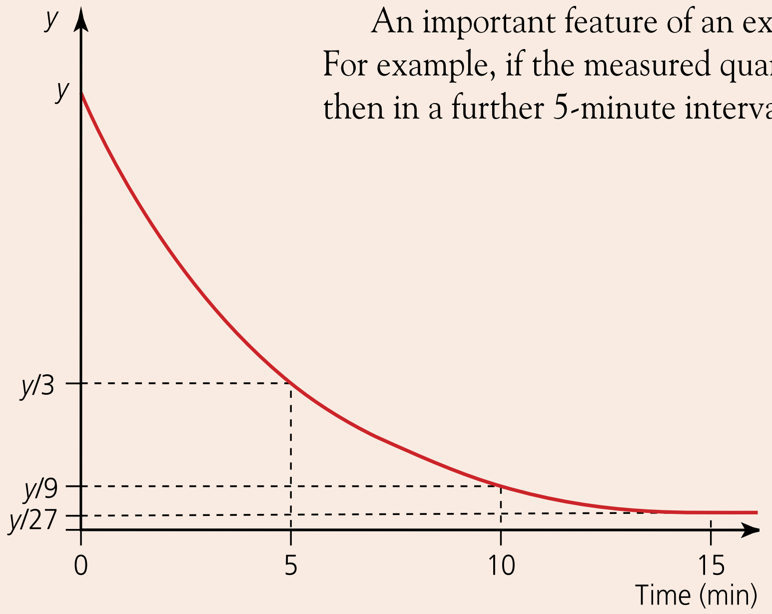


Figure 16.37 An exponential decay curve

An important feature of an exponential relationship is that the fractional change is constant. For example, if the measured quantity reduces to 1/3 of its original value in a 5-minute interval, then in a further 5-minute interval it will reduce to 1/3 of the value at the start of that interval, to 1/9 of the original value, and so on (Figure 16.37). This is true for any chosen fraction. Half-life is the time interval for which the constant fractional change is 1/2.

If the relationship between a quantity y and time is exponential, a plot of $\ln y$ (or $\log y$) against time will always produce a straight line.

Question

1 Construct a spreadsheet that plots an exponential decay curve where the user can adjust the half-life and decay constant.

Decay constant and half-life

B.2.6 Derive the relationship between decay constant and half-life.

Using the equation $N = N_0 e^{-\lambda t}$, we can derive an equation which relates the half-life, $T_{1/2}$, to the decay constant, λ.

For any radioactive nuclide, the number of undecayed nuclei after one half-life is, by the definition of half-life, equal to $N_0/2$, where N_0 represents the original number of undecayed nuclei. Substituting this value for N in the radioactive decay equation at time $t = T_{1/2}$ we have:

$$\frac{N_0}{2} = N_0 e^{-\lambda T_{1/2}}$$

Dividing each side of the equation by N_0:

$$\frac{1}{2} = e^{-\lambda T_{1/2}} \quad \text{or} \quad 2 = e^{\lambda T_{1/2}}$$

Taking natural logarithms (to the base e):

$$\ln 2 = \lambda T_{1/2}$$

So that:

$$T_{1/2} = \frac{\ln 2}{\lambda}$$

This equation is in the IB Physics *data booklet.*

or

$$T_{1/2} = \frac{0.693}{\lambda}$$

Worked examples

21 A radioactive sample gives a count rate of $100\,s^{-1}$ at a certain instant of time. After 100 s the count rate drops to $20\,s^{-1}$. The background count rate is measured to be $10\,s^{-1}$. Calculate the half-life of the sample. Assume that the count rate is a measure of the activity.

Initial count rate due to sample = $100\,s^{-1} - 10\,s^{-1} = 90\,s^{-1}$
Count rate due to sample after 100 s = $20\,s^{-1} - 10\,s^{-1} = 10\,s^{-1}$

$$A = A_0 e^{-\lambda t}$$

$$10 = 90 e^{-100\lambda}$$

$$\lambda = \frac{-\ln \frac{10}{90}}{100}$$

$$T_{1/2} = \frac{100 \ln 2}{-\ln \frac{1}{9}} = 32\,s$$

22 The radioactive element A has 6.4×10^{11} atoms and a half-life of 2.00 hours. Radioactive element B has 8.0×10^{10} atoms and a half-life of 3.00 hours. Calculate how much time will pass before the two elements have the same number of radioactive atoms.

For A, $\lambda_A = \frac{0.693}{2.00} = 0.347\,h^{-1}$ for B, $\lambda_B = \frac{0.693}{3.00} = 0.231\,h^{-1}$

At time t: $N_A = (6.4 \times 10^{11})e^{-0.347t}$ $N_B = (8.0 \times 10^{10})e^{-0.231t}$

For $N_A = N_B$: $(6.4 \times 10^{11})e^{-0.347t} = (8.0 \times 10^{10})e^{-0.231t}$

$$8 = \frac{e^{-0.231t}}{e^{-0.347t}}$$

$8 = e^{0.116t}$

$\ln 8 = \ln (e^{0.116t})$; $2.079 = 0.116t$

$t = 18\,h$

After 18 hours the two elements will have the same number of active atoms.

52 Radioactive carbon in a leather sample decays with a half-life of 5770 years.
a What is the decay constant?
b Calculate the fraction of radioactive carbon remaining after 10 000 years.

Measurement of half-life

B.2.7 Outline methods for measuring the half-life of an isotope.

The method used to measure the half-life of a radioactive element depends on whether the half-life is relatively long or short. If the activity of the sample stays virtually constant over a few hours then it has a relatively long half-life. However, if its activity decreases during a few hours then the radioactive element has a relatively short half-life.

Isotopes with long half-lives

$$\frac{\Delta N}{\Delta t} = -\lambda N$$

If the activity ($\Delta N/\Delta t$) of a source can be determined, then the decay constant (and therefore the half-life) can be calculated if the number of undecayed atoms, N, is known. Theoretically this is straightforward, but determining practically the number of atoms of a particular isotope in a sample containing a mixture of isotopes is not easy and requires sophisticated equipment like a mass spectrometer.

However, for a *pure* sample of mass m, the number of atoms of the isotope can be determined from the relative atomic mass (A_r) and Avogadro's contant, N_A as follows:

$$N = \frac{mN_A}{A_r}$$

Therefore if the activity ($\Delta N/\Delta t$) is measured we can calculate the half-life, $T_{1/2}$, from the equation:

$$\frac{\Delta N}{\Delta t} = \frac{-\lambda m N_A}{A_r} = \frac{-0.693 m N_A}{T_{1/2} A_r}$$

53 The experimentally determined activity from 1.0 gram of radium-226 is 1.14×10^{18} alpha particles per year. Calculate its half-life.

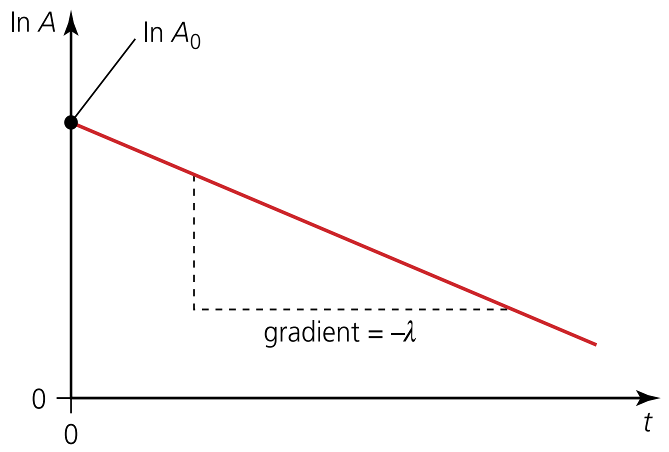


Figure 16.38 A logarithmic–linear graph to show exponential decay of a radioactive nuclide

Isotopes with short half-lives

If an isotope's half-life is less than several days then it can be measured directly. This is done by measuring the count rate over a short period of time at regular intervals. The measured count rate is assumed to be proportional to activity. A graph of activity, *A* (or count rate) against time, *t*, is plotted and the half-life obtained directly from the graph (Chapter 7). (In practice the graphs will be count rates, not activities.)

Alternatively, a graph can be plotted of natural logarithm of the activity ($\ln A$) against time, *t*. This will give a straight line with a gradient of $-\lambda$ (Figure 16.38). Since $T_{1/2} = \frac{0.693}{\lambda}$, the half-life can be calculated. This is a better method because the reliability of the data can be more readily assessed by seeing how close to the straight line of best fit the data points lie.

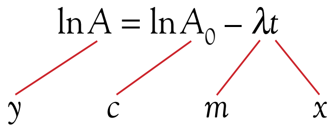
This approach relies on transforming the equation $A = A_0 e^{-\lambda t}$ which describes the activity in radioactive decay.

$$A = A_0 e^{-\lambda t}$$

Taking natural logarithms:

$$\ln A = \ln A_0 - \lambda t$$

The equation can be compared to the equation for a straight line ($y = mx + c$):



Thus the gradient is equal to $-\lambda$.

In accurate experiments where measurements of count rate are taken, an allowance should be made for the effect of background radiation. The two graphs in Figure 16.39 show the effect of allowing for the background count. The corrected count rate is assumed to be proportional to the activity.

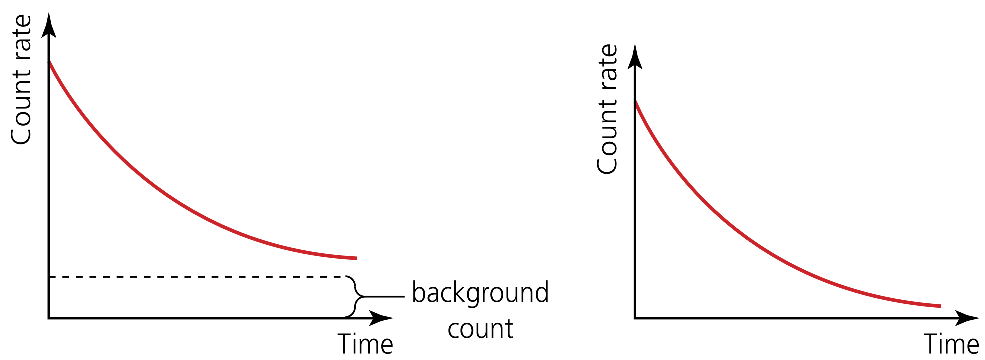


Figure 16.39 Uncorrected and corrected count rates for a radioactive isotope

When the half-life of the isotope is very small, less than a second, then both of these methods are unsuitable. Such half-lives may be found from tracks in a cloud or bubble chamber.

B.2.8 Solve problems involving radioactive half-life.

54 Radium-226 has a long half-life (>1000 years); radium-227 has a half-life of 42 minutes. Outline how the two half-lives of these radium isotopes can be determined experimentally.

55 The following data were obtained from the decay of caesium-130:

Time/s	Activity of ^{130}Cs/disintegrations s^{-1}
0	200
500	165
1500	113
2500	79
3500	54
4500	38
5500	26

Use a spreadsheet to plot a graph of $\ln A$ against t to determine the decay constant and hence the half-life (to the nearest integer).

56 A sample contains the atoms of radioactive element A and another sample contains the atoms of a radioactive element B. After a fixed length of time, it is found that $\frac{7}{8}$ of atoms A and $\frac{3}{4}$ of atoms B have decayed.

Calculate the value of the ratio $\frac{\text{half-life of element A}}{\text{half-life of element B}}$.

SUMMARY OF KNOWLEDGE

B1 Quantum physics

- Electrons may be emitted from cleaned metal surfaces if the metal is illuminated by electromagnetic radiation, usually ultraviolet radiation. This phenomenon is called photoelectric emission.
- Electron energies are often measured in electronvolts (eV), where $1\,\text{eV} = 1.6 \times 10^{-19}\,\text{J}$.
- For photoelectric emission to occur, there is a threshold frequency (energy) below which no electrons are released. Above the threshold frequency photoelectrons are released at a rate proportional to the intensity of the light. The energy of the emitted electrons is independent of the intensity of the incident radiation.
- Photoelectric emission cannot be explained by the wave model of light. It is necessary to apply quantum theory, in which electromagnetic radiation is described as consisting of packets of energy called photons.
- The energy of a photon, E, is given by the Planck relationship: $E = hf$, where h represents Planck's constant and f represents the frequency of the electromagnetic radiation.
- The work function, ϕ, of a metal is the minimum energy needed to free an electron from the surface of a metal. Different metals have different work functions.
- The Einstein photoelectric equation is: $hf = \phi + \frac{1}{2}m_e v_{max}^2$ or $hf = \phi + E_{max}$
- The threshold frequency for photoelectric emission, f_0, is given by: $hf_0 = \phi$
- The maximum kinetic energy of the photoelectrons is determined by the frequency of the electromagnetic radiation and not by the intensity.
- If a reverse potential is applied the kinetic energy lost by photoelectrons = electric potential energy gained by those electrons: $\frac{1}{2}m_e v^2 = eV$. Therefore:

$$hf = hf_0 + eV \qquad \text{or} \qquad V = \left(\frac{h}{e}\right)f - \frac{hf_0}{e}$$

- All moving particles, in principle, show wave-like properties. However, only electrons and atoms, which have very small mass, will show their wave nature (diffraction and interference) in experiments.
- The de Broglie wavelength, λ, is given by $\frac{h}{p}$, where p represents the momentum of the electron and h represents Planck's constant.
- If a charged particle carrying a charge of q coulombs is accelerated by applying a potential difference of V volts, then the de Broglie wavelength of the particle is given by the following relationship:

$$\lambda = \frac{h}{\sqrt{2m_e qV}}$$

- de Broglie waves are matter waves. The amplitude of the matter wave at a point represents the probability of the particle being at that position.
- The Davisson–Germer experiment involved directing a beam of electrons at a thin metal foil. The electrons were diffracted, providing experimental evidence for de Broglie's hypothesis.
- At certain angles there were peaks in the intensity of the scattered electron beam. These peaks indicated wave behaviour for the electrons, (and could be interpreted to give values for the lattice spacing in the metal crystal.
- Light is regarded as behaving as a stream of particles (quanta) or as a wave, depending on the phenomenon being described.
- A continuous spectrum has a complete range of wavelengths.
- An emission spectrum can be produced by passing light from a sample of excited gas through a slit to generate a narrow beam which is then directed onto a prism or diffraction grating, which causes the light to undergo dispersion.
- An emission spectrum is produced when matter emits electromagnetic radiation. In the visible region it consists of a series of sharp coloured lines on a black background.
- An absorption spectrum is produced using the same apparatus (spectrometer) except that a beam of white light is passed through a sample of cool gas (maintained at low pressure).
- An absorption spectrum is produced when electromagnetic radiation travelling through matter is absorbed. In the visible region it consists of a series of sharp black lines on a coloured background.
- An atom is a quantized system and has definite electron energy levels.
- The lowest energy level is the ground state, and all the other energy levels are excited states. The highest possible energy level occurs when the atom is ionized.
- The values of the energy levels are governed by the principal quantum number, n.
- The energy level at ionization is given the value of 0 eV; all other energy levels have negative values.
- An electronic transition is the process in which an atom changes its quantum state by absorbing or emitting a certain discrete amount of energy.
- The amount of energy, ΔE, absorbed or emitted in an electronic transition is given by:

$$\Delta E = E_2 - E_1 = hf = \frac{hc}{\lambda}$$

- Spectra can be interpreted in terms of the transition of electrons in atoms between different energy levels.
- A line in an absorption spectrum is formed when an electron moves from a low energy level to a higher energy level. This is known as excitation.
- A line in an emission spectrum is formed when an electron moves from a high energy level to a lower energy level.

- For a photon to be absorbed it must have exactly the correct amount of energy to raise an electron from a lower energy level to a higher energy level.
- Absorption and emission spectra support the concept that atoms have quantized energy levels.
- The 'electron in a box' is a simple quantum mechanical model which shows how boundary conditions that a wavefunction must satisfy lead to the quantization of energy.
- The system consists of an electron with a mass m in a one-dimensional region of space of length L: the potential is zero for $0 \leq x \leq L$ and infinite elsewhere.
- The de Broglie waves associated with the electron will be standing waves of wavelength $\frac{2L}{n}$ where L is the length of the box and n is a positive integer.
- The kinetic energy of the electron in the box is given by: kinetic energy $(E_K) = \frac{n^2h^2}{8m_eL^2}$, where m_e is the mass of the electron and n is the energy level.
- The Schrödinger wave model builds on the concept of de Broglie's matter waves.
- The electron is described in terms of a wavefunction, Ψ, where at any instant of time, the wavefunction has different values at different points in space.
- A wavefunction is a solution of Schrödinger wave equation and is a mathematical description of an electron as a wave.
- The probability of finding the electron at any point in space is given by the square of the absolute amplitude of the wavefunction at that point.
- Orbitals are three dimensional regions in space where electrons are likely to be found. Orbitals have different energies and shapes.
- The wavefunctions for electrons in different energy levels can be determined by solving the Schrödinger wave equation.
- The Heisenberg uncertainty principle states that values cannot be assigned, with full precision, for position and momentum, or for energy and time, for a particle.
- If a particle has a well-defined de Broglie wavelength, then its momentum is known precisely, but there is no knowledge of its position.
- Measurements of time and energy, and position and momentum are linked variables and described by the energy–time and position–momentum uncertainty principles:

  $$\Delta x \Delta p \geq \frac{h}{4\pi} \quad \text{and} \quad \Delta E \Delta t \geq \frac{h}{4\pi}$$

  where ΔE represents the uncertainty in the measurement of energy, Δx represents the uncertainty in the measurement of position, Δp represents the uncertainty in the measurement of momentum and Δt represents the uncertainty in the measurement of time.

B2 Nuclear physics

- When a fast-moving alpha particle approaches a gold nucleus head-on, it will be directed straight back along the same path. It will get close to the nucleus but not collide with the nucleus, owing to the action of repulsive electrostatic forces. This effect is known as Coulomb scattering.
- An alpha particle approaching a gold nucleus slows down as it gains electrical potential energy and loses kinetic energy.
- At the closest approach, the alpha particle is temporarily stationary and all its energy is electrical potential energy.
- Applying Coulomb's law at closest approach:

  $\frac{q_1q_2}{4\pi\varepsilon_0 r^2} = E_K$, where q_1 is the charge on an alpha particle, q_2 is the charge on the gold nucleus and r is an upper estimate for the radius of the gold nucleus.
- The $1/r$ hill is a gravitational model showing how the electrical potential varies round a charged particle. The elevation of the hill above the bench top represents the potential and the steepness of the hill represents the field.

- A Bainbridge mass spectrometer can be used to measure the specific charge to mass ratio of a positively charged particle and to detect isotopes of an element and measure their percentage abundances.
- The principle of the mass spectrometer is to use a magnetic field to deflect moving positive ions (in the gaseous state). If a moving ion (accelerated by an electric potential) enters a magnetic field of constant strength it will follow a circular path.
- The magnetic force provides the centripetal force required:

$$Bqv = \frac{mv^2}{r} \quad \text{or} \quad r = \frac{mv}{Bq}$$

  where B is the magnetic field strength, q the charge on the positive ion, r is the radius of the ion's path, m is the mass of the ion and v is its velocity.
- If the ions have the same charge, q, and they are all selected to be travelling at the same velocity, v, then the radius of the circle will depend on the mass of the ion. An ion with larger mass will travel in a circle with a larger radius.
- The nucleus is a quantum system and its energy values are quantized into discrete energy levels.
- When an alpha particle or a gamma photon is emitted from the nucleus only discrete energies are observed. These energies correspond to the difference between two discrete nuclear energy levels.
- Beta energy spectra are continuous – beta particles are emitted with a range of kinetic energies.
- Neutrinos are small particles with no charge and negligible mass. They pass through matter with little interaction. They carry variable amounts of energy when formed during beta decay.
- Beta positive decay causes no change to the nucleon number of the parent nuclide but causes an increase of one in the proton number.
- The half-life, $T_{1/2}$, of a radioactive nuclide is the time taken for the number of undecayed nuclei to be reduced to half the original number. This is constant for a given isotope.
- The activity of a radioactive source is the number of decays per unit time.
- The activity $\frac{\Delta N}{\Delta t}$ of a radioactive source is related to the number N, of undecayed nuclei by the equation: $\frac{\Delta N}{\Delta t} = -\lambda N$, where λ is the decay constant.
- The decay constant is defined as the probability of decay per unit time of a nucleus. The larger the value of the decay constant the more rapid the radioactive decay.
- The number N of undecayed nuclei in a radioactive sample after time t is given by the equation $N = N_0 e^{-\lambda t}$, where N_0 is the number of undecayed nuclei at the start of timing.
- Since the activity, A, of a source is directly proportional to N, it follows that $A = A_0 e^{-\lambda t}$, where A_0 is the initial activity of the source. Taking logarithms to the base e gives $\ln A = \ln A_0 - \lambda t$.
- A graph of $\ln A$ against t is a straight line, with the gradient equal to $-\lambda$.
- The half-life of a short-lived isotope can be found by measuring the corrected count rate (proportional to activity) over a period of time. The decay constant can be found from a plot of $\ln A$ against t.
- The half-life, $T_{1/2}$, and the decay constant, λ, are related by the equation:
  $T_{1/2} = \frac{0.693}{\lambda}$, which is derived from $N = N_0 e^{-\lambda t}$.
- If the half-life is long, then the activity will be effectively constant over a period of time. The number of nuclei can be calculated from the mass of a pure sample, its atomic mass and Avogadro's constant. The decay equation can then be used to calculate the half-life.

Examination questions – a selection

All of the IB questions and IB style questions from Papers 1 and 2 which are to be found at the end of Chapter 13 are suitable for the revision of Option B, although the actual option examination paper (Paper 3) does *not* contain any multiple choice type questions.

Paper 3 IB questions and IB style questions

Q1 This question is about the spectrum of atomic hydrogen.

a The diagram represents the principal lines in the visible spectrum of atomic hydrogen.

Outline how the spectrum can be produced and observed in the laboratory. [3]

b Calculate the difference in energy in eV between the energy levels in the hydrogen atom that give rise to the red line in the spectrum. [2]

Standard Level Paper 3, November 2010, QB2

Q2 This question is about nuclear physics and radioactive decay.

a Define the decay constant of a radioactive nuclide. [1]

b **i** Plutonium-239 (Pu-239) has a half-life of 2.4×10^4 years. Show that the decay constant of Pu-239 is approximately 3×10^{-5} year^{-1}. [1]

ii Calculate the time taken for the activity of a freshly-prepared sample of Pu-239 to fall to 0.1% of its initial value. [2]

Standard Level Paper 3, May 2009 TZ2, QB3

Option

C

Chapter 17
Digital technology

STARTING POINTS

- A flow of electrons is an electric current.
- Light travels in straight lines from a source in the form of progressive waves.
- Light waves are described by their amplitude (brightness or intensity), wavelength (colour), frequency and speed.
- The frequency of a wave is the number of waves that pass a point in one second.
- The wave equation relates the speed, frequency and wavelength of a wave. Speed = frequency × wavelength. $v = f\lambda$
- Light waves can exhibit the properties of reflection and interference.
- Light (and other electromagnetic radiations) can be regarded as a stream of energy 'packets' known as photons.
- The energy of a photon is directly proportional to its frequency: $E = hf$
- Interference patterns may be produced by the superposition of two or more waves.
- Constructive interference occurs when two sets of waves meet at positions where the waves arrive in phase; crests/troughs of the waves meet and produce larger crests/troughs.
- Constructive interference occurs at positions where the path difference is equal to 0, λ, 2λ, etc., where λ represents the wavelength.
- Destructive interference occurs when two sets of waves meet at positions where the waves arrive exactly out of phase; crests of a wave meet the troughs of the other and cancel each other out.
- Destructive interference occurs where the path difference is an odd multiple of $\frac{\lambda}{2}$.
- Diffraction of a wave is greatest when it passes through a gap comparable to its wavelength.
- The angles at which waves passing through an aperture are diffracted depend on the ratio of wavelength to aperture size.
- Resolution is the ability to distinguish the images of two separate objects from one another.
- The electromotive force (emf) of a source of electrical energy is the total energy transferred when unit charge passes through it.
- In an ohmic component, the current, I, is proportional to the voltage (p.d.), V, across it, as expressed by Ohm's law.
- Voltmeters are connected in parallel with components in order to measure the potential difference across them; ammeters are connected in series to measure the currents through components.
- Resistance (ohm, Ω) = $\frac{\text{voltage (V)}}{\text{current (A)}}$; $R = \frac{V}{I}$.
- A resistor is an electrical component used in a circuit to provide electrical resistance.
- Internal resistance is the resistance to the flow of current within a source of electrical energy.
- The sum of currents into and out of a junction in an electrical circuit must be equal to zero.
- The sum of the voltages (p.d.s) around a closed loop of an electric circuit must equal zero.
- The total resistance of resistors in series is equal to the sum of the individual resistances.
- The inverse of the total resistance of resistors in parallel is equal to the sum of the inverses of the individual resistances.
- The voltages (p.d.s) across resistors in series are in proportion to the values of the resistances. A circuit designed to utilize the voltage across one of two resistors connected in series is called a potential divider.
- Many electronic circuits are designed to respond to signals that change continuously over a range of values. These are known as analogue circuits.

Introduction

In earlier chapters we met examples of uses of electricity for lighting, heating, cooling (air conditioners and refrigerators) and in electric motors. More recently, electricity has also been vital to digital technology, which has made huge changes to the way people study, communicate, work and relax.

Digital technology is also the basis for digital devices, such as cameras and video recorders, MP3 players, CD players, games machines, mobile phones, tablets and computers. Digital devices also include input, output and communication devices that allow computer access to the vast amount of information and data that can be found on the Internet (the World Wide Web).

C1 Analogue and digital signals

When we communicate, we send each other information. Information comes in many forms, for example:

- text (letters, numbers and symbols)
- still pictures, for example, drawings, paintings and photographs
- speech, sound and music
- video and animation.

The transfer of information needs an agreed and recognized code. For example, the sounds of words, smoke signals, written symbols and the sequence of coloured lights at traffic signals are all examples of codes. The transfer of information between modern electronic devices requires the use of standard codes (e.g. the ACSCII code). Standard codes are very important because they allow digital devices from different manufacturers to share data. The use of electronic technology for the storage, processing, validation (checking) and transfer of information involves converting the original information into a different form, usually an electrical signal, using a coding process, as shown in Figure 17.1. The electrical signal can be converted back when needed. This can be done using digital or analogue techniques.

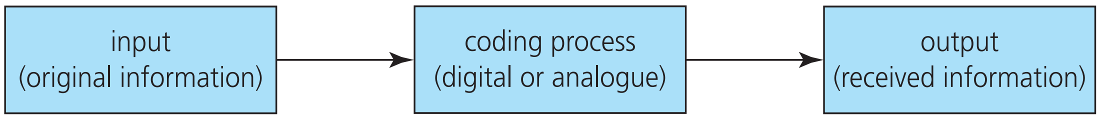


Figure 17.1 The coding of information

Any information signal that has the same variations with time as the information itself is known as an **analogue signal**. For example, the voltage produced by a microphone in a recording studio varies in the same way as the sound waves that are detected by the microphone. The voltage variation is due to the loudness or amplitude of sound detected by the microphone. The voltage output from the microphone is an analogue signal. The graph in Figure 17.2 represents an analogue signal from a microphone that varies continuously within the range +6 V and –6 V.

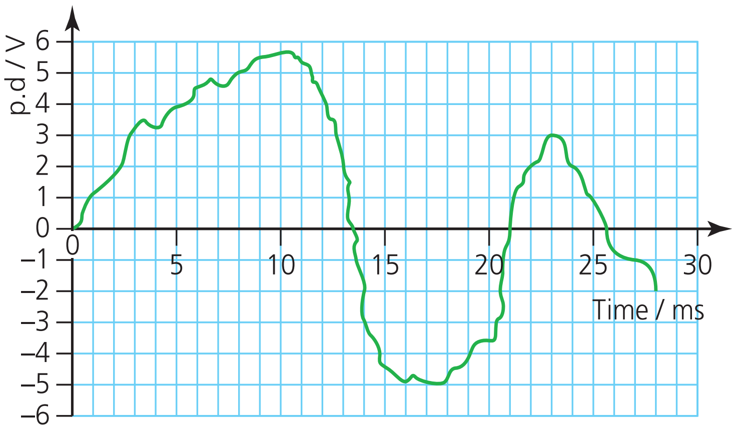


Figure 17.2 An analogue signal

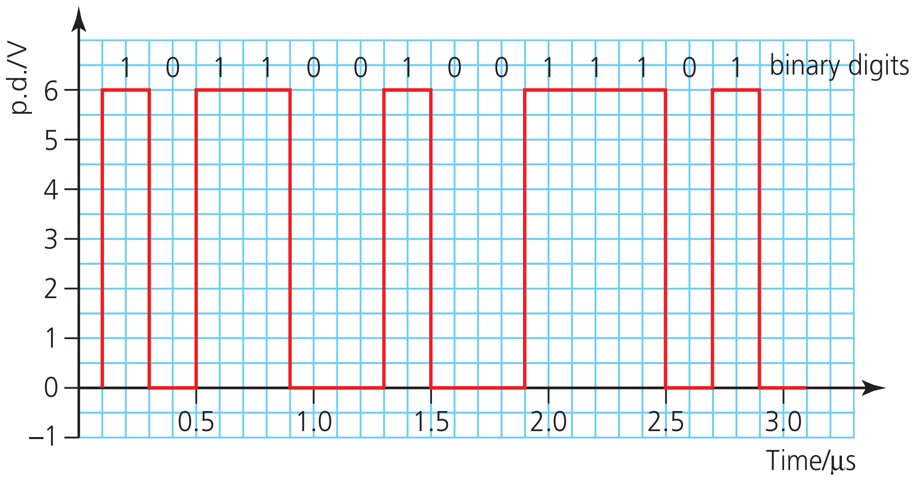


Figure 17.3 A digital signal

A **digital signal** consists of a series of 'highs' and 'lows' with no values between the 'highs' and 'lows'. The discrete data in the signal is transmitted as a series of binary ones (1s) and binary zeros (0s). The binary number system (base 2) represents numbers using two symbols, 0 and 1.

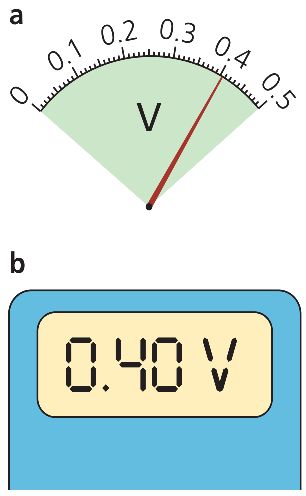


Figure 17.4 Two different voltmeters reading the same value of potential difference:
a analogue voltmeter
b digital voltmeter

The signal in Figure 17.3 is a digital signal which has two values only: +6 V or 0 V, representing respectively a binary one or binary zero. Each '1' or '0' lasts for 0.2 microseconds.

Figure 17.4 shows an analogue and a digital voltmeter reading the same value of potential difference. The analogue voltmeter shows *all* values of potential difference from the minimum to the maximum value on the scale. Intermediate values can be found by interpolating (reading between) the gradations (marks) of the scale. The digital voltmeter is accurate to ± 0.01 V. The digital voltmeter reading is said to be quantized in steps of 0.01 V.

There are many advantages in using digital electronic circuits to store, process and transmit data rather than analogue circuits. Today virtually all communication systems and electronic devices are digital.

Analogue signals can be converted to digital signals by an analogue-to-digital converter (ADC). This encodes or changes the analogue signal into a digital form via a process known as quantization. Computers can only process and store digital signals. Figure 17.5 shows how a computer can be used to log, convert and display data from a temperature sensor.

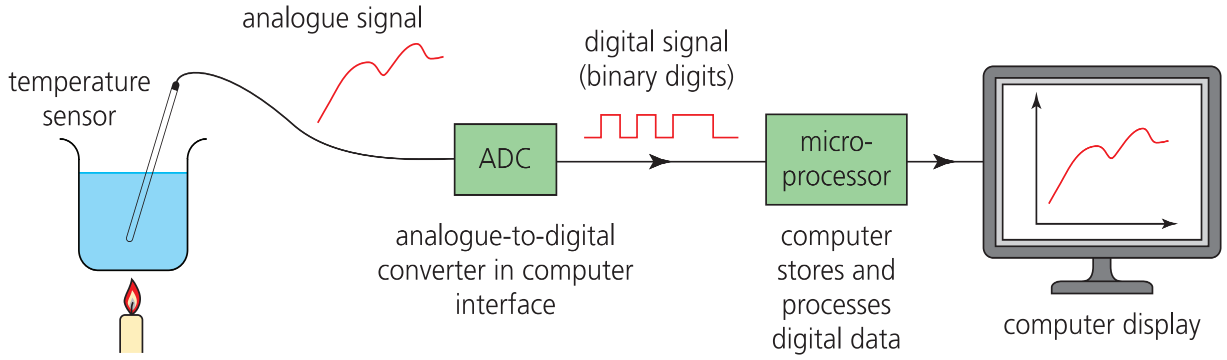


Figure 17.5 The use of a temperature sensor and a data logger

1 Research the use of semaphore flags and Morse code in coding text messages.

2 Give examples of analogue and digital measurements from your practical investigations.

3 Find out about the Google projects to digitally store, by scanning, thousands of works of literature and famous artwork from libraries and art galleries all over the world.

Binary numbers

C.1.1 **Solve** problems involving the conversion between binary numbers and decimal numbers.

A binary number is a base 2 number; a decimal number is a base 10 number. Table 17.1 shows some decimal numbers and their equivalents in base 2 or binary notation. A binary number is formed from a number of binary digits, or **bits**. All the binary numbers shown in Table 17.1 are four-bit numbers. Zeros are added as extra bits to the left of any number that is less than four bits. Binary numbers may also be called digital numbers. Digital signals are often four-bit binary numbers, or multiples of them: eight-bit, 16-bit, 32-bit, or even 64-bit.

Larger numbers would require digital numbers with more bits. When reading a digital number, the bit on the left-hand side of the digital number is the **most-significant bit** (MSB). This binary digit or bit has the highest value. The binary digit or bit on the right-hand side has the least value and is known as the **least-significant bit** (LSB).

Table 17.1 Decimal and four-bit binary numbers

Decimal number	Binary or digital number
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111

When the LSB is 1 and all other bits are 0, this corresponds to decimal number 1. When the second bit is binary 1, and all other bits are 0, this corresponds to the decimal number 2. When successive bits show binary 1 they correspond to decimal numbers 4, 8, 16, 32, 64, etc. (Table 17.2).

The binary number 1101 corresponds to the decimal number 8 + 4 + 0 + 1, or 13. The decimal number 11, which equals 8 + 0 + 2 + 1, corresponds to the binary number 1011.

Figure 17.6 shows how the place value you are familiar with in base 10 is used in base 2. A binary signal containing a series of binary 1s and 0s is known as a **word**. An eight-bit binary number or word occupies one **byte** of computer memory. Values stored on a compact disc (CD) are represented by a two-byte number (16 bits). Computer memory is measured in kilobytes (KB), megabytes (MB) and gigabytes (GB). 1 KB = 1024 (2^{10}) bytes; 1 MB = 1 048 576 bytes (2^{20}) and 1 GB = 1 073 741 824 (1024^3 or 2^{30}) bytes.

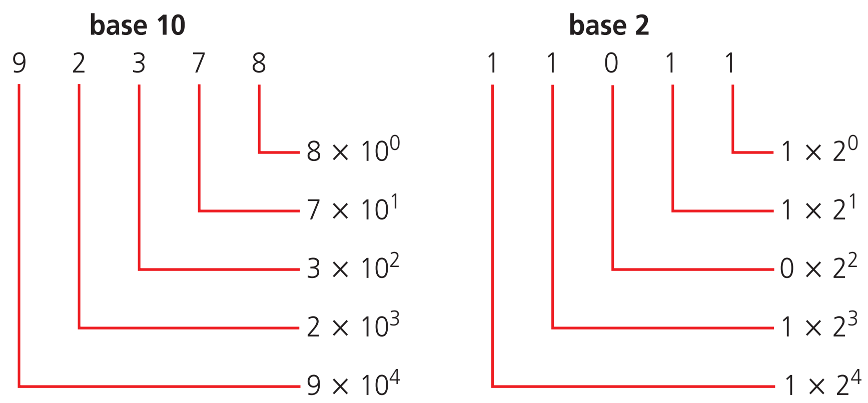


Figure 17.6 A decimal number and a binary number showing base 10 and base 2 place values

Table 17.2 Decimal numbers and their five-digit binary equivalents showing the powers to the base 2

Base 10	2^4	2^3	2^2	2^1	2^0	Base 2
0	0	0	0	0	0	00000
1	0	0	0	0	1	00001
2	0	0	0	1	0	00010
3	0	0	0	1	1	00011
4	0	0	1	0	0	00100
5	0	0	1	0	1	00101
6	0	0	1	1	0	00110
7	0	0	1	1	1	00111
8	0	1	0	0	0	01000
9	0	1	0	0	1	01001
10	0	1	0	1	0	01010
11	0	1	0	1	1	01011
12	0	1	1	0	0	01100
13	0	1	1	0	1	01101
14	0	1	1	1	0	01110
15	0	1	1	1	1	01111
16	1	0	0	0	0	10000
17	1	0	0	0	1	10001
18	1	0	0	1	0	10010
19	1	0	0	1	1	10011
20	1	0	1	0	0	10100

■ Additional Perspectives

ASCII code

A particular number of binary digits (bits) represents a fixed number of different values. Each additional bit doubles the number of different possible values that can be represented (Table 17.3).

English text contains a mixture of letters (combined to form of words), numbers, symbols and punctuation marks. Additional characters are also needed to represent formatting, such as new line and new paragraph. Computers and the Internet use a code known as ASCII (American Standard Code for Information Interchange). It is an eight-bit code representing 256 different possible characters and formatting codes. For example, 1000000 represents the 'at' symbol @ and 1000011 represents the letter C.

ASCII has been superseded by Unicode, a double byte (16-bit) character system designed to store and display a much wider range of letters and symbols. The extras include foreign languages, such as Mandarin, and mathematical/scientific symbols, plus space for future expansion.

Question

1 Research the problems that Unicode has in representing Chinese characters.

Table 17.3 Number of binary digits with the associated number of different possible digital values

No. of bits used	Bits used	No. of different possible values
1	0, 1	2
2	00, 01, 10, 11	4 (2^2)
3	000, 001, 010, 011, 100, 101, 110, 111	8 (2^3)
4	etc.	16 (2^4)
5	etc.	32 (2^5)
6	etc.	64 (2^6)
7	etc.	128 (2^7)
8	etc.	256 (2^8)
9	etc.	512 (2^9)
10	etc.	1024 (2^{10})

Worked examples

1 What is the least number of binary digits needed to express each of the decimal numbers (base 10) 10 and 2038 as a binary number?

10 – The *smallest* power of 2 that exceeds 10 is 4, since $2^4 = 16$ and is greater than 10. Hence four bits must be used in the binary representation of 10 (1010).

2038 – The smallest power of 2 that exceeds 2038 is 11, since $2^{11} = 2048 > 2038$. Hence 11 bits must be used in the binary representation of 2038 (11111110110).

2 How many decimal numbers can be represented with eight-bit words?

With eight bits the maximum number is $2^8 = 256$. These are the numbers from 0 to 255.

3 Express the six-bit binary number 11101 as a decimal number.

The binary number is shown below. The correct power of 2 has been placed under each binary digit.

1	1	1	0	1
2^4	2^3	2^2	2^1	2^0

$$\begin{aligned} \text{So } 11101 &= (1 \times 2^4) + (1 \times 2^3) + (1 \times 2^2) + (0 \times 2^1) + (1 \times 2^0) \\ &= 16 + 8 + 4 + 0 + 1 \\ &= 29 \end{aligned}$$

An alternative presentation of the working is shown in Figure 17.7.

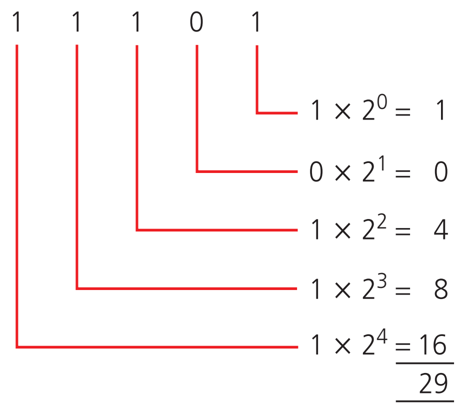


Figure 17.7 Converting binary 11101 into a decimal number

4 Write the decimal number 62 as a binary number using eight binary digits (bits).

You are told to use eight bits, so the highest power of 2 is 2^7 (128).

Draw up a table of powers of 2 up to 2^7. Then subtract the largest possible power of 2 from 62, and keep subtracting the next largest possible power from the remainder, marking 1s in each column in the table where this is possible and 0s where it is not (see Figure 17.8).

$$\begin{array}{r} 62 \\ -32 \\ \hline 30 \\ -16 \\ \hline 14 \\ -8 \\ \hline 6 \\ -4 \\ \hline 2 \\ -2 \\ \hline 0 \end{array}$$

128	64	32	16	8	4	2	1
0	0	1	1	1	1	1	0

Figure 17.8 Converting decimal 62 (base 10) to binary (base 2)

This gives you the number in terms of its coefficients and powers of 2:

$$62 = (0 \times 2^7) + (0 \times 2^6) + (1 \times 2^5) + (1 \times 2^4) + (1 \times 2^3) + (1 \times 2^2) + (1 \times 2^1) + (0 \times 2^0)$$
$$(0 \times 128) + (0 \times 64) + (1 \times 32) + (1 \times 16) + (1 \times 8) + (1 \times 4) + (1 \times 2) + (0 \times 1)$$
$$= 00111110$$

5 The letter A is typed every second into a text document. The letter A is stored as an eight-bit binary digit (ASCII code). How long would it take in days (to the nearest integer) to fill a 2 GB thumb drive?

$2\,\text{GB} = (2\,\text{GB} \times 1024\,\text{MB} \times 1024\,\text{kB} \times 1024\text{ bytes} \times 8)\text{ bits} = 1.7179869184 \times 10^{10}\text{ bits}$

One 'A' uses eight bits, so (171 798 691 840 / 8) = 2 147 483 648 bytes.

This will take $\dfrac{2147483648}{(60 \times 60 \times 24)} = 24\,855$ days

4 Convert the following decimal numbers into eight-bit binary numbers:
a 8 **b** 14 **c** 17 **d** 68 **e** 125

5 Convert the following six-bit binary numbers to base 10 (decimal) numbers:
a 000110 **b** 100101 **c** 110010 **d** 111111

6 The hexadecimal system (base 16) is particularly important in computer programming since four binary digits can be easily expressed using a single hexadecimal digit. Describe the hexadecimal system.

7 Table 17.4 shows part of the ASCII code. Write the word CAB in ASCII code.

Table 17.4

Letter	ASCII code
A	0100 0001
B	0100 0010
C	0100 0011
D	0100 0100
E	0100 0101

8 Find out how multiplication and division are performed in binary (base 2).

9 Find out about the history and development of counting systems and numerals. Ensure that your research includes the work of the Babylonians, Egyptians, Arabs, Indians and the Maya.

10 Construct a spreadsheet using Excel that will interconvert decimal and binary numbers. Use the BIN2DEC and DEC2BIN functions. You will need to make sure the Analysis Toolpak is installed.

11 Find out about Grid computing and the processing and storage of digital data at CERN in Geneva, where high-energy particle physics experiments are performed.

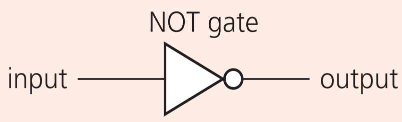
12 Make a list of digital devices that you use regularly and sort the devices into work/learning/study and leisure/fun/entertainment. Suggest what new digital devices may be developed in the future.

Additional Perspectives

Logic gates

Logic gates are switching circuits found in computers and other electronic devices. They 'open' and give a 'high' voltage (a signal, 6 V, represented by binary 1) depending on the combination of voltages at their inputs. There are three basic types (NOT, OR and AND) and the behaviour of each is described by a *truth table* showing what the output is for all possible inputs. 'High' (e.g. 6 V) and 'low' (e.g. 0 V) outputs and inputs are represented by a binary 1 and 0 respectively, and are referred to as logic levels 1 and 0.

The NOT gate is the simplest logic gate, with only one input and one output. It produces a 'high' output if the input is 'low'. Whatever the input signal, the gate *inverts* the signal. The symbol and truth table are given in Figure 17.9.



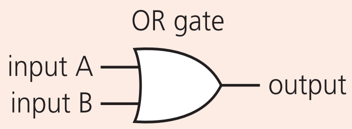
Input	Output
0	1
1	0

Figure 17.9 NOT logic gate symbol and truth table

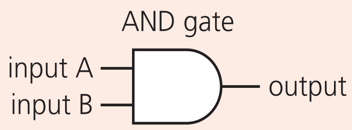
The AND and OR logic gates have two inputs and one output. They behave according to the following statements:

OR output is 1 if input A OR input B OR both are 1
AND output is 1 if input A AND input B are 1

The truth tables and diagrams used for the OR gate as well as the AND gate are shown in Figure 17.10.



Input A	Input B	Output
0	0	0
0	1	1
1	0	1
1	1	1



Input A	Input B	Output
0	0	0
0	1	0
1	0	0
1	1	1

Figure 17.10 Symbols and truth tables for an OR and an AND gate

Logic gates can be used as processors in electronic control systems. The block or system diagram that might be used by a jeweller to protect an expensive gold watch is shown in Figure 17.11. The gold watch is placed on a push switch (pressure sensor) which sends a binary 1 to the NOT gate *unless* the gold watch is lifted, in which case a binary 0 is sent. If a binary 0 is sent, then the output from the NOT gate is a binary 1, which rings the buzzer.

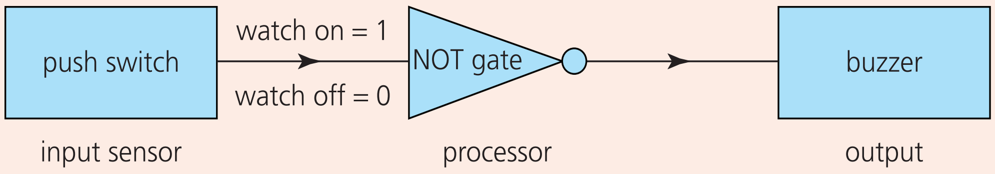


Figure 17.11 A simple alarm system

Question

1 Deduce the truth table for an AND gate whose output is connected to a NOT gate.

Storage of information

C.1.2 Describe different means of storage of information in both analogue and digital forms.

Analogue and digital data must be stored in a permanent 'machine-friendly' form. This then allows the data to be accessed and used many times. Data storage can be physical – the shape of a vinyl record groove or the tracks of pits on a CD or DVD – or it can be magnetic, as in the magnetic patterns on a cassette tape or computer disk.

LPs

Vinyl records, gramophone records or LPs (long players) and cassettes (Figure 17.12) were popular formats for storing recorded music until compact discs (CDs) were introduced in the 1980s. The analogue musical information in an LP is stored in a groove cut into the plastic (vinyl) record by a sharp stylus (needle). The shape of the wiggles in the groove is an analogue representation of the musical signal. Loud sounds are stored as large wiggles, and quiet sounds as small wiggles. The closer together the wiggles are, the higher the frequency of a sound. The average LP contains approximately 470 metres of groove on each side.

The LP is played back by placing another stylus into the groove and then rotating the record at constant speed on a turn table (Figure 17.13). The groove in the LP spirals inwards so the needle moves through the groove faster when it is playing music on the edge of the LP. This means that the groove is more squashed in the centre of the LP than at the outer edge to compensate for the decrease in speed as the record continues to play.

Figure 17.12 Audio cassette tape and LP (vinyl record)

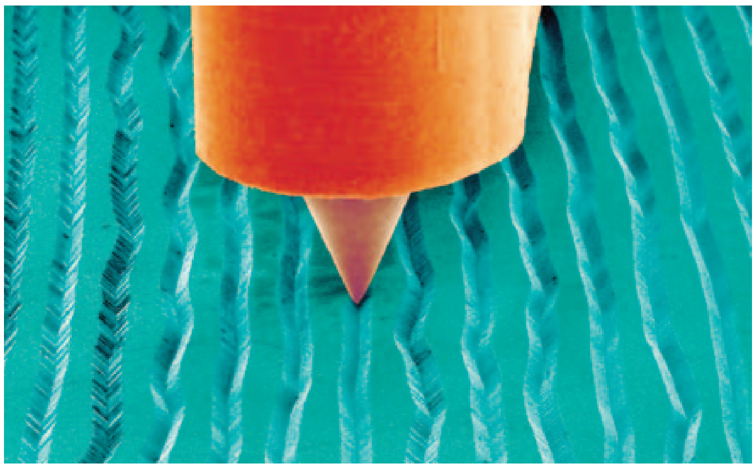


Figure 17.13 Retrieval of analogue information from an LP record

The stylus vibrates in the groove and produces an electrical signal (due to the presence of a piezoelectric crystal), which is played back through an amplifier and speakers. Every time the LP is played, the stylus slightly damages and changes the shape of the groove. In addition, dust and smoke particles settle in the groove. The distorted and dust-filled groove causes unintended vibrations of the stylus and so unwanted electronic 'noise' is added to the signal. The noise is heard as a background hissing sound when the record is played. A relatively large LP may also warp (bend), which distorts the recording.

Cassette tapes

Music (or video) can be recorded on a long thin plastic tape coated with a fine powder of magnetic material (oxides of iron or chromium) which becomes magnetized during recording. A chain of tiny permanent magnets is produced on the tape in a pattern which represents the original sound (or picture). Figure 17.14 shows the simple magnet patterns for single 'high' and 'low' frequencies.

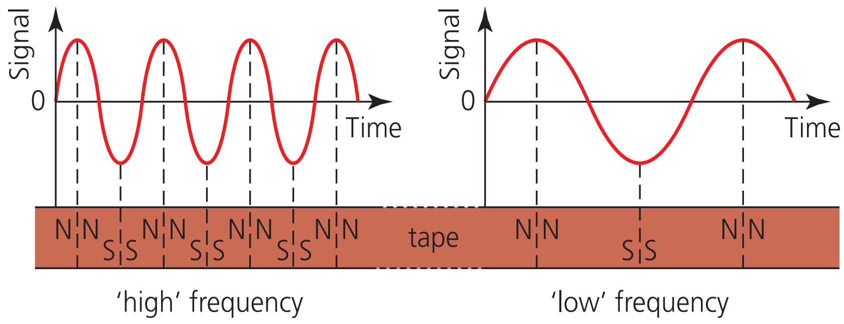


Figure 17.14 Magnetization of audio or video tape

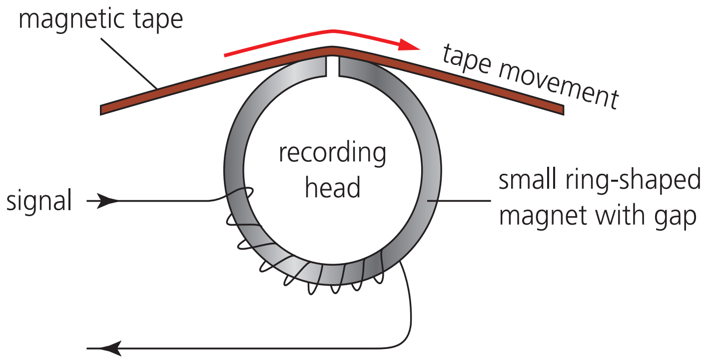


Figure 17.15 The recording head in a cassette tape recorder

The magnetizing is done by the recording head (Figure 17.15), which consists of a small electromagnet. The information to be recorded is sent as a small, changing magnetic field which magnetizes the tape particles in the same pattern as the current in the coil.

To retrieve the information from the tape, the tape is pulled past the electromagnet (which also acts as the playback head). The pattern of magnetization on the tape produces a varying magnetic field in the core of the head, which then induces a varying electric current in the coil. This current is amplified, so the signal can be output through a loudspeaker. Cassette tapes are generally analogue in nature, but digital cassette systems have also been developed.

13 Find out about the use of tape drives to back up computer systems or large data files.

Floppy disks

Figure 17.16 A 3½ inch floppy disk

The floppy disk (Figure 17.16), like the cassette, is a magnetic form of data storage. The first floppy disks, developed in the 1960s, were 8 inches in diameter. By the end of the 20th century they were 3½ inches in diameter and could store 1.44 megabytes of digital data. Floppy disks are now virtually obsolete because of their relatively low capacity to store data, but have some use with old computer systems.

The data on a floppy disk is stored in patterns of magnetic particles arranged in concentric rings (called tracks). The floppy disk reader is able to access data on any track without having to search sequentially through the other tracks as on an analogue cassette tape. The floppy disk (and hard disk, see below) are examples of direct access storage devices. In a direct access storage device bits of data are stored at precise locations, enabling the computer to retrieve information directly without having to scan a series of records.

14 Find out how 'flash memory' works. Describe the forms it takes and the types of devices that use it.

Hard disks

Large amounts of data can be stored in magnetic form on a **hard disk**, for example in computers and in digital video cameras. The hard disk of a personal computer contains a stack of platters (Figure 17.17) which spin at high speed. The platters are rigid (stiff) and are coated with iron oxide particles. A head, which is an electromagnet that has reversible polarity, is used to put digital information on the disk surface. In one polarity the head aligns the magnetism in a tiny area of the disk in one direction so that information is stored as a binary one ('on'), and in the opposite polarity the information acts as a binary zero ('off').

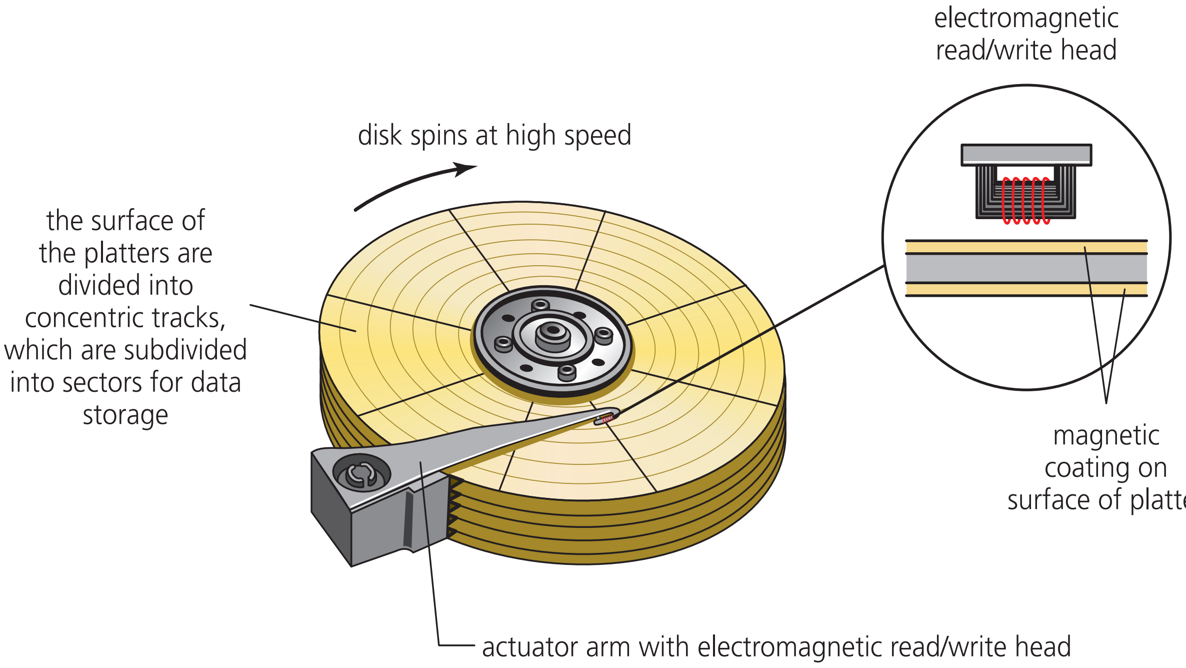


Figure 17.17 The spinning platters in the hard disk of a personal computer

To read the data, the platter or disk spins past the head, which induces a current in the electromagnet. If dust or smoke particles were to get into the drive, the disk surface would be irreversibly damaged. For this reason, the platter and head assembly is sealed during manufacture.

Compact discs

An optical storage medium uses light, usually in the form of a laser, to read and/or write data. The compact disc (CD) is an optical storage medium. CDs have been one of the most popular formats for storing digital sound recordings (music), large computer files and movies. CDs can store about 0.75 GB (gigabytes) of digital data and so are very convenient for transferring and storing digital information.

The data on a CD is stored in a track of microscopic 'pits' and 'lands', which are moulded into a thin layer of transparent plastic (Figure 17.18). These indentations are then covered with a thin layer of reflective aluminium. The pits and lands are arranged one after another in a spiral track, starting from the centre of the disc.

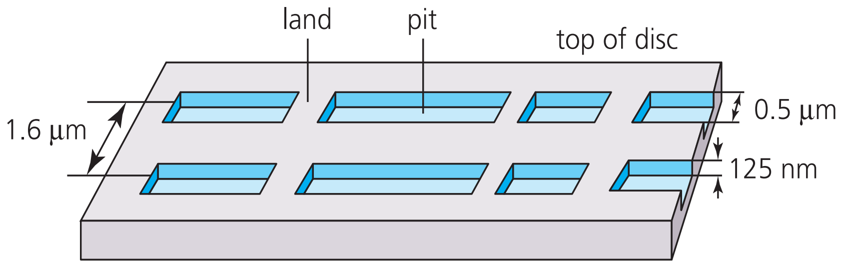


Figure 17.18 Diagram showing the three dimensional arrangement of 'pits' and 'lands' in a track on a CD (not to scale). The length of the pits ranges from 830 to 3560 nm

The disc rotates at about 500 revolutions per second, and a low power laser beam 'reads' the data off the track with its series of pits and lands. (The laser reads from the bottom of the disc, so the pits are in fact raised compared to the lands.) The pattern of pits forms a coded signal of 16 binary numbers, each representing one feature of the sound wave (if the CD stores music).

The laser starts reading the data from the centre and moves outward along a radial line as the disc rotates rapidly next to it. There is no mechanical contact between the laser and the CD disc, so it can be replayed repeatedly without introducing noise.

It is important that the laser reads the pits at a constant rate, otherwise it might misread the digital information in the pits and lands. If the CD is rotated at a constant rate, the speed at which the pits and lands pass the head would increase as the laser moved to the outer edge, so the CD rotation is slowed down (by a motor) as the laser reader moves outward.

15 Describe the similarities and differences between LPs, cassette tapes and CDs.

16 Find out how music from an LP can be transferred to a computer.

Additional Perspectives

Error correction in CD players

The music on an audio CD does not play in sequence along the track. The binary digits corresponding to a small piece of a music track are scattered around the disc. This means that if a portion of the CD is damaged, only a very small part of any section of music is lost. This is known as interleaving and it means that the laser reading head has to be accurately guided or steered around the disc; it is moving laterally as the disc is playing. This means that the disc speed is constantly changing, which requires high-quality motors and excellent control electronics. The CD contains digital information that allows the laser reading head to keep track of its location.

Question

1 Find out about the role of error correcting code (ECC) on an audio CD.

DVDs

DVD (digital video disc or digital versatile disc) is an optical disc storage format similar to the CD. Its main uses are video and data storage. DVDs are of the same dimensions (just under 12 cm) as compact discs, but are capable of storing almost seven times as much data.

DVD uses a 650 nm wavelength laser diode light as opposed to 780 nm for CD. This allows a smaller pit to be etched on the media surface compared with CDs (0.74 μm for DVD versus 1.6 μm for CD), which gives the DVD a greater storage capacity. The track length of a DVD is about twice as long as the track length of a CD.

DVD+R DL is a DVD format that contains double the information of a DVD by having two layers of pits. The top layer is coated with a semi-reflective coating, enabling light to also pass through to read the bottom layer. This DVD format is able to store 17 GB of data.

TOK Link: Digital representation

René François Ghislain Magritte (1898–1967) was a Belgian surrealist artist. He became well known for a number of thought-provoking images including *La trahison des images*, which shows a pipe and the words *Ceci n'est pas une pipe*, which translates into 'This is not a pipe'. This seems a contradiction, but is actually true: the painting is not a pipe, it is an image of a pipe. This reminds us that the digital information stored in digital storage devices is a *representation* of words, music and images. The image is reduced to binary 0s and 1s and the 'essence' of the image is lost. It is not the 'real' pipe.

Question

1 What important reminders does this painting give us about how we represent the world in physics?

Recovery of information on a CD

C.1.3 **Explain** how interference of light is used to recover information stored on a CD.
C.1.4 **Calculate** an appropriate depth for a pit from the wavelength of the laser light.

The laser beam is focused on the CD surface, and is reflected back onto a detector (photodiode). How it is reflected depends on whether it falls on a pit or a land. If the laser beam is entirely incident on a pit (or on a land) then all the waves in the reflected beam are in phase with each other. Constructive interference takes place, and a strong signal (a binary one) is detected by the photodiode.

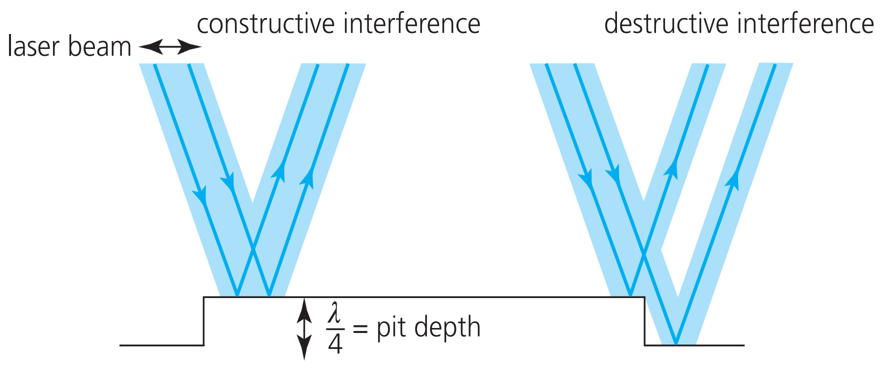


Figure 17.19 The reflection of the laser beam from a pit and from a land. Here the light is shown falling at an angle, but in a CD player it is *almost* normal to the disc surface

If part of the light beam is incident on a pit and part on a land, then there is a path difference between the two parts of the laser beam. The pit depth is such that, for a laser beam of wavelength λ, the path difference is $\lambda/2$, which means that destructive interference occurs and there is no signal produced (a binary zero) at the photodiode. Since the path difference is twice the pit depth, to obtain a path difference of $\lambda/2$:

$$\text{pit depth} = \frac{\lambda}{4}$$

As the laser beam reflects off the rotating spiral track, the signal received by the photodiode changes as the beam travels from pit to land to pit. This produces digital signals of 1s and 0s varying according the whether the interference is constructive or destructive, and also according to the lengths of the pits and lands. If the CD is an audio CD storing music, then a digital-to-analogue converter (DAC) converts this digital signal back into the corresponding analogue signal.

Calculating the depth for a pit from the wavelength of the laser light

We have seen that the depth of the pit on a CD track must be a quarter wavelength of the laser light used to read the CD so that destructive interference can occur. The lasers used for CDs, DVDs and Blu-ray all have different wavelengths, so the discs themselves must be made differently.

Worked example

6 Laser light of frequency 3.80×10^{14} Hz is used in the laser of a CD-ROM reader. Calculate an appropriate depth of a pit on a CD.

$$\lambda = \frac{c}{f} = \frac{3.00 \times 10^{8}}{3.80 \times 10^{14}} = 7.89 \times 10^{-7}\,\text{m}$$

$$\text{Depth of pit} = \frac{\lambda}{4} = \frac{7.89 \times 10^{-7}}{4} = 1.97 \times 10^{-7}\,\text{m}$$

17 If the laser light had a wavelength of 620 nm in plastic, what depth of pit would the CD have? Explain your answer.

18 If the CD rotates at a rate of 600 revolutions per minute (rpm) when the laser pickup is 2 cm from the centre, how fast should it rotate when the pickup is 6 cm from the centre?
(Recall from the discussion about circular motion in Chapter 4 that the speed of a rotating body is given by: $v = \omega r$ where $\omega = 2\pi f$.)

19 Find about holographic discs and describe how digital data is stored in them.

■ Additional Perspectives

Conversion of analogue to digital signals

Analogue signals can be easily and quickly converted into digital signals. In an analogue-to-digital converter, the analogue voltage is *sampled* at regular intervals of time. The sampling frequency (rate) is the number of samples in unit time. The value of the sample voltage measured at each sampling time is converted into a binary (digital) number that represents the voltage value.

For example, if a four-bit number is being used, then the number representing a signal that is sampled as 5.0 V would be 0101. When sampling, the number representing the sample would be the whole number (integer) below the actual value of the sampled voltage. If the signal were to be sampled as 11.4 V, then the four-bit number would be 1011. A sampled signal of 11.6 V would also be 1011.

Figure 17.20a shows an analogue signal that is to be sampled at a sampling rate of 10 kHz. A four-bit system is used for the binary (digital) numbers generated. The sample voltages are shown in Figure 17.20b. These sample voltages are converted into a digital signal, shown in Figure 17.20c, by the analogue-to-digital converter (ADC). After this digital signal has been transmitted, it is converted back into an analogue signal using a digital-to-analogue converter (DAC). Complex programs (algorithms) are required for converting analogue signals into digital signals and vice versa.

Figure 17.20d shows the analogue signal that has been recovered. The recovered signal has large 'steps'. It is known as a **pulse amplitude modulated** (PAM) signal. The size of these steps can be reduced and hence the accuracy of the reproduction of the initial analogue signal can be improved by using more voltage levels (greater quantization) and sampling at a higher frequency.

The number of bits in each binary number limits the number of voltage levels (**quantization levels**). In this conversion there are four bits and 16 (2^4) levels. An eight-bit number would give 256 (2^8) levels and a 16-bit number would give 65 536 (2^{16}) levels.

For good quality reproduction of music, the higher audible frequencies must be present; that is, frequencies up to about 20 kHz. For compact discs (CDs) the sampling frequency is 44.1 kHz. This quality of reproduction is not required for speech, and it would be expensive. For walkie-talkies, intercoms and the telephone system the sampling frequency is 8 kHz and the highest frequency to be transmitted is limited to 3.4 kHz.

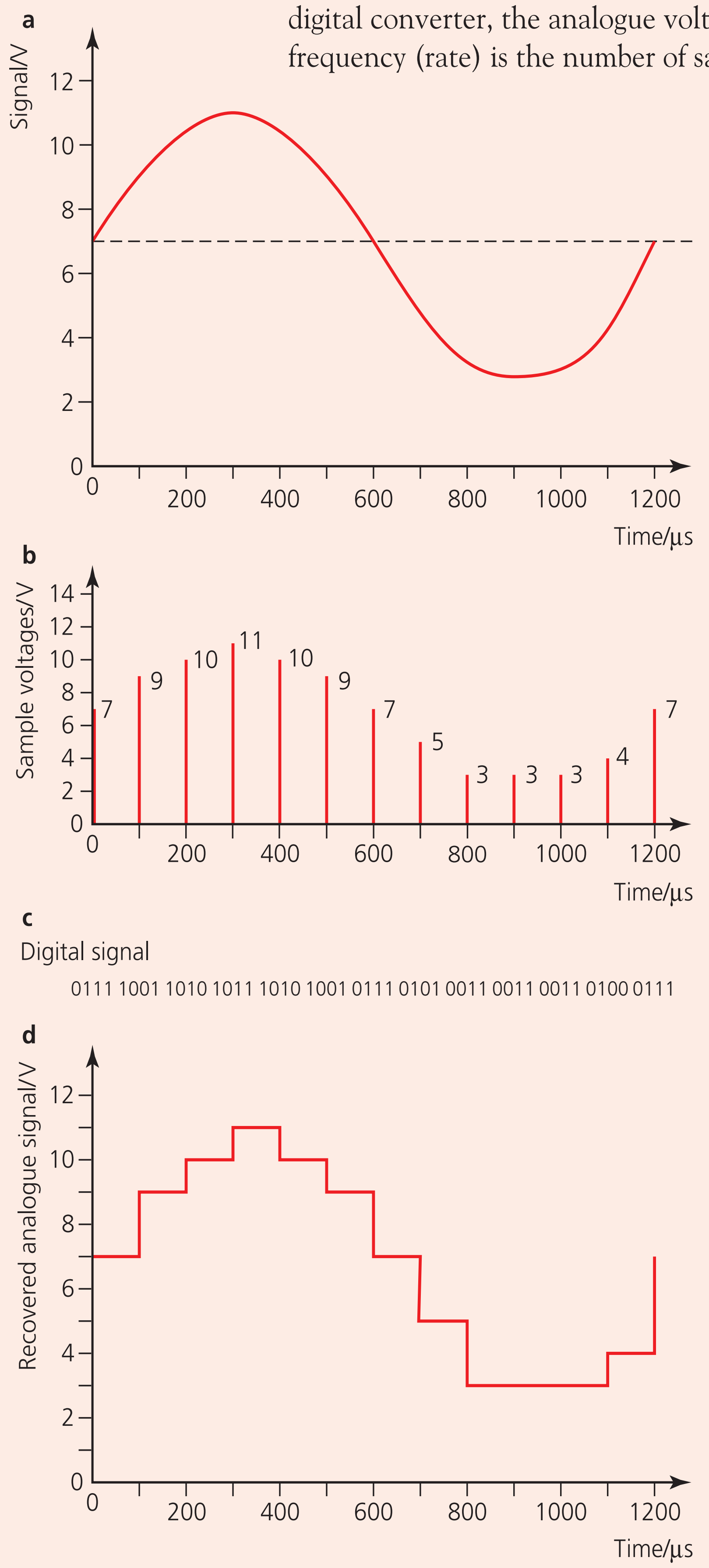


Figure 17.20 Analogue-to-digital and digital-to-analogue conversion

Questions

1 a Figure 17.21 shows a graph of the varying potential difference from a microphone. Convert this signal to a PAM signal by measuring the potential difference at a sampling rate of 200 Hz. Round off all of the potential difference values to the nearest tenth of a volt. Draw a graph showing the digital signal.

b Repeat with a sampling rate of 400 Hz.

2 To measure the variation of a 100 Hz alternating current (ac) signal using a digital device, what is the minimum sampling rate you should use?

3 The following string of binary 1s and 0s is three-bit binary data sampled at 2 Hz:

011 100 101 001 011 111

a Present this data in the form of a table displaying time and a number to represent the potential difference.

b Use this data to draw the digital signal.

4 Find out how digital data can be compressed and encrypted.

5 Find out about digital television (DTV). Outline the advantages of DTV over traditional analogue TV.

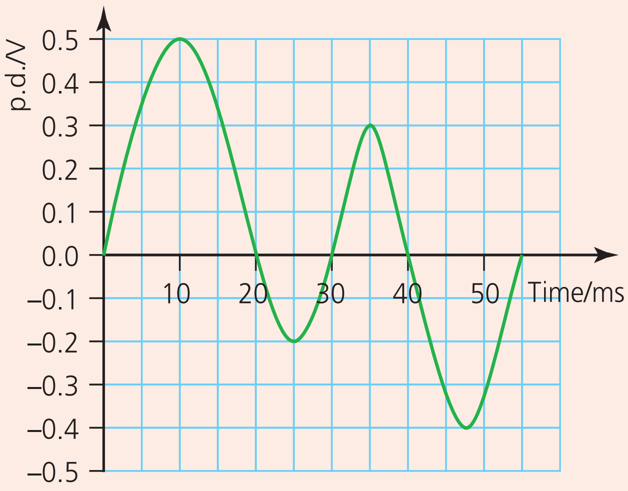


Figure 17.21 A graph of the varying potential difference from a microphone

Optical data storage capacity

C.1.5 Solve problems on CDs and DVDs related to data storage capacity.

Problems about optical data storage capacity can be about how much data is stored, how quickly it can be retrieved, the play time of a disc, or how long the tracks are that store the data. You will need to use what you have learned about the relationship between bits and bytes. You also need to know that the sampling rate is the rate at which an analogue signal is sampled (its value recorded). Thus 16-bit sampling at a rate of 44 kHz means that each second $44\,000 \times 16$ bits of data are recorded.

Worked examples

7 Information is imprinted (during the manufacturing process) on a music CD at a rate of 44 100 words per second. The digitized information consists of 32-bit words (two stereo channels of 16-bit samples each).

A mini-CD single contains 24 minutes of music. Calculate the capacity of a mini-CD single to the nearest megabyte.

Number of bits imprinted on the CD $= 44\,100 \times 32 \times 24 \times 60 = 2.03 \times 10^9$ bits

Since one byte equals eight bits this corresponds to:

$$\frac{2.03 \times 10^9}{8} = 2.54 \times 10^8 \text{ bytes} \approx 242\text{ MB (since 1 MB} = 1\,048\,576 \text{ bytes)}$$

8 A track on a music CD moved from a radius of 21 mm to 54 mm with an average radius of 40 mm. The distance between spirals is 1.6 μm.

a Estimate the length of the track (in centimetres).

b The average scanning speed by the laser is 1.4 m s^{-1}. Estimate the play length of the audio CD in minutes.

c The CD can store 737 MB of information. What is the average length of track in micrometres per bit of information?

a Number of turns $= \frac{(54-21)\times 10^{-3}}{1.6\times 10^{-6}} = 2.06\times 10^4$

Track length $= 2.06\times 10^4 \times 2\times \pi \times 4.0 = 5.2\times 10^5$ cm

b CD playing time $= \frac{5.2\times 10^3}{1.4} = 3.7\times 10^3$ s $=$ 62 minutes

c Average length per bit $= \frac{5.2\times 10^3}{7.37\times 10^2 \times 2^{20}\times 8} = 8.4\times 10^{-7}$ m $=$ 84 µm

9 Estimate the playing time of a 737 MB CD storing stereo music using 16-bit sampling. The maximum human audio frequency is 20 kHz, hence the sampling rate is 40 kHz.

Number of bits every second for each channel (of the stereo) $= 40\,000\times 16 = 6.4\times 10^5$ bits
Total number of bits per second for stereo $= 2\times 6.4\times 10^5$ bits $= 1.28\times 10^6$ bits
Total storage capacity of CD $= 7.37\times 10^2\times 2^{20}$ bytes $= 6.18\times 10^9$ bits

Maximum play time for CD $= \frac{6.18\times 10^9}{1.28\times 10^6} = 4830$ s $=$ 81 minutes

20 The track on a 12 cm audio CD is 5.7 km long and is made of a series of pits and lands that individually are a minimum of 0.8 µm long.
- **a** How many pits are there on a CD track (remembering that each pit is followed by a land)?
- **b** Each short pit has two edges so represents two bits of data. Calculate the number bits present on a standard audio CD.
- **c** How many megabytes (MB) of data are there on a standard 12 cm audio CD?

21 A standard 12 cm audio CD can store 74 minutes of stereo music.
- **a** If 16 bits are recorded at 44 100 samples per second, how many bits are recorded in 74 minutes?
- **b** The music to be stored is recorded as 'surround sound' so there are six channels. Calculate the number of bits.
- **c** 'Surround sound' music is supplied on a Super Audio CD (SACD), which has a much greater storage capacity than a standard audio CD. Calculate the number of megabytes stored in the 'surround sound' on the SACD.

Storage of information

C.1.6 **Discuss** the advantage of the storage of information in digital rather than analogue form.

■ Additional Perspectives

Transmission of information

When any electrical signal is transmitted over a long distance it will pick up noise. **Noise** is any unwanted random signal disturbance that adds to the signal being transmitted. It often occurs when two wires are next to each other. In addition, the power of the signal becomes progressively reduced with distance. We say that the signal has become **attenuated**. For long distance transmission, the signal has to be amplified at regular intervals. But the problem is that when an analogue signal is amplified, the noise is also amplified. The signal becomes distorted or 'noisy'. This effect is shown in Figure 17.22.

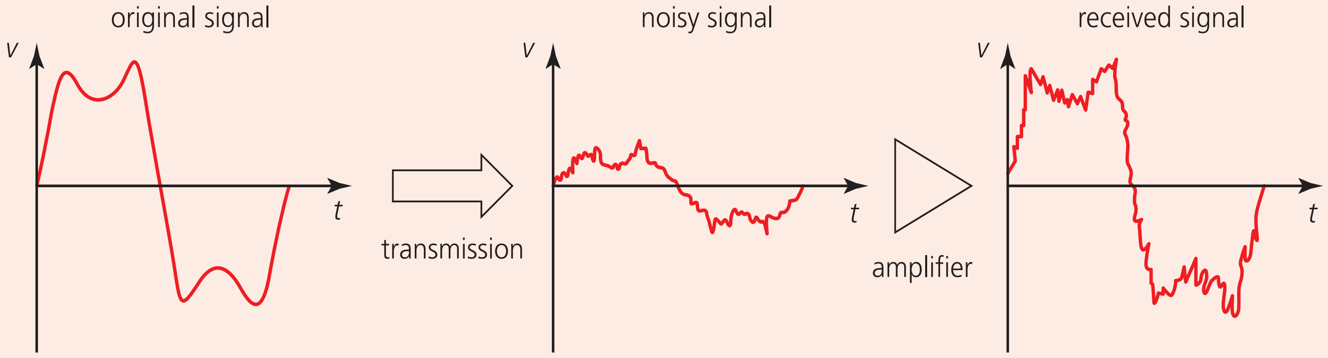


Figure 17.22 Amplification of a 'noisy' analogue signal

A digital signal still has noise and will be attenuated. However, when amplified the noisy 1s and 0s can be reshaped or regenerated to return the signal to its original form. Such amplifiers are known as regenerator amplifiers. They 'filter out' any noise and restore the digital signal (Figure 17.23). In contrast to an analogue signal, a digital signal can be transmitted over a long distance with regular regenerations without the signal becoming degraded.

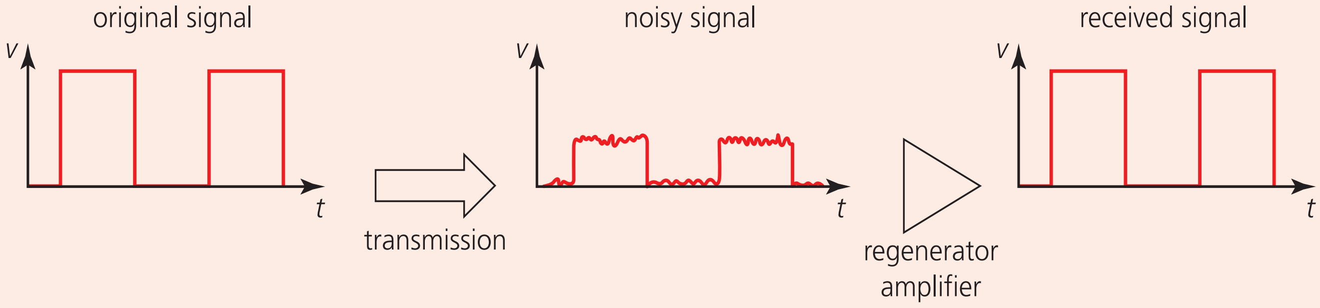


Figure 17.23 Amplification of a 'noisy' digital signal

Modern digital electronic circuits are more reliable, smaller and cheaper to produce than analogue circuits. An added advantage of digital systems is that extra information, or data known as a checksum, can be added to the transmission. These extra data are a simple code for the receiving system so that the transmitted signal may be checked and corrected before the signal is finally reproduced.

Digital data can be quickly and cheaply manipulated, compressed (to reduce storage space), processed, edited or made secure via a process of encryption. Analogue data is more difficult to process and encrypt.

Modern miniaturization techniques based on microprocessors mean that large amounts of digital data can be stored on a physically small device. Although stored analogue data can be compact, for example analogue tapes, many analogue storage systems are relatively large. Compare, for example, an MP3-player with an LP. A 120 GB MP3 player can store the music of about 1500 LPs.

The retrieval speed for digital data is usually much more rapid than the retrieval speed for comparable analogue data. The process of retrieving analogue data often affects the quality of future data retrieval. For example, the quality of music on magnetic tape and LPs decreases with the number of times the data is retrieved and the music played. However, optical techniques, as used in CD-ROM readers, can ensure that the data is identical each time it is accessed and there is no degradation in its quality.

Data storage

C.1.7 **Discuss** the implications for society of ever-increasing capability of data storage.

Table 17.5 on page 647 outlines *some* of the implications for the ever-increasing capability of data storage. It is claimed computing power (the number of operations a computer can perform per second) doubles about every 18 months and data storage costs are rapidly declining. Networking advances and the Internet make copying data from one location to another and accessing personal data from remote locations much easier. Some of the moral, ethical, social and environmental implications of this global issue are considered.

Table 17.5 Implications for increasing data storage

Moral/ethical	• Issues concerning the privacy and anonymity of personal digital data. • Issues related to access and ownership of personal digital data.
Cultural	• Some people are concerned about cultural integrity of their regional or lingual groups due to the widespread use of English on the Internet.
Economic	• Many of the recent financial scandals involved the secret modification of accounting and financial data. • The integrity of currencies may be undermined by digital or e-cash. • Increased access to training and education via computer-assisted teaching and learning. • Economic decision-making by consumers will be faster and more accurate. This may lead to more stable markets for goods and services and perhaps a tendency towards low inflation. There is also a counter argument, that the 'herd instinct' seen in stock markets, based on instant digital information, results in a high degree of instability in the system.
Social	• The emergence of new forms of communities due to social networking sites (such as Facebook and Twitter), which have both positive and negative features.
Environmental	• Electronic data storage could replace traditional techniques, thus saving in resources, e.g. e-books do not use wood pulp. However, data centres consume electricity (and hence in most cases contribute to global warming and climate change). • The resources needed for the maintenance of electronic data will be in higher demand, e.g. fossil fuels and silicon.

22 Find about the new emerging science of bioinformatics, a combination of biology and information technology.

23 Find about any laws in your country, such as the United Kingdom's Data Protection Act 1998, which govern the protection of personal data, including digital data.

24 Research the use of 'cookies' and 'web bugs' by websites on the Internet.

25 Give one additional example of your own for each of the five headings in Table 17.5.

C2 Data capture; digital imaging using charge-coupled devices (CCDs)

Capacitance

C.2.1 Define *capacitance.*

Capacitors (Figure 17.24) are devices that can store charge (electrons). They store energy in an electric field. Capacitors are common components of electrical circuits and perform a variety of functions. Camera flash units use capacitors to store energy and capacitors are an essential part of radio tuners. Digital cameras contain a sensor known as a charge-coupled device (CCD) which is made up of millions of tiny pixels, which behave like tiny capacitors.

Two parallel metal plates separated by a small gap filled with air form a simple capacitor, as illustrated in Figure 17.25. The metal plates are connected to a battery (a source of a potential difference). When the switch is closed, a current will flow for a very short time. The capacitor becomes charged. Negative charge (electrons) will accumulate on the bottom plate, leaving behind an equal amount of positive charge on the top plate.

The amount of charge stored on the surface of the plates depends on the design of the capacitor and the p.d. across them. In fact, the charge, q, is directly proportional to the

Figure 17.24 A capacitor (with a capacitance of 500 μF)

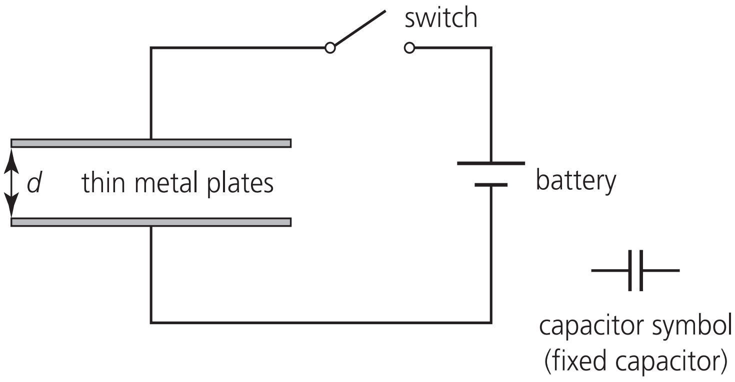


Figure 17.25 A simple circuit with a capacitor, and the symbol for a capacitor (with fixed capacitance)

potential difference between the metal plates (i.e. $q \propto V$). The constant in this relationship is called the **capacitance** C of the plates.

$$q = CV$$

Capacitance is the charge per unit potential difference that can accumulate on a conductor. It is defined as the ratio of charge stored to the applied potential difference:

$$C = \frac{q}{V}$$

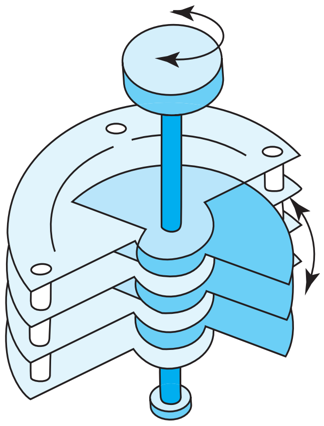


Figure 17.26 A variable capacitor

The SI unit of capacitance is the farad (F), with one **farad** (1 F) being a capacitance of one coulomb per volt ($1\,\text{C}\,\text{V}^{-1}$); one farad (1 F) is a relatively large capacitance. The unit is named after Michael Faraday.

Smaller multiple units of the farad are often used: $1\,\mu\text{F} = 10^{-6}\,\text{F}$ (a microfarad), $1\,\text{nF} = 10^{-9}\,\text{F}$ (a nanofarad) and $1\,\text{pF} = 10^{-12}\,\text{F}$ (a picofarad). Variable capacitors (Figure 17.26) can be constructed whose capacitance can be varied by changing the overlap of the plates. A variable capacitor is a critical part of a radio tuning system that selects the frequency of a radio station, thus it is often referred to as a tuning capacitor. Resonance occurs when the discharge frequency of the capacitor equals the radio frequency.

Worked example

10 The capacitance of two parallel metal plates is 9.50 pF. Calculate the charge on one of the plates when a potential difference of 18.0 V is established between the plates.

$$q = CV = 9.50 \times 10^{-12} \times 18.0 = 1.71 \times 10^{-10}\,\text{C} = 171\,\text{pC}$$

Additional Perspectives

Capacitors

The voltage rises as we charge up a capacitor, and falls as the capacitor discharges. The current falls from a high value as the capacitor charges up, and falls as it discharges. During the discharge of a capacitor the current can cause the heating of a resistor or even drive a small electric motor for a short period of time. So the charged capacitor is acting as an energy store. A graph of potential difference (V) across the capacitor against charge (q) is a straight line, with the gradient being a constant for the capacitor, and equal to 1/C. For a capacitor the energy stored is equal to the area under the graph of potential difference versus charge (Figure 17.27).

Thus:

$$\text{energy stored} = \text{area under } V\text{–}q \text{ graph}$$
$$= \frac{1}{2} \times \text{final voltage} \times \text{final charge}$$
$$= \frac{1}{2}Vq$$

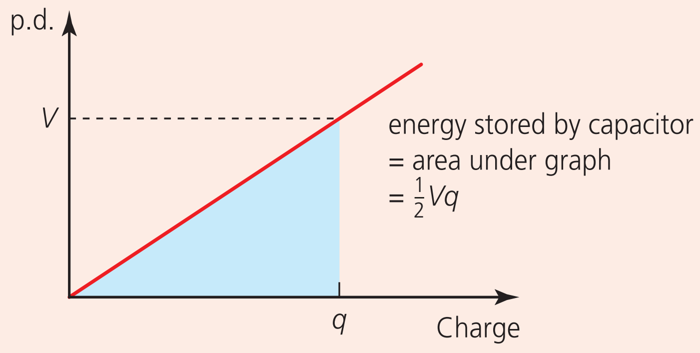


Figure 17.27 Energy stored in a capacitor

Since $C = \frac{q}{V}$,

$$\text{energy stored} = \frac{1}{2}\frac{q^2}{C} \quad \text{or} \quad \frac{1}{2}CV^2$$

When a capacitor is discharged the charge decays exponentially (Figure 17.28). A graph of charge q against time t is an exponential decay curve (compare with radioactive decay – Chapter 7 and Option B). The rate of decay, or steepness of the curve, depends on the values of R (resistance) and C (capacitance) and on the charge on the capacitor at that moment. Such a proportional relationship between the rate of change of value and the value itself always leads to an exponential relationship.

The exponential decay curve for a capacitor is described by the following equation:

$$q = q_0 e^{-t/RC}$$

where RC is known as the *time constant*.

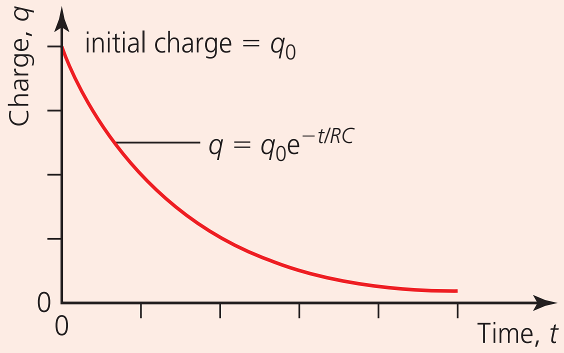


Figure 17.28 A decay curve for a discharging capacitor

Question 1

1 Use a spreadsheet to model capacitor discharge: use $R = 47\,k\Omega$ and $C = 22\,\mu F$. In Excel the EXP function returns e raised to the nth power, where e = 2.718. The syntax for the EXP function is: EXP (number). Use the generated numbers and plot a graph of charge versus time. Plot a graph of $\ln q$ against time and show that the slope of the graph will be $-1/RC$.

26 What will be the potential difference between the plates of a 20 μF capacitor if there is a charge of 5 μC on the plates?

27 The capacitance of a device known as a charge-coupled device (CCD) is 100 μF. What charge (in μC) will cause a potential difference of 1.2 V across it?

28 If there are 500 electrons stored on a 100 nF photodiode, what will be the potential difference across it?

29 Find out about the use of capacitors in the tuning of a simple radio with a tuning circuit.

30 A capacitor is charged with a 20 V battery. Calculate the capacitance of the capacitor if the maximum charge on the plates is ± 500 μC.

31 A graph of the variation of charge with potential difference (p.d. or voltage) across a capacitor is shown in Figure 17.29. What is the capacitance of the capacitor?

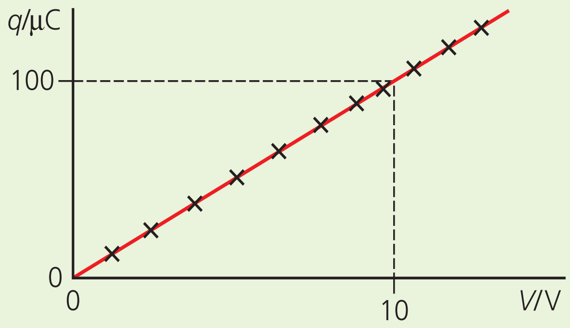


Figure 17.29

Charge-coupled devices

A charge-coupled device (CCD) is a highly sensitive light detector originally developed in astronomy for digital imaging. These devices doubled the ability of telescopes to detect light and allowed astronomers to obtain colour images of very faint stars. They are now to be found in digital cameras, web cams, digital video recorders and scanners.

Image capture in a charge-coupled device

C.2.2 **Describe** the structure of a charge-coupled device (CCD).
C.2.3 **Explain** how incident light causes charge to build up within a pixel.

A charge-coupled device (CCD) is a small silicon chip with a surface covered with a very large number of light sensitive elements called **pixels** (picture elements). The pixels are defined on the surface of the silicon wafer by insulating channels in one direction and by rows of transparent electrodes deposited on the surface in the other direction (Figure 17.30).

When exposed to light, for example by opening the shutter on a digital camera, each pixel within the CCD releases electrons as result of a process similar to the photoelectric effect (Option B). The greater the intensity of the light on a pixel, the higher the number of photons that are incident on the pixel per second, and so the higher the number of electrons released.

The voltages that develop on the electrodes across the array are set so that each individual pixel in the array behaves as a small capacitor, trapping the electrons released in a potential 'well' (Figure 17.31). This results in a build up of charge within the pixel, and a potential difference V develops across the pixel given by $V = q/C$, where q is the charge of the electrons and C is the capacitance of the pixel. Because this potential difference is proportional to the number of electrons released, it is also proportional to the intensity of light falling on that pixel.

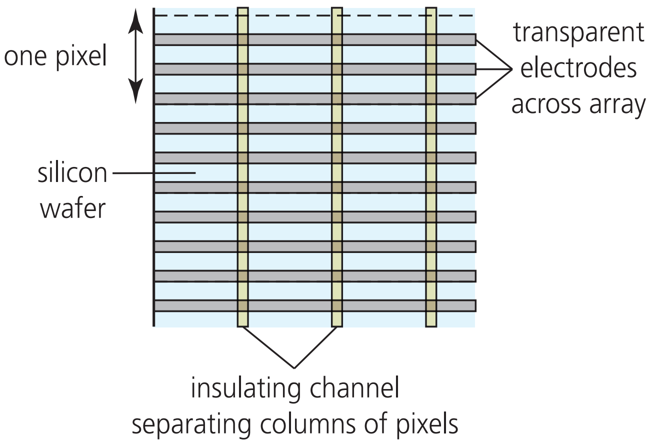


Figure 17.30 Part of the pixel array on the surface of a CCD sensor

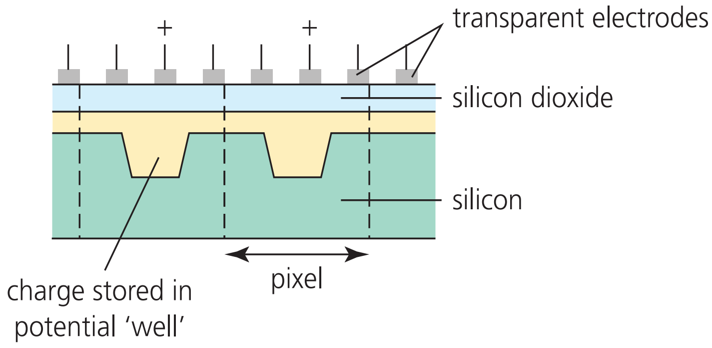


Figure 17.31 Storage of charge in a pixel array. Electrons released by the incident photons are stored in the potential 'well' beneath the positively charged electrode

Once the exposure to light has stopped, for example by closing the shutter on a camera, the charge that has built up in each pixel is stored.

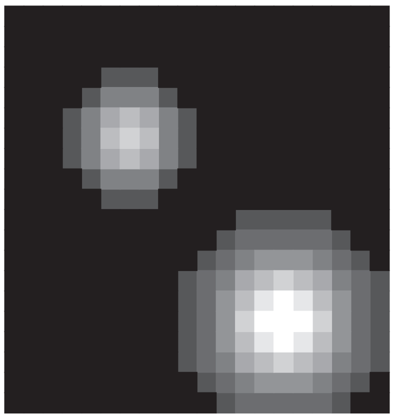
32 Find out how charge passes through a semiconductor, like silicon. (Investigate both the flow of electrons and of positively charged 'holes'.)

Processing a CCD image

C.2.4 **Outline** how the image on a CCD is digitized.
C.2.10 **Outline** how the image stored in a CCD is retrieved.

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	3	3	3	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	3	5	6	5	3	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	3	6	7	6	3	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	3	5	6	5	3	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	3	3	3	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	1	0	0
0	0	0	0	0	0	0	0	0	0	1	2	3	4	4	4	3	2	1	0
0	0	0	0	0	0	0	0	0	1	2	3	5	6	7	6	5	3	2	1
0	0	0	0	0	0	0	0	0	1	2	4	6	8	9	8	6	4	2	1
0	0	0	0	0	0	0	0	0	1	2	4	7	9	9	9	7	4	2	1
0	0	0	0	0	0	0	0	0	1	2	4	6	8	9	8	6	4	2	1
0	0	0	0	0	0	0	0	0	1	2	3	5	6	7	6	5	3	2	1
0	0	0	0	0	0	0	0	0	0	1	2	3	4	4	4	3	2	1	0
0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	1	0	0

a



b

Figure 17.32 **a** Data from a CCD (as decimal numbers) **b** Interpretation of numbers as light intensity levels

Image formation

The stored charges across the pixel array represent the variation in intensity of the light that fell upon the detector surface. It is analogue information. By measuring and digitizing the voltage developed across each pixel, at the same time recording the position of each pixel, a digital intensity 'map' of the CCD surface is built up. This intensity 'map' is in effect an image of the light focused onto the CCD (Figure 17.32).

Figure 17.32a shows the typical data from a CCD consisting of an array of numbers, running from 0 to 9 in this simplified example (rather than using binary numbers). Each number represents the intensity of the light incident on that specific pixel. The larger the number, the greater the intensity of the incident light. These numbers correspond to the intensity levels shown in Figure 17.32b.

Retrieving the information from a CCD

As shown in Figure 17.30, the pixels are arranged in columns that are separated from one another by an insulator. By applying a potential difference to corresponding electrodes on each row of pixels, the charges in each row are pushed down to the row below. The charge on the bottom row moves off the array onto the **serial register** (Figure 17.33).

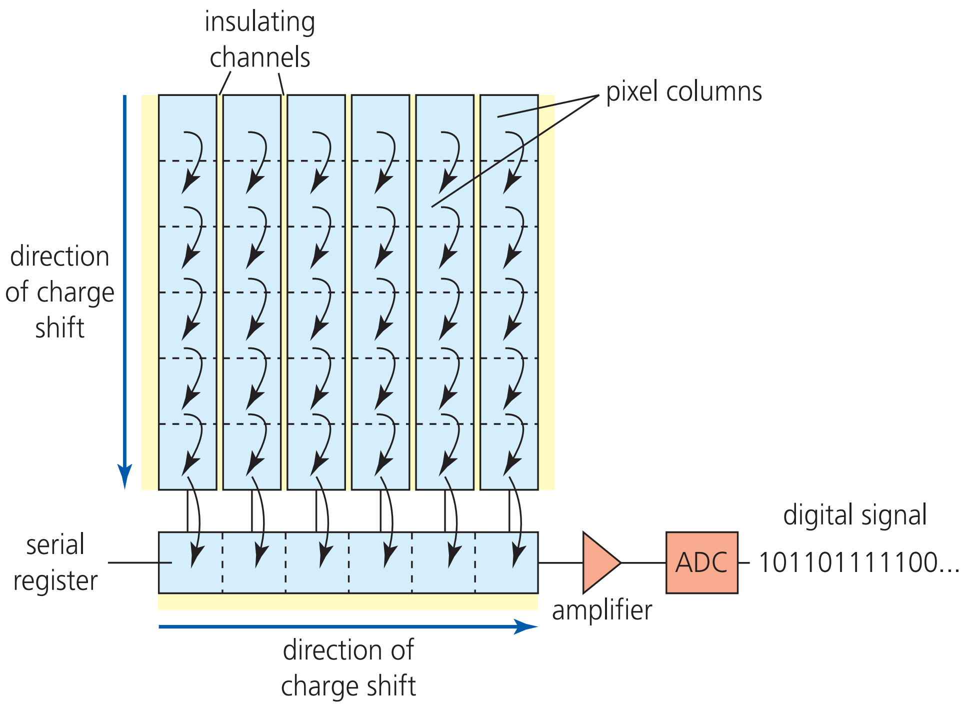


Figure 17.33 A clocking potential shifts the charges down the array

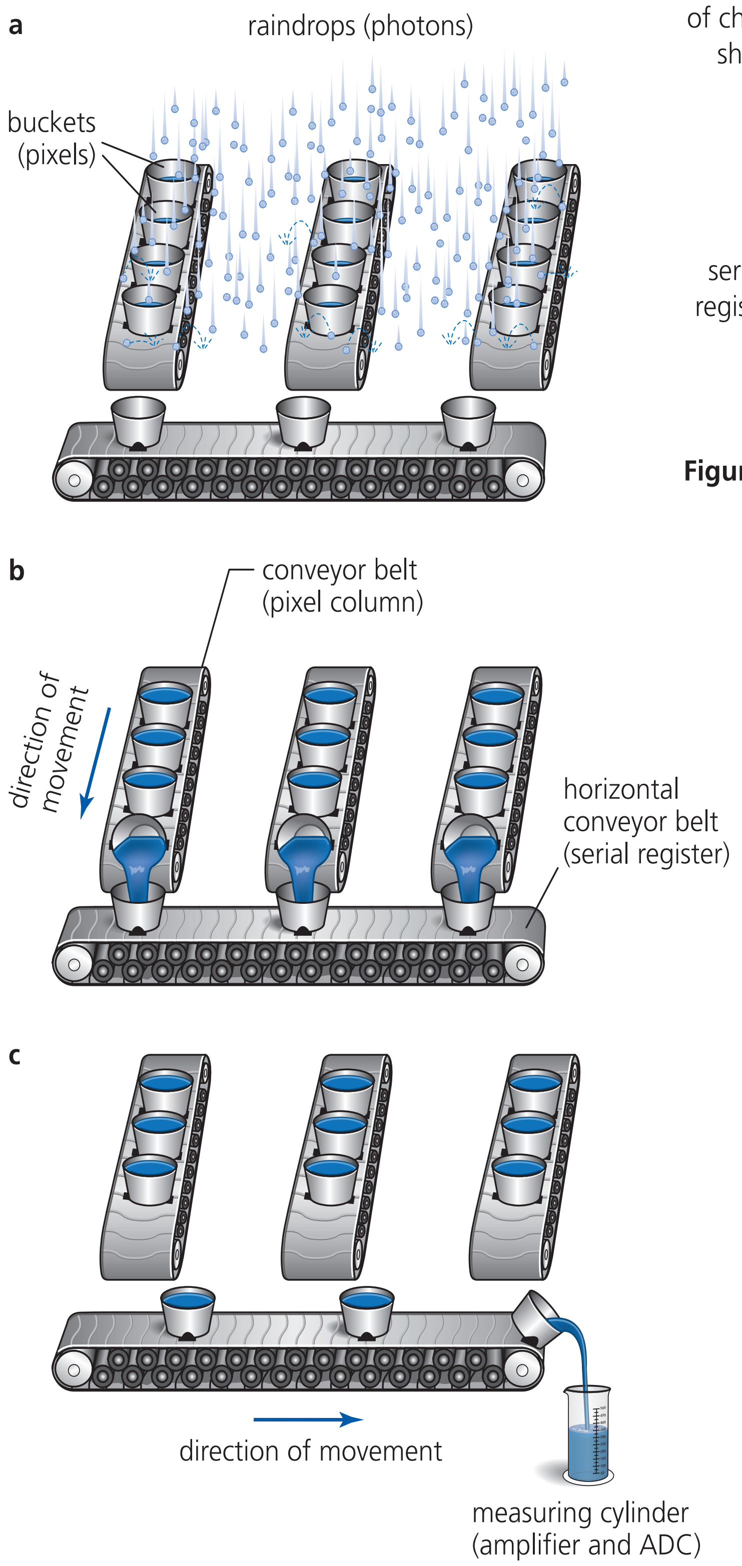


Figure 17.34 A simple mechanical analogy for the clocking of a CCD

The packets of charge on the serial register are then moved sideways. One by one they pass through an amplifier and then to an analogue-to-digital converter (ADC), which outputs the value in digital form. This process continues until the charge in the entire row of pixels has been read and converted to a digital signal. This process is known as 'clocking' the CCD.

The process is repeated for the next row, until the entire array has been processed. The name 'charge-coupled' device comes from this process, where the stored charge passes down from one row to the next to be read. Once the data from whole array has been read, digitized and stored, which can happen very quickly, the pixels can be reused to record another image.

Figure 17.34 shows a simple mechanical analogy (using rain and buckets) describing how a CCD is clocked and its charge collected. Once the exposure to rain is finished the buckets will contain samples of rain water (Figure 17.34a). The conveyor belt starts turning and transfers the water in the first buckets (only) to other buckets on a stationary 'horizontal' belt at the end (Figure 17.34b). The conveyer belts stop and the horizontal conveyor starts up and tips each bucket in turn into the measuring cylinder (Figure 17.34c). After each bucket has been measured, the measuring cylinder is emptied, ready for the next bucket load.

To produce a black and white image, the only information required is the overall light intensity on each pixel and the pixel's position (its *x*- and *y*-coordinates). Computer software can use the digital data retrieved from the CCD to generate a black and white image of the scene recorded. This can be displayed on a computer screen and be further processed or manipulated.

A colour image requires more information. One type of colour CCD has the pixels arranged into groups of four, with green, red and blue filters, known as the Bayer array filter (Figure 17.35). There are two green filters because the human eye is more sensitive to green light than to red or blue light. The intensity of the light in each of the four pixels is measured. The group of four pixels acts as a sensor unit giving information for all three primary colours for that area. Computer software combines the information to produce digital data that can recreate a full colour image.

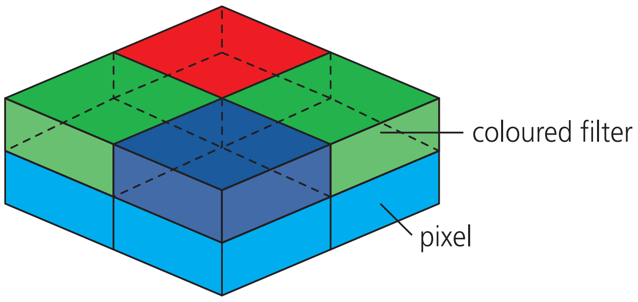


Figure 17.35 A group of four pixels with green, red and blue filters (Bayer array filter)

Characteristics of a CCD image

Quantum efficiency

C.2.5 Define *quantum efficiency* of a pixel.

Not every photon that is incident on a pixel of a CCD will cause an electron to be released. Some photons are reflected and other photons may pass through the pixel.

The **quantum efficiency** of a pixel is definded as the ratio of the number of emitted electrons to the number of incident photons. It is an accurate measurement of a CCD's electrical sensitivity to light.

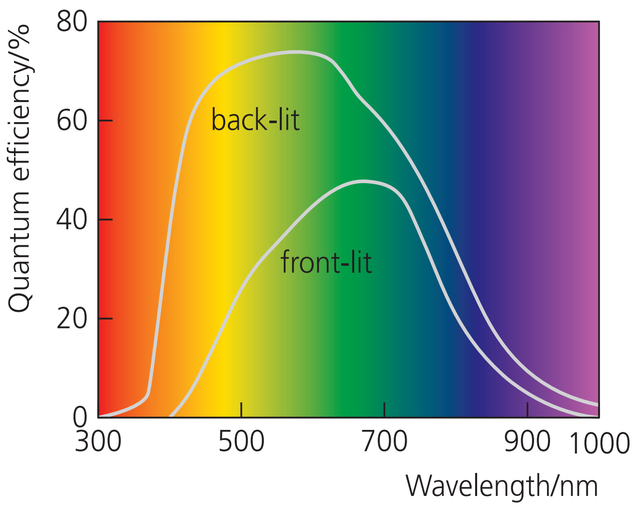


Figure 17.36 Comparison of the quantum efficiency of front-lit and back-lit CCDs over the range 300–1000 nm

A higher quantum efficiency means that a clear image is formed even at low light intensity. This is very important in astronomical imaging where very faint light signals are being studied. Since the energy of a photon varies with wavelength, quantum efficiency is often measured over a range of different wavelengths to quantify a CCD's efficiency at each photon energy.

Charge-coupled devices have very high values of quantum efficiency ranging between 70% and 80%. The human eye has a quantum efficiency of only about 20%; photographic film (an analogue medium) has a quantum efficiency of around 10%. However, the quantum efficiency is not constant for all wavelengths of light and is generally higher for back-lit CCDs than for those that are illuminated from the front (Figure 17.36). The use of back-lit CCDs in digital cameras allows the user to take high-quality night shots.

33 One-fifth of the photons incident on a CCD do not result in electrons being emitted. Calculate the quantum efficiency of the CCD.

34 There are 50 photons incident on a CCD pixel (photodiode) and 30 electrons are released. Calculate the quantum efficiency.

Magnification

C.2.6 Define *magnification*.

The magnification of the image on a CCD is defined as the ratio of the length of the image as it formed on the CCD to the actual length of the object. Magnification has no units.

$$\text{magnification} = \frac{\text{length of the image on CCD}}{\text{actual length of the object}}$$

The magnification of a CCD system is determined by the overall properties of the lenses that are used to focus the light from the CCD. A greater magnification means that more pixels are used for a given section of the image. Hence, the image will be more detailed.

Worked example

11 A digital camera is used to take a photograph of a small insect embryo. The area of the embryo is $0.012\,cm^2$ and the area of the image is $9.8 \times 10^{-6}\,m^2$. Calculate the magnification of the CCD.

The ratio of the image to the object areas is:

$$\frac{9.8 \times 10^{-6}}{1.2 \times 10^{-6}} = 8.2$$

So the ratio of corresponding linear sizes is $\sqrt{8.2} = 2.9$
Therefore the magnification is 2.9.

Resolution

C.2.7 State that two points on an object may be just resolved on a CCD if the images of the points are at least two pixels apart.

An important characteristic of a CCD is its ability to *resolve* two points very close to each other on an object whose image is required (resolution was discussed in Chapter 11).

Two points are resolved (distinct) if their images are at least two pixels apart (Figure 17.37), so that there is a noticeable decrease in intensity between them.

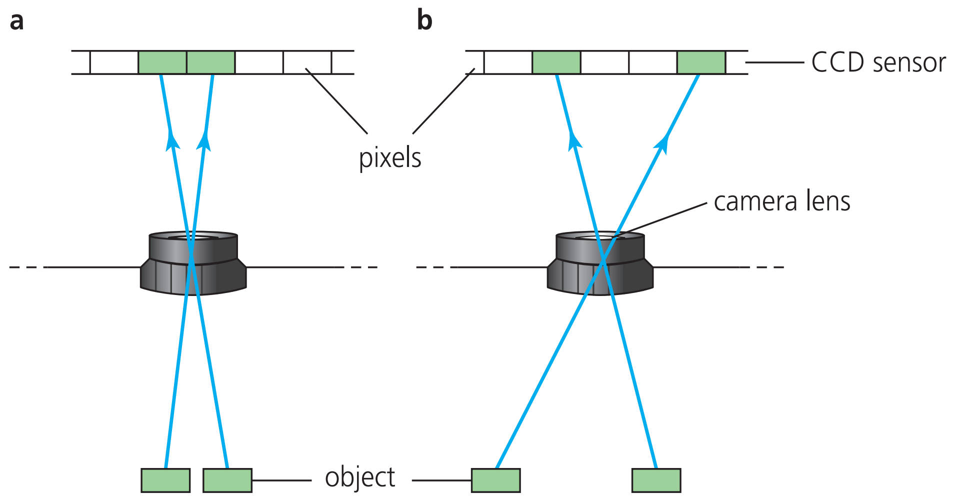


Figure 17.37 A pair of objects which are **a** not resolved and **b** resolved by a CCD

Worked example

12 The magnification produced by a 8.0 megapixel digital camera with a light-collecting area of $14\,mm^2$ is 1.8. Determine if this digital camera can resolve two points that are $1.2 \times 10^{-3}\,mm$ apart.

$$\text{The area of a pixel} = \frac{14}{8.0 \times 10^{-6}} = 1.8 \times 10^{-6}\,mm^2$$

and so the length of a pixel is $\sqrt{1.8 \times 10^{-6}} = 0.0013\,mm = 1.3 \times 10^{-3}\,mm$
The distance between the two points is:
$1.8 \times 1.2 \times 10^{-3}\,mm = 2.2 \times 10^{-3}\,mm$
This is less than two pixel lengths and so the points are not resolved.

Factors affecting the quality of a CCD processed image

C.2.8 Discuss the effects of quantum efficiency, magnification and resolution on the quality of the processed image.

A higher quantum efficiency means that the image produced by a CCD will require less time to form if the incident light intensity is very low. This is important in astronomy because the Earth is rotating relative to the stars in the night sky.

The greater the magnification of an object, the greater the length of the image on the CCD surface. This results in a larger number of pixels that will accumulate charge due to the incident light. This means that the image will be more detailed and have a higher resolution.

The resolution is greatest with a high pixel density (that is, number of pixels per unit area). An image of high resolution is of high quality because it includes more detail than an image of low resolution.

Practical uses of CCDs

C.2.9 Describe a range of practical uses of a CCD, and list some advantages compared with the use of film.

Charge-coupled devices have a range of practical uses, as described below.

Endoscopes

An endoscope is a thin and flexible tube with a light and lenses attached at the end that is used to look inside the body. The endoscope can make use of the body's natural openings, so it is often inserted through the mouth, nose, ear or anus. Modern electronic endoscopes are equipped with CCDs that produce high-quality colour images in real time.

Medical X-ray imaging

Special CCDs have been developed which can detect X-rays. This means that exposure times to X-rays are shorter, benefiting patients and operators. At present these devices are relatively expensive and many X-ray machines still use photographic film.

Digital cameras and video cameras

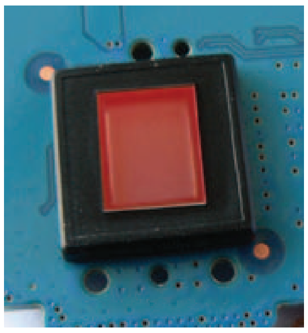


Figure 17.38 The CCD of a disposable polaroid camera

Digital cameras (Figure 17.38) are widely used and have many advantages over photographic film. The image produced by a CCD can be enhanced and edited using electronic processing techniques. The storage, archiving and sorting of a large number of photographs from a digital camera is relatively easy and cheap. The images taken by a digital camera can be viewed immediately – the processing time is very fast. A high definition digital video camera generally takes about 30 pictures per second. This means that the information from the CCD must be read very quickly, which limits the number of pixels.

Telescopes

Figure 17.39 An HST image of the Butterfly Nebula, 3800 light years away from Earth, taken using the Wide Field Camera 3

Charge-coupled devices are very useful in data collection in astronomy because they can respond to a wide range of electromagnetic radiation, and their response is in electrical form. They are also very sensitive to low intensities of light.

Electromagnetic radiation from distant sources is more intense in space than it is at the bottom of the Earth's atmosphere. The Hubble Space Telescope was launched into orbit in 1990 and does not suffer from the effects of atmospheric refraction and scattering that affect telescopes on Earth. The two cameras currently in use have specialized CCD arrays for recording different parts of the electromagnetic spectrum. Figure 17.39 shows an HST image of hot gases ejected from a dying star, a so-called planetary nebula.

35 Find about Super CCD developed by Fujifilm.

Solving problems involving CCDs

C.2.11 Solve problems involving the use of CCDs.

A variety of problems can be solved relating to CCDs. The calculations may involve energy, power, surface area and the energy carried by a photon, or the charge carried by a specific number of electrons.

Worked examples

13 Calculate the number of megapixels on a 20 mm × 20 mm CCD where the pixel size is 25×10^{-6} m.

The collecting area of the CCD is
$20\,\text{mm} \times 20\,\text{mm} = 4.0 \times 10^{2}\,\text{mm}^{2} = 4.0 \times 10^{2} \times 10^{-6}\,\text{m}^{2} = 4.0 \times 10^{-4}\,\text{m}^{2}$

and the area of each pixel is
$25 \times 10^{-6}\,\text{m} \times 25 \times 10^{-6}\,\text{m} = 6.25 \times 10^{-10}\,\text{m}^{2}$

The number of pixels is therefore

$$\frac{4.0 \times 10^{-4}}{6.25 \times 10^{-10}} = 6.4 \times 10^{5} = 0.64 \text{ megapixels}$$

14 Light of intensity $6.9 \times 10^{-6}\,\text{W m}^{-2}$ and wavelength 6.0×10^{-7} m is incident on a pixel with an area of $6.25 \times 10^{-10}\,\text{m}^{2}$. Calculate the number of photons incident on each pixel in a period of 35 ms.

Power incident on pixel area $= 6.9 \times 10^{-6}\,\text{W m}^{-2} \times 6.25 \times 10^{-10}\,\text{m}^{2} = 4.3 \times 10^{-15}\,\text{W}$
The total energy incident on the pixel in a time of 35 ms $= 4.3 \times 10^{-15}\,\text{W} \times 35 \times 10^{-3}\,\text{s}$
$= 1.5 \times 10^{-16}\,\text{J}$
The energy of one photon is

$$E = \frac{hc}{\lambda} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{6.0 \times 10^{-7}} = 3.3 \times 10^{-19}\,\text{J}$$

and hence the number of photons per pixel is equal to the total energy incident on the CCD/energy of one photon:

$$\frac{1.5 \times 10^{-16}}{3.3 \times 10^{-19}} \approx 450$$

15 The area of a pixel in a CCD is $8.5 \times 10^{-10}\,\text{m}^{2}$ and its capacitance is 54 pF. Light of intensity $3.2 \times 10^{-3}\,\text{W m}^{-2}$ and wavelength $4.6 \times 10^{-7}\,\text{m}^{2}$ is incident on the collecting area of the CCD for 150 ms. Calculate the potential difference established at the ends of the pixel, assuming that 80% of the incident photons cause the emission of electrons.

The energy incident on a pixel is
$3.2 \times 10^{-3} \times 8.5 \times 10^{-10} \times 150 \times 10^{-3} = 4.08 \times 10^{-13}\,\text{J}$

The energy of one photon is

$$E = \frac{hc}{\lambda} = \frac{6.63 \times 10^{-34} \times 3.00 \times 10^{8}}{4.6 \times 10^{-7}} = 4.32 \times 10^{-19}\,\text{J}$$

The number of incident photons is then equal to

$$\frac{4.08 \times 10^{-13}}{4.32 \times 10^{-19}} = 9.44 \times 10^{5}$$

Since the quantum efficiency of the CCD is 80%, the number of ejected electrons is
$0.80 \times 9.44 \times 10^{5} = 7.55 \times 10^{5}$

The charge corresponding to this number of electrons is
$7.55 \times 10^{5} \times 1.60 \times 10^{-19} = 1.21 \times 10^{-13}\,\text{C}$

The potential difference is then

$$V = \frac{q}{C} = \frac{1.21 \times 10^{-13}}{54 \times 10^{-12}} = 2.2 \times 10^{-3}\,\text{V} = 2.2\,\text{mV}$$

36 **a** A square CCD has a diagonal measurement of 6.0 mm. Calculate the width of the CCD.
b If each pixel is a square of side 12 μm, how many pixels will this CCD have?
c A picture of a 5.00 m high tree is taken with a digital camera containing a 6.0 mm CCD. What is the magnification of the camera if the tree image just fits on the camera's CCD?

37 **a** If 10^{11} photons enter a digital camera and are incident on the CCD of a 6 megapixel camera, how many photons land on each pixel?
b If the quantum efficiency is 75%, how many electrons are liberated in each pixel of a back-lit CCD by the photoelectric effect?

38 The rate of reading each pixel in a CCD is 8 MHz. How long will it take to read (clock) all of the pixels in a 1 megapixel CCD?

39 A video camera takes 25 frames per second (smooth action video must be shot at a minimum of 25 frames per second). If the rate of reading pixels is 6 MHz, how many pixels does the digital video camera have?

40 Find about complementary metal oxide semiconductors (CMOS) and their use as image sensors. Summarize their uses and relative advantages and disadvantages compared to CCDs.

41 Outline current research into 'vision chips' using CCDs and CMOS. Could such devices be used to help blind people (with defective retinas) see in the future? How are they currently being used in robotics?

42 CCDs used in astronomy are often cooled down to low temperatures. Find out why this is done. Refer to dark currents.

C3 Electronics

Operational amplifiers

An **operational amplifier** (widely known as an **op-amp**) is a linear amplifier used extensively in analogue circuits. An amplifier increases the amplitude of an electrical signal (Figure 17.40). A **linear amplifier** increases the amplitude of all the frequencies in the same proportion.

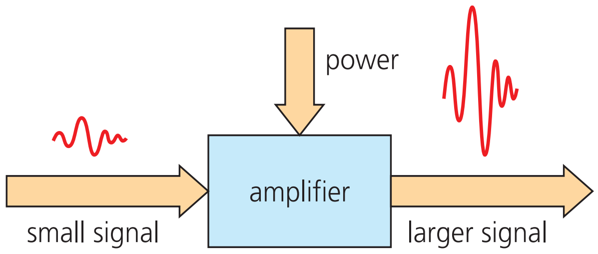


Figure 17.40 The action of an amplifier

Computers, tablets, televisions, MP3 players, mobile phones and radios all use amplifiers to increase signal amplitudes. For example, an op-amp may be used to make a voltage large enough to cause the diaphragms of a pair of loudspeakers to oscillate to generate sound waves.

Although *digital* devices are widely used in communications and data storage, they often contain *analogue* circuits, which are used when devices receive or transmit analogue data. The first computers were analogue devices.

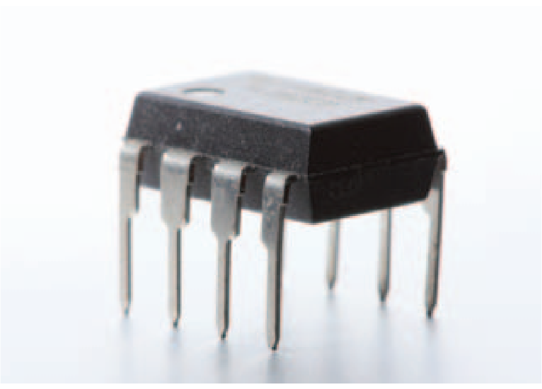


Figure 17.41 An integrated circuit op-amp

The op-amp is described as an *integrated circuit* (IC) because all the components are formed on one small slice of a silicon-based semiconductor, with 'pin' connections that allow the op-amp to be connected into a circuit. The whole of the circuit is encased in a plastic case (Figure 17.41). We are not concerned in this course with the circuitry *inside* the op-amp.

The different operations of an op-amp include:

- amplifying alternating voltages
- comparing two voltages and giving an output that depends on the result of that comparison (comparator)
- acting as a switch when a voltage reaches a certain level (Schmitt trigger).

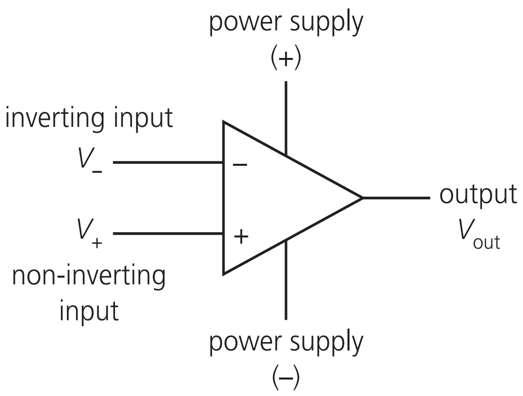


Figure 17.42 An op-amp and its circuit symbol (with power connections)

Figure 17.42 shows the symbol for an op-amp with its power connections.

The positive and negative power supplies to the op-amp provide fixed voltages, typically ±15 V. Sometimes the power supply lines are not included in diagrams. An op-amp has two inputs, called the **inverting input** (V_-) and the **non-inverting input** (V_+).

The basic function of an op-amp is as a high-gain **differential amplifier**. This means that the output of the op-amp, V_{out}, is equal to the *difference* between the two inputs multiplied by a number (much greater than one) called the **open loop gain**, G_0, which is typically 10^5 or more. (In this case it is called the *open loop* gain because no connection has been made between the output and the input.)

$$V_{out} = G_0(V_+ - V_-)$$

If a *very small* voltage is applied to the *non-inverting input*, V_+, while at the same time the inverting input is **earthed/grounded** (kept at 0V), an output voltage equal to G_0V_+ appears between the output terminal and 0V. If the input is positive, then the output is positive; if the input is negative, the output is negative. This is shown in Figure 17.43, but note that the scales used for voltages on the two axes are very different. It should also be stressed that there is only a small region in which the difference between the two input voltages is proportional to the output voltage. Beyond this region the output voltage is constant (*saturated*) and equal to the supply voltage (or, in practice, a little less).

The reverse is true when using the *inverting input* while keeping the non-inverting input grounded at 0V. A very small positive input, V_-, produces a negative output $-G_0V_-$, and a negative input produces a positive output $+G_0V_-$.

Figure 17.44 summarizes these basic op-amp characteristics.

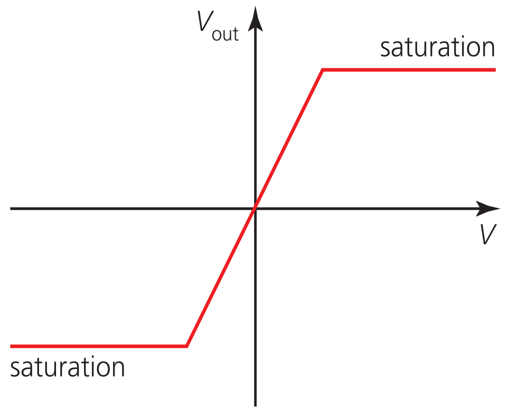
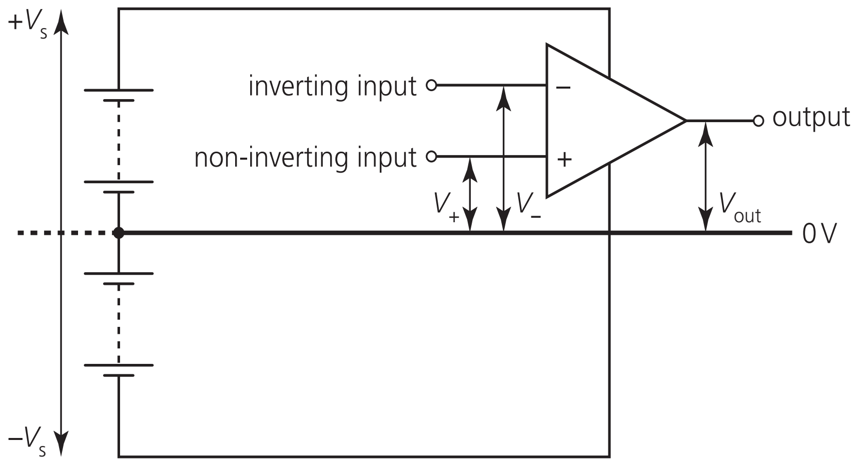


Figure 17.43 The variations in input and output voltages for an op-amp (using the non-inverting input)



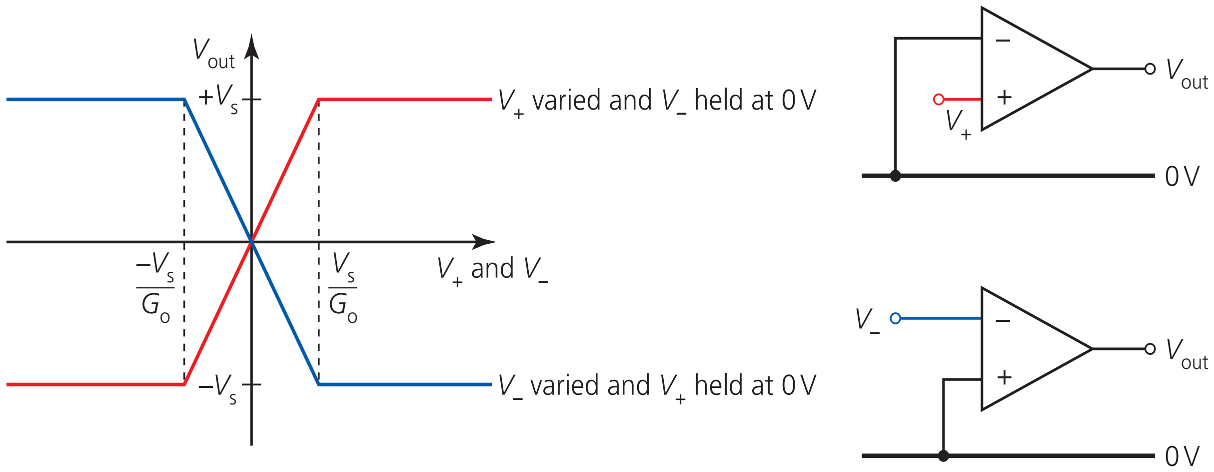


Figure 17.44 The characteristics of an op-amp showing how the output voltage varies with input voltage for the inverting and non-inverting inputs (V_s represents the saturation voltage, usually about ±13V)

When used to amplify varying analogue signals, the output must be kept below saturation. Figure 17.45 compares the inputs and outputs to an op-amp amplifying a sinusoidal waveform.

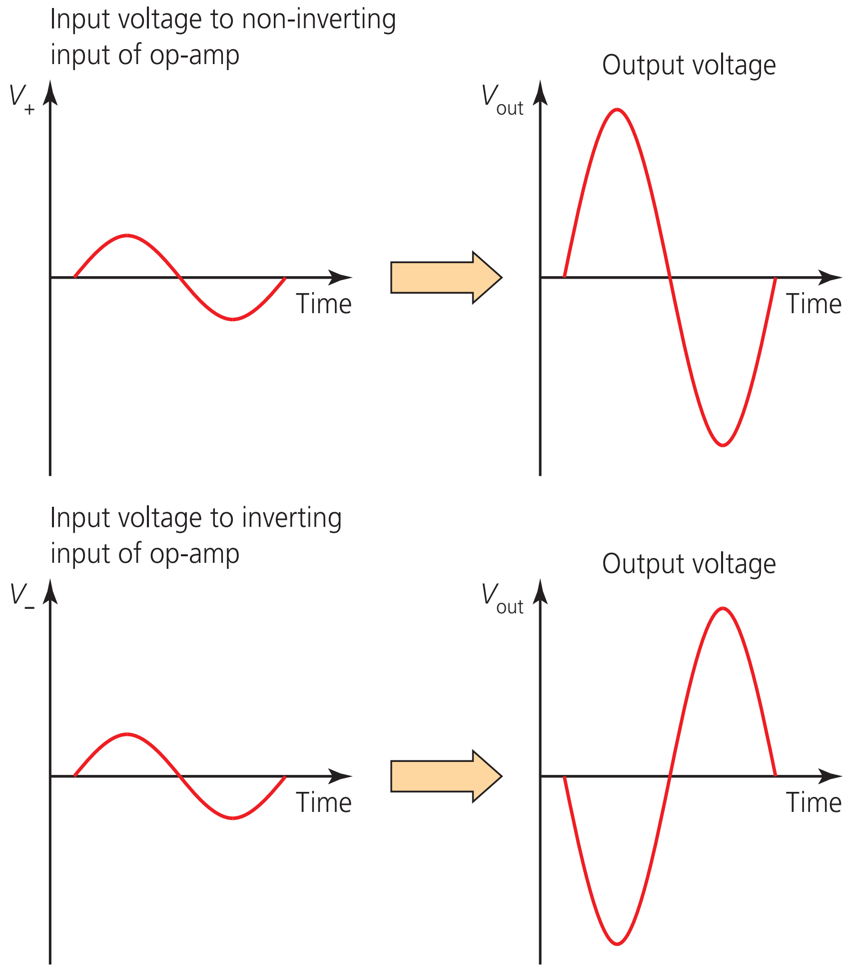


Figure 17.45 The output from the op-amp is in phase with the input if the non-inverting input is used, and in antiphase if the inverting input is used (input and output graphs are not to the same scale)

C.3.6 **Solve** problems involving circuits incorporating operational amplifiers.

43 An op-amp which saturates at ±12 V has an open loop gain of 5.0×10^5.
- **a** Calculate the output voltage when there is a difference between the inputs ($V_+ - V_-$) of:
  i 14 μV **ii** 240 μV **iii** −19 μV **iv** −58 μV.
- **b** Calculate the output voltage when:
  - **i** V_+ = 932 μV and V_- = 910 μV
  - **ii** V_+ = 730 μV and V_- = 743 μV
  - **iii** V_+ = 68 μV and V_- = 123 μV.

44 One input to an amplifier has a peak voltage of 0.1 mV and the other input is grounded. The output has a peak voltage of 1 V. What is the open loop gain of the amplifier?

45 An op-amp has a gain of 1×10^6 and is connected to a ±15 V power source, with one input grounded. What is the voltage output if:
- **a** the non-inverting input is 4 μV
- **b** the non-inverting input is 4 V
- **c** the inverting input is 4 V?

46 **a** An operational amplifier has an open loop gain of 120 000. The voltage supply is between +15 V and −15 V. With the inverting input grounded, approximately what voltage change on the non-inverting input in (μV) is needed for the output to go from positive saturation to negative saturation?
b What is the significance of the answer to **a**?

Properties of an ideal operational amplifier

C.3.1 **State** the properties of an ideal operational amplifier (op-amp).

To simplify the analysis and calculations involving op-amps they can be assumed to behave in an *ideal* way. The **ideal operational amplifier** (op-amp) has the following properties:

- *Infinite input resistance* (so that negligible current enters or leaves either of the inputs). Consider connecting the input of an op-amp across one of the resistors in a potential divider (Chapter 5): the op-amp would be in parallel with the resistor, so that when making the connection it would be expected to make the voltages in the potential divider circuit change.

However, if the op-amp has infinite input resistance, then connecting it to a potential divider will not affect the voltage input.

- *Zero output resistance* (so that the whole of the output voltage appears across any resistance that is connected across the output, which is usually called the *load resistance*).
  If the output connection has some resistance, then the output voltage will be divided between the output and the load resistance. This is similar to the effect of the internal resistance of a battery in a simple circuit.
- *Infinite open loop gain*, G_0 (so that that when there is only a very small input voltage, the amplifier will saturate and the output will have its maximum value).

It is important to realize that these assumptions are made in order to greatly simplify the analysis of the behaviour of op-amps but a typical op-amp has an input resistance of about 1 MΩ and an output resistance of about 100 Ω.

C.3.6 Solve problems involving circuits incorporating operational amplifiers

47 a Resistors of 500 kΩ and 1 MΩ are connected in series with a 12 V battery of negligible internal resistance. What is the p.d. across the 1 MΩ resistor?

b The input to an op-amp, which has an input resistance of 2 MΩ, is connected across the larger resistance. What is the input voltage?

c What would be the input voltage if the op-amp had infinite input resistance (as is often assumed)?

48 Use the Internet or an electronics catalogue to find out the specifications of the widely used 741 op-amp.

49 Strictly speaking, the op-amp should be described as having infinite input 'impedance' and zero output 'impedance'. Research the concept of 'impedance' in ac circuits.

Additional Perspectives

Decibels, bandwidth and slew rate

The *decibel* is a common unit of measurement in electronics for the comparison of two power or voltage levels. (It is also widely used in the comparison of sound levels.) The expression for the power gain of an amplifier in dB is $10\log_{10}(P_{out}/P_{in})$ and the voltage gain is expressed as $20\log_{10}(V_{out}/V_{in})$. The decibel level is the same regardless of whether it is derived from a power gain or from a voltage gain.

If an alternating voltage is applied to the input of an op-amp, then the output will have the same frequency but a larger amplitude. The range of frequencies that is amplified by the same amount (the input signals of different frequencies that all have the same gain) is known as *bandwidth* (Figure 17.46). An ideal linear op-amp has infinite bandwidth, which means all frequencies are amplified equally. (Note that the term *bandwidth* has an alternative meaning in computing, as the rate of transfer of data, for example an Internet connection might be 600 Mbits per second.)

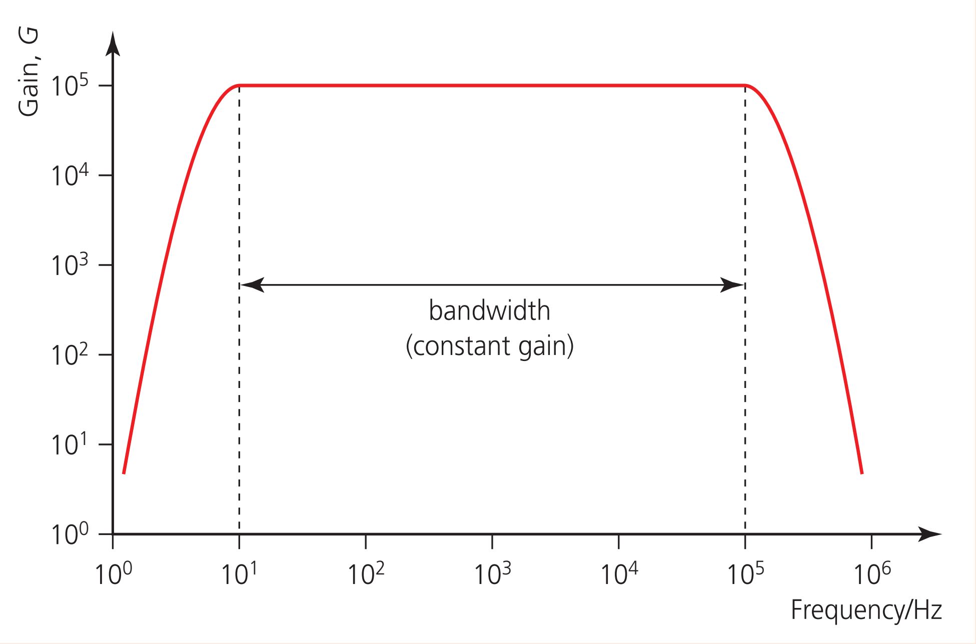


Figure 17.46 Frequency response for an op-amp

When the input signal into an op-amp is changed, then the output signal will also change. The *slew rate* is a measure of the time delay between the changes to the input and output. A high slew rate implies a short time delay. With an infinite slew rate there is no delay. An ideal op-amp operates with an infinite slew rate.

Questions

1 The input to an amplifier is 2.4×10^{-3} V and the output voltage is 4.3 V. Calculate the gain of the amplifier in decibels.

2 Show that the expressions gain (dB) $= 20\log_{10}\left(\frac{V_2}{V_1}\right)$ and gain (dB) $= 10\log_{10}\left(\frac{P_{out}}{P_{in}}\right)$ are equivalent.

Amplifiers and negative feedback

C.3.2 **Draw** circuit diagrams for both inverting and non-inverting amplifiers (with a single input) incorporating operational amplifiers.
C.3.3 **Derive** an expression for the gain of an inverting amplifier and for a non-inverting amplifier.

Feedback is the name given to the technique in which the output of a process is then used to alter the input, and so affect future outcomes. In an example from everyday life, the management of a hotel may ask customers for their opinions after they have stayed overnight. The information can then be used to improve the service provided. In general, feedback may be either *positive* or *negative*, having the primary effect of increasing or decreasing the output.

Negative feedback is widely used in amplifiers, as represented by Figure 17.47. Negative feedback greatly reduces gain, but this has a number of significant benefits for the performance of op-amp circuits:

- greater stability
- less distortion of the output
- (increased bandwidth).

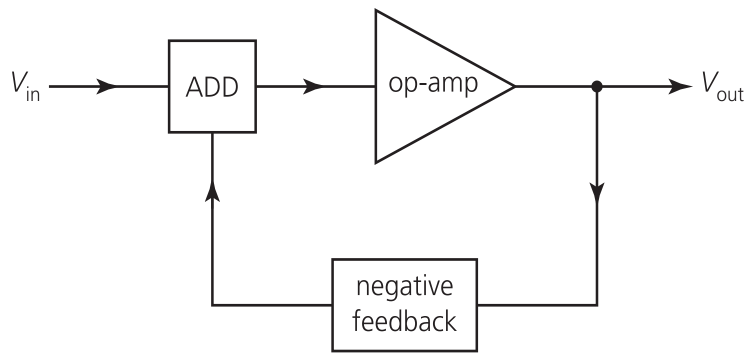


Figure 17.47 An op-amp with negative feedback

Inverting amplifiers

A circuit for an **inverting amplifier** incorporating an op-amp with negative feedback is shown in Figure 17.48. For simplicity, the power supplies are not shown. The 0 V line can be kept at 0 V by connecting it to the Earth (grounding it), and it is then described as an **earth connection**. It may be represented by the symbol shown in the figure. The non-inverting input is kept at 0 V.

The resistor R_F introduces feedback into the circuit. Because it is connected to the inverting input this is negative feedback. The resistors R and R_F act as a potential divider between the input and the output of the op-amp.

Since the open loop gain of the op-amp is very large, the difference in inputs must be very small (assuming that the output is not saturated), so that the input voltages at the non-inverting (V_+) and the inverting (V_-) inputs must be almost the same. Since the non-inverting input is connected to 0 V, the inverting input must also be approximately 0 V. Point P in Figure 17.48 is known as a **virtual earth**.

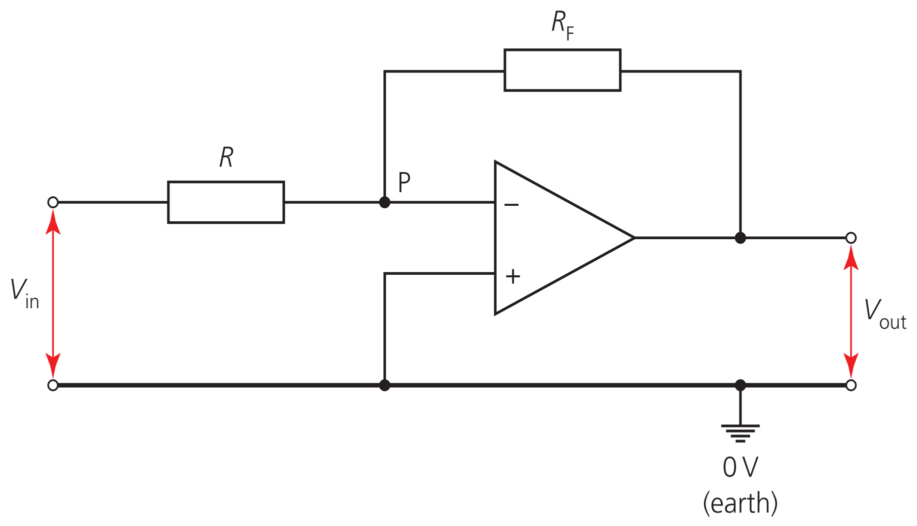


Figure 17.48 An inverting amplifier incorporating an op-amp

The input resistance of the op-amp is very large, which means that there is almost no current flowing in either the inverting or non-inverting inputs. So the current to or from the input signal is assumed to pass through the feedback resistor to or from the output. This is shown by the arrows in Figure 17.49.

A positive input signal produces a negative output and a negative input signal produces a positive output. The output is the inverse of the input and the amplifier is referred to as an inverting amplifier.

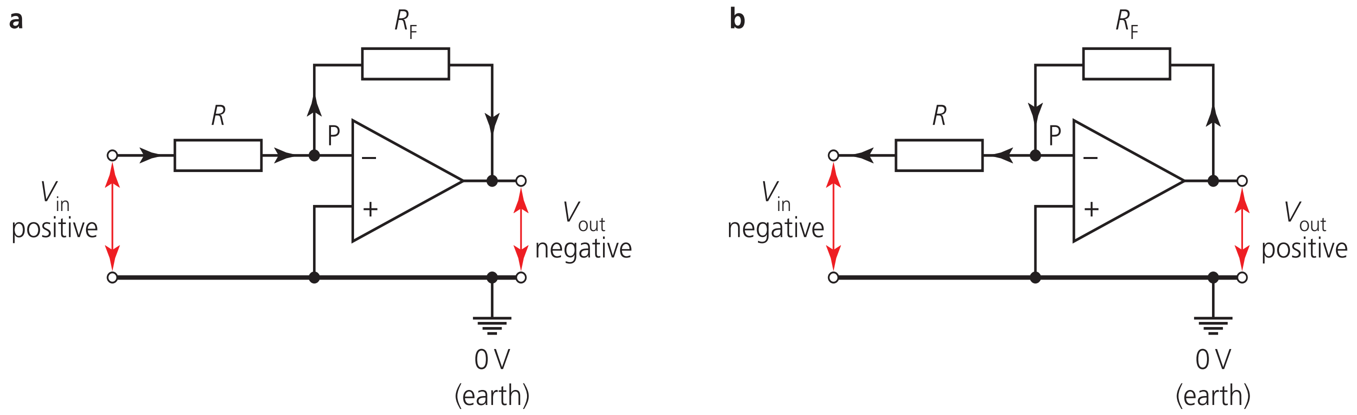


Figure 17.49 Current passes through the feedback resistor in an inverting amplifier

Referring to Figure 17.49a, since the input resistance is assumed to be infinite, the current in R equals the current in R_F. Assuming that the resistors are ohmic, $I = \frac{V}{R}$, so that:

$$\frac{\text{p.d. across } R}{R} = \frac{\text{p.d. across } R_F}{R_F}$$

P is a virtual earth at 0 V, so that:

$$\frac{V_{in} - 0}{R} = \frac{0 - V_{out}}{R_F}$$

The overall voltage gain, (G), of the amplifier circuit is given by the following expression. Because the circuit has feedback, G is called the **closed loop gain**.

$$G = \frac{V_{out}}{V_{in}} = -\frac{R_F}{R}$$

This equation is in the IB *Physics data booklet*. Note that this equation can only be used if the op-amp is not saturated.

The magnitude of the closed loop gain of the op-amp can be selected by appropriate choice of resistors and typical values might be $R = 10^4\,\Omega$ and $R_F = 10^5\,\Omega$.

Figure 17.50 shows the out-of-phase sinusoidal input and output voltages for an inverting amplifier.

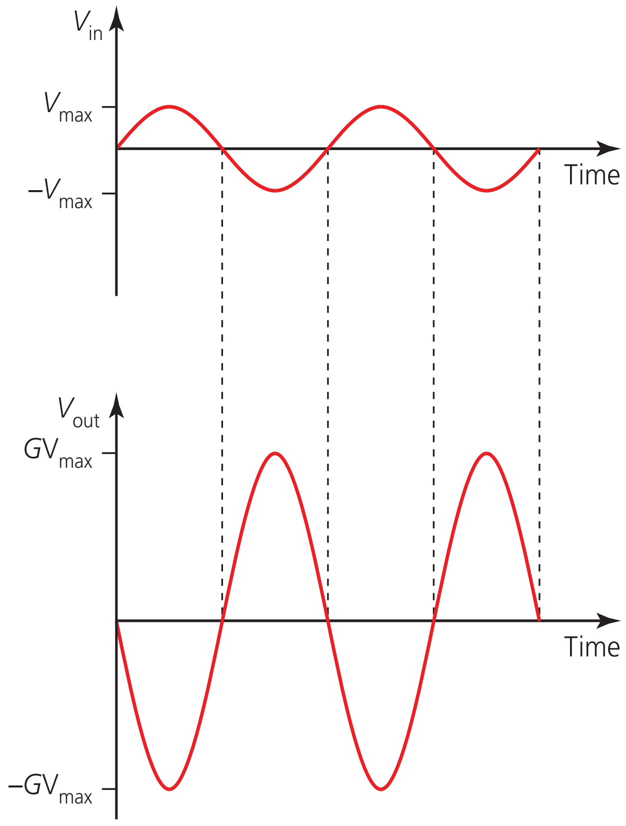


Figure 17.50 Input and output signals for an inverting amplifier

Worked example

16 An inverting op-amp has a feedback resistor of 200 kΩ and a resistor of 10 kΩ connected to the input. It receives an input signal of +0.5 V. Calculate the output voltage.

$$G = \frac{V_{out}}{V_{in}} = -\frac{R_F}{R} = -\frac{200}{10} = -20$$

$V_{out} = G \times V_{in}$

$V_{out} = -20 \times 0.5 = -10\,V$ (assuming this is below saturation)

C.3.6 Solve problems involving circuits incorporating operational amplifiers.

50 What value of feedback resistance will result in an output of 6.7 V when an input voltage of −1.2 V is applied to the inverting input of an op-amp using a resistor of 2.5 kΩ connected to the input?

51 An inverting amplifier has a resistor of 20 kΩ connected to the input and a feedback resistance of 1.0 MΩ. The supply voltages to the op-amp are ±10 V and the non-inverting input is kept at 0 V. Calculate the output voltage of the amplifier circuit for a voltage at the inverting input of:

a +40 mV
b −80 mV
c −1.2 V

52 Consider Figure 17.48. If V_{in} = 25 mV, R = 100 kΩ, and R_F = 1.2 MΩ, calculate:

a the voltage at P
b the current in R
c the current in R_F
d V_{out}.

53 Sketch the input/output voltage characteristic of the inverting amplifier discussed in Worked example **16**. Assume that it saturates at ±12 V.

54 Find out about the 'maximum power theorem' – a useful idea from dc circuits that helps in the design of efficient electronic systems.

Non-inverting amplifiers

Figure 17.51 shows the circuit for a non-inverting amplifier.

The input voltage, V_{in}, is applied directly to the non-inverting input. Negative feedback occurs because a connection is made from the potential divider across the output (R_F and R), back to the inverting input. If the amplifier is not saturated, then the potential difference between the inputs of the op-amp will be almost 0 V, so that *both* inputs can be assumed to be equal to V_{in}. Then the voltage at point P will also be equal to V_{in}.

Since we can assume that no current flows into the inputs, the current in R_F equals the current in R.

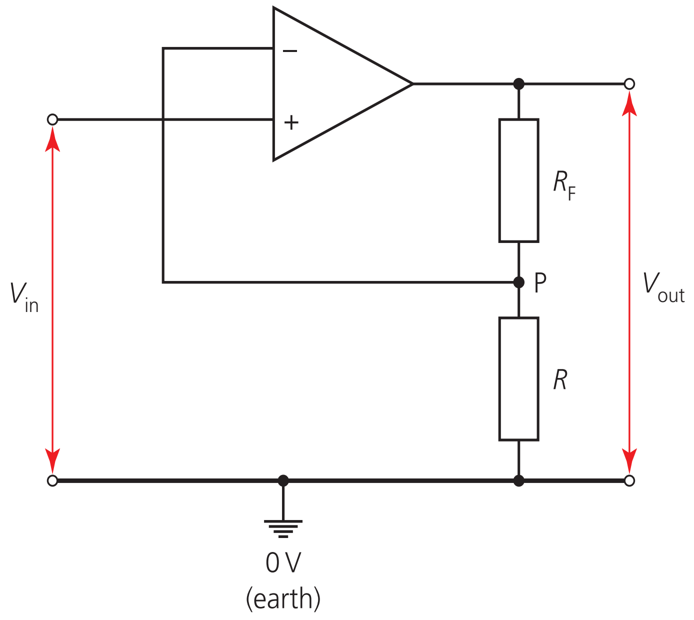


Figure 17.51 A non-inverting amplifier

$$I = \frac{V_{in}}{R} = \frac{V_{out} - V_{in}}{R_F}$$

$$V_{out} = \left(\frac{R_F}{R} \times V_{in}\right) + V_{in} = V_{in}\left(1 + \frac{R_F}{R}\right)$$

$$G = \frac{V_{out}}{V_{in}} = 1 + \frac{R_F}{R}$$

This equation is in the IB *Physics data booklet*.

The non-inverting amplifier produces an output that is in phase with the input. If the input is positive, the output is positive, and if the input is negative, the output is negative.

As with the inverting amplifier, the magnitude of the closed loop gain of the non-inverting amplifier can be selected by appropriate choice of resistors, but note that this equation can only be used if the op-amp is not saturated.

For most amplifier applications it is not important whether the output signal is inverted or not, so that the non-inverting amplifier may be the better choice, since it has a slightly higher gain (for resistors of the same value).

Worked example

17 Figure 17.52 shows a circuit for a non-inverting amplifier. Calculate possible values of R_1 and R_2 so that the gain may be set at any value between 5 and 10.

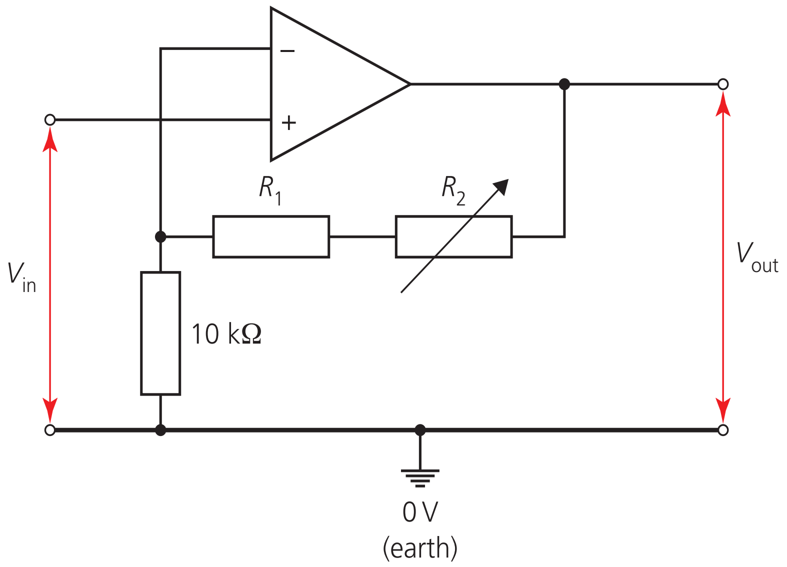


Figure 17.52

$$G = \frac{V_{out}}{V_{in}} = 1 + \frac{R_F}{R} = 1 + \left(\frac{R_1 + R_2}{10^4}\right)$$

If the gain is to be at its lowest value (5), then R_2 must be set to its lowest value (0), so that:

$$5 = 1 + \frac{R_1}{10^4}$$

$$R_1 = 4 \times 10^4 \Omega = 40\,\text{k}\Omega$$

If the gain is to be at its highest value (10), R_2 can be determined from:

$$10 = 1 + \left(\frac{4 \times 10^4 + R_2}{10^4}\right)$$

$$R_2 = 5 \times 10^4 \Omega = 50\,\text{k}\Omega$$

So, R_2 is a variable resistor with a range from 0–50 kΩ.

C.3.6 Solve problems involving circuits incorporating operational amplifiers.

55 Figure 17.53 shows an alternative way of drawing a non-inverting amplifier.
- **a** Calculate the gain of this amplifier.
- **b** Calculate the output voltage if the input voltage is 0.84 V.
- **c** What assumption did you make in answering **b**?

56 Find out:
- **a** what is meant by a *voltage follower* circuit
- **b** typical uses of such circuits
- **c** about the use of op-amps in voltage followers.

57 Figure 17.50 compares input and output signals for an inverting amplifier. Sketch similar graphs to represent the behaviour of a non-inverting amplifier.

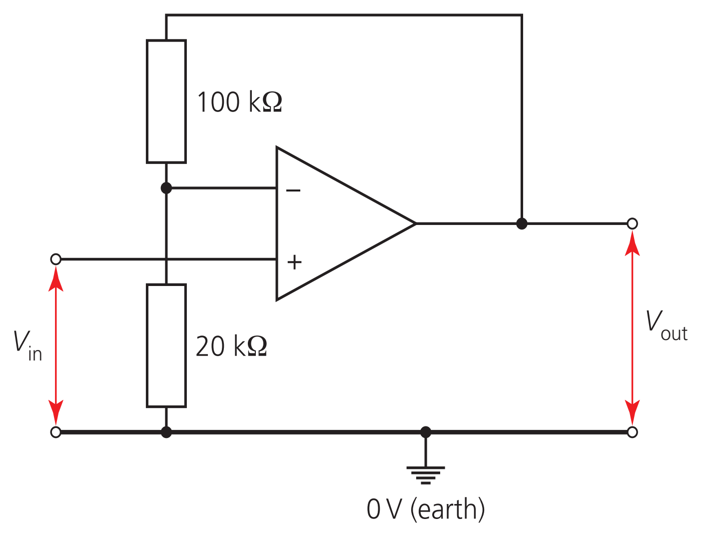


Figure 17.53

58 Figure 17.54 shows a non-inverting amplifier. The supply voltage is 9 V, $R_F = 10\,k\Omega$ and $R_1 = 500\,\Omega$.

a Calculate the gain of the amplifier.

b Calculate the output voltage if the input voltage is 0.20 V.

c Calculate the current through R_F.

d Calculate the potential difference across R_F.

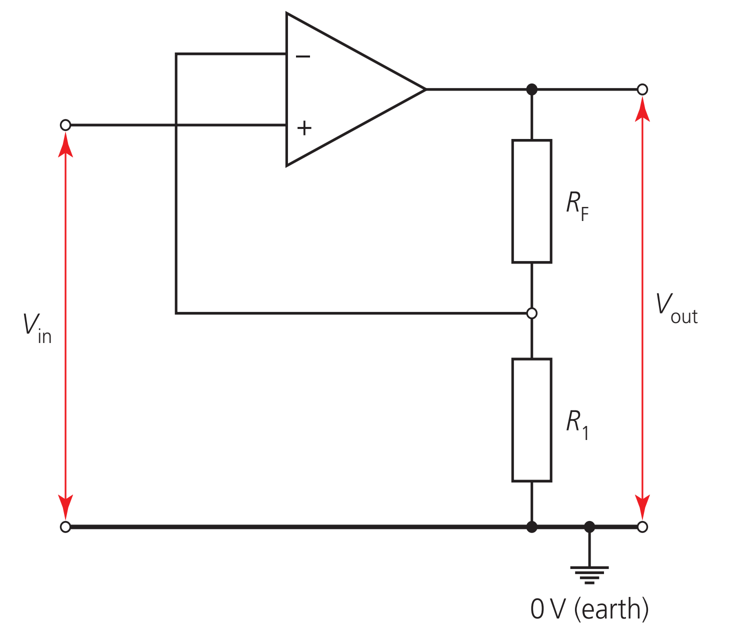


Figure 17.54

C.3.4 **Describe** the use of an operational amplifier circuit as a comparator.

The operational amplifier as a comparator

As we have seen, the basic function of an op-amp is to give an output which depends on the *difference* between its two inputs: $V_{out} = G_0(V_+ - V_-)$ for an op-amp without feedback.

For example, an op-amp may have a non-inverting input, V_+, of 0.95 V, an inverting input, V_-, of 0.94 V, a gain of 10^5 and supply voltages of ±15 V. Then:

$$V_{out} = 10^5 \times (0.95\,V - 0.94\,V) = +1000\,V$$

But the output voltage cannot exceed the power supply voltage, so the amplifier is saturated and the output voltage may be about +13 V.

But, making just a very small change, if the non-inverting input, V_+, is 0.94 V and the inverting input, V_-, is 0.95 V:

$$V_{out} = 10^5 \times (0.94\,V - 0.95\,V) = -1000\,V$$

The amplifier is saturated and the output voltage may be about −13 V.

Essentially, for an op-amp without feedback,

- if V_+ is greater than V_- then the output is approximately $+V_{supply}$
- if V_- is greater than V_+ then the output is approximately $-V_{supply}$.

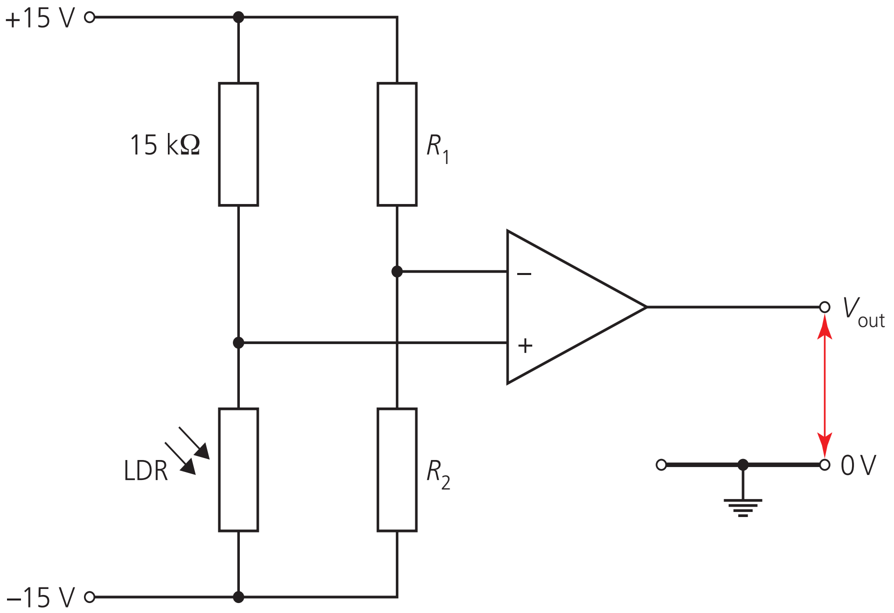


Figure 17.55 The op-amp used as a comparator to monitor illumination

If one input to an op-amp is kept at a fixed voltage (usually using a potential divider), then a changing voltage applied to the other input can be used to switch the saturation output voltage on or off. An op-amp used in this way is described as a **comparator** circuit. Typically, the input varies in response to the behaviour of some kind of transducer, such as a thermistor or an LDR (light-dependent resistor). The change of output voltage is then used to switch some kind of output device, such as a light or buzzer, on or off.

Figure 17.55 shows an example of a comparator circuit illustrating how an op-amp could be used in a circuit with an LDR to monitor illumination.

In this example there is a p.d. of 30 V across both potential dividers. Suppose that the two resistors, R_1

and R_2, have been chosen such that there is an input voltage of 5 V at the inverting input of the op-amp. If, with a certain light level, the resistance of the LDR is greater than 30 kΩ, then the voltage at the non-inverting input will be more than the 5 V on the other input, and the output voltage will saturate at about +13 V. If the light level is higher, the resistance of the LDR will be less than 30 kΩ and the voltage at the non-inverting input will be less than 5 V, and the output voltage will saturate at about −13 V.

The output therefore depends on the light intensity and this can be used to switch an output device such as a light-emitting diode (LED) on or off (depending on how it is connected). If the values of the resistors in the potential divider connected to the inverting input are changed, the voltage at which the circuit switches from +13 V to −13 V can be altered. Thus, the level of light intensity at which the circuit switches can be changed. This is often achieved by having a variable resistor instead of one of R_1 or R_2. Alternatively a three-terminal potentiometer could be connected to the inverting input.

The circuit shown in Figure 17.55 switches from +13 V to −13 V when the level of light intensity increases. But by swapping the connections to the two inputs, the output could be made to switch from +13 V to −13 V when the level of light intensity decreases.

C.3.6 Solve problems involving circuits incorporating operational amplifiers.

Other devices could also be fitted into the comparator circuit. For example, a thermistor (Chapter 5) could be used so that the circuit provides a warning for either high or low temperatures. A simple buzzer could be connected to the output in order to provide an audible alarm.

59 Calculate suitable values for resistors R_1 and R_2 in Figure 17.55 so that $V_- = 5.0$ V.

60 Make a copy of Figure 17.55 but replace R_1 and R_2 with a potentiometer. Add two LEDs (with protective resistances in series), so that one LED comes on, and the other turns off, at a particular light level.

61 If a circuit similar to Figure 17.56 was used to sound an alarm when the light level fell below a certain value, explain why a diode would have been needed in series with the alarm.

62 Figure 17.56 shows an op-amp being used as a comparator. The resistance of the thermistor at 20 °C is 50 kΩ and at 100 °C it is 5 kΩ.

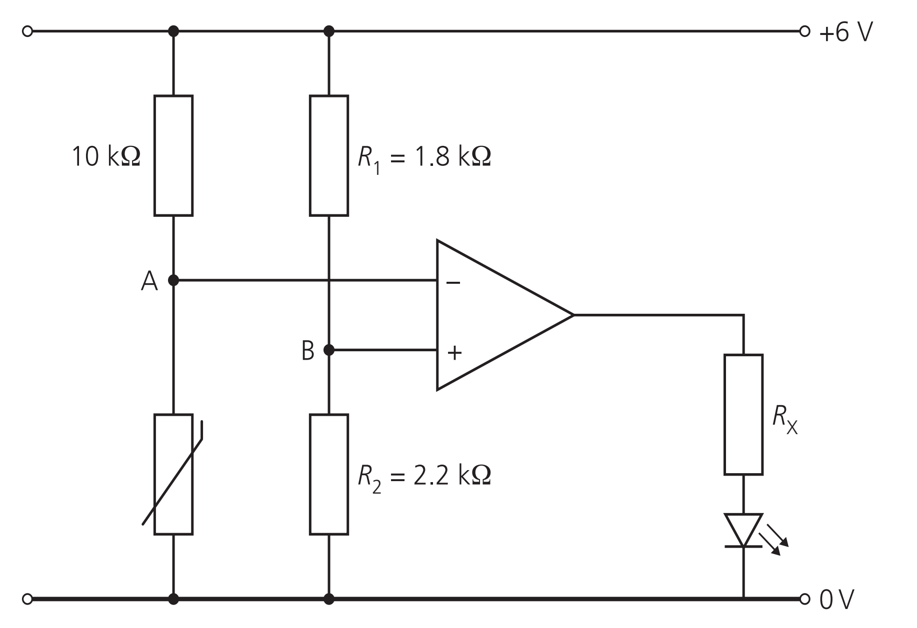


Figure 17.56

a Calculate the voltage at point A when the temperature is:
i 20 °C
ii 100 °C.
b Calculate the voltage at point B.
c The LED has a maximum voltage of 2 V across it, limiting the current to 15 mA. Calculate the value of the protective resistor, R_X, placed in series with the LED.
d Describe the function of the circuit.

63 a Adapt the circuit shown in Figure 17.56, using a power supply capable of providing ±9 V and a potentiometer, so that red and green LEDs can be used to give an indication when the temperature goes above or below an adjustable pre-set value.
b Explain how it is possible to change the temperature at which the LEDs turn on/off.

C.3.5 Describe the use of a Schmitt trigger for the reshaping of digital pulses.

The Schmitt trigger

In the transmission of a digital signal, electronic *noise* (unwanted signals) and dispersion may cause the signal to become distorted and corrupted. However, such signals can be easily regenerated with a device based on the op-amp, known as a **Schmitt trigger** (Figure 17.57), which has a single input and a single output.

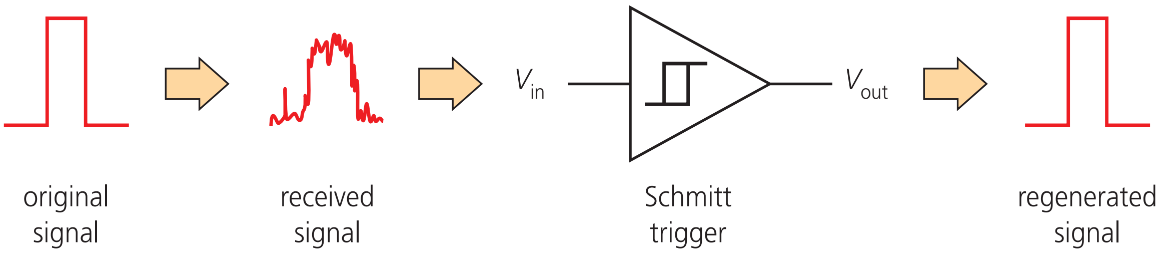


Figure 17.57 A very corrupted digital signal can be regenerated with a Schmitt trigger

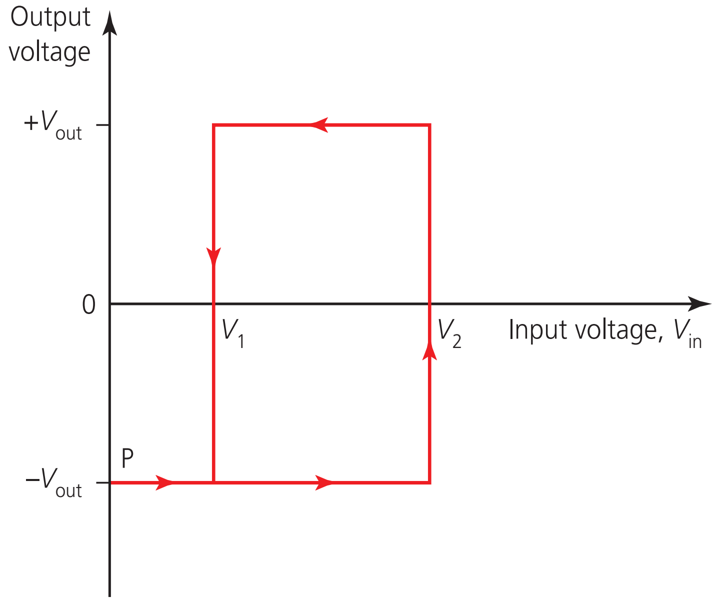


Figure 17.58 Input and output voltage characteristic of the Schmitt trigger

Figure 17.58 shows how the output voltage of a Schmitt trigger, V_{out}, changes with varying positive input voltages, V_{in}. The output can only have two possible values.

Consider the point P on the graph. For an input voltage of zero, the output is $-V_{out}$. As the input voltage rises, the output voltage is constant at this value until it reaches a value of V_2, which is called the *upper threshold input voltage*. The output then increases very quickly to the value $+V_{out}$. (This is shown in the figure as a vertical line, although in practice a very small voltage difference will be involved.) If the input signal then decreases, the output remains constant at $+V_{out}$ until the *lower threshold*, V_1, is reached, when the output decreases rapidly to $-V_{out}$.

For example, consider Figure 17.59, in which part of a corrupted digital signal is shown by the curved line. This signal has been connected as the input to a Schmitt trigger which was designed so that the threshold voltages are 0.5 V and 1.0 V. Every time the input voltage *rises* above 1.0 V the output of the Schmitt trigger changes to +3 V and every time the input *falls* below 0.5 V the output switches to –3 V. In this way the digital signal is regenerated. Because the Schmitt trigger compares the input with reference values, it can be considered as another example of a comparator circuit.

Figure 17.60 shows how an op-amp can be used in a non-inverting Schmitt trigger arrangement. The two threshold voltages are determined by the values of the resistors R_1 and R_2, and the value of the voltage at the inverting input, which can be provided from a suitable potential divider (not shown).

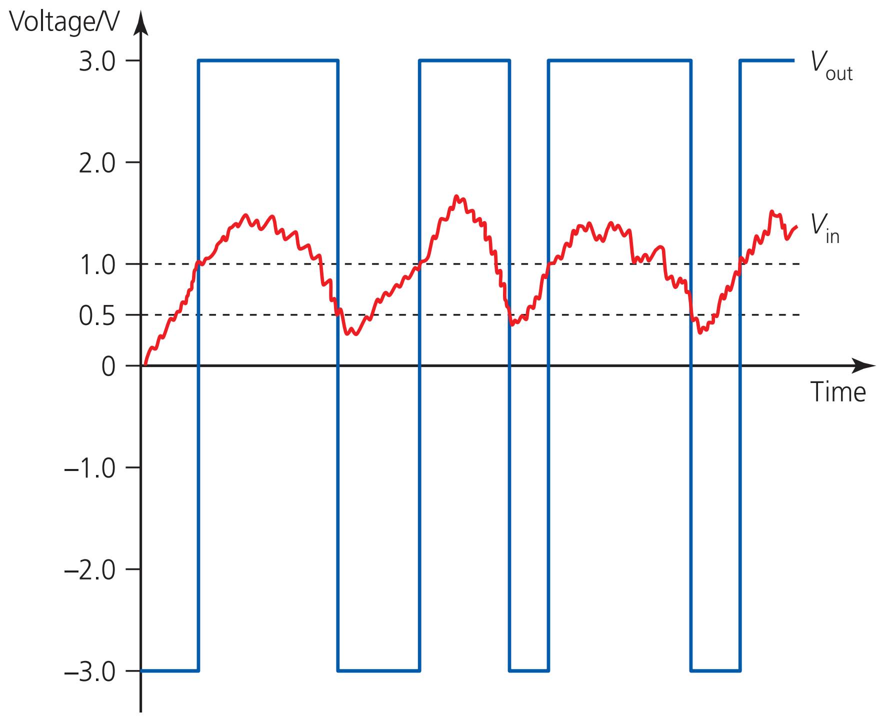


Figure 17.59 Regeneration of a corrupted digital signal by a Schmitt trigger

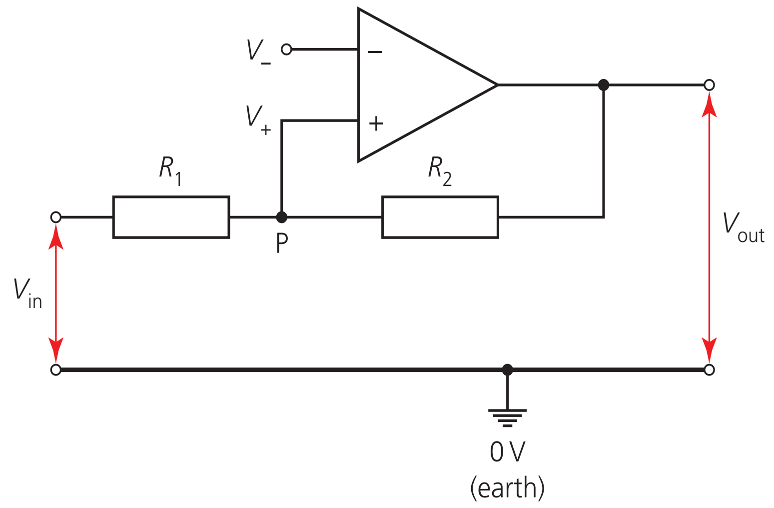


Figure 17.60 An op-amp connected as a Schmitt trigger

Worked example

18 Consider Figure 17.60. Suppose $R_1 = 5.0\,\text{k}\Omega$, $R_2 = 40\,\text{k}\Omega$, $V_- = 2.0\,\text{V}$ and the output can be switched between +12 V and –12 V. Assume also that V_{in} is such that the voltage at point P (= V_+) is less than 2.0 V, so that the output, V_{out} = –12 V. If V_{in} is increased, at what value does the output switch to +12 V?

The circuit will switch when the voltage at P (V_+) rises above 2.0 V. At that time the p.d. across R_2 will be 2 – (–12) = 14 V. So,

$$\text{current through } R_2 \text{ (using } I = \frac{V}{R}\text{)} = \frac{14}{4.0 \times 10^4} = 3.5 \times 10^{-4}\,\text{A}$$

The same current passes through R_1 (assuming the op-amp has infinite input resistance), so that:

$$\text{p.d. across } R_1 \text{ (using } V = IR\text{)} = (3.5 \times 10^{-4}) \times (5.0 \times 10^3) = 1.75\,\text{V}$$

Since the voltage at P is 2.0 V, V_{in} = 2.0 + 1.75 = 3.75 V

C.3.6 Solve problems involving circuits incorporating operational amplifiers.

64 a Determine the input voltage at which the circuit discussed in Worked example **18** will switch the output back to –12 V.

b If R_2 was a variable resistor and was changed to a higher value, how would that affect the answer to **a**?

65 Design a Schmitt trigger circuit which would have threshold switching voltages of approximately 1 V and 4 V.

■ Additional Perspectives

Summing amplifiers

Consider the circuit shown in Figure 17.61, which is an inverting amplifier with two inputs.

If each of the inputs was connected separately to the op amp, the outputs would be –8.4 V and –1.2 V.

If both inputs are applied at the same time (as in Figure 17.61), then the output can be shown to be –9.6 V as follows.

The non-inverting input of the op-amp is a virtual earth. The current through the op-amp itself is negligible, so that:

$$I_1 + I_2 = I_F$$

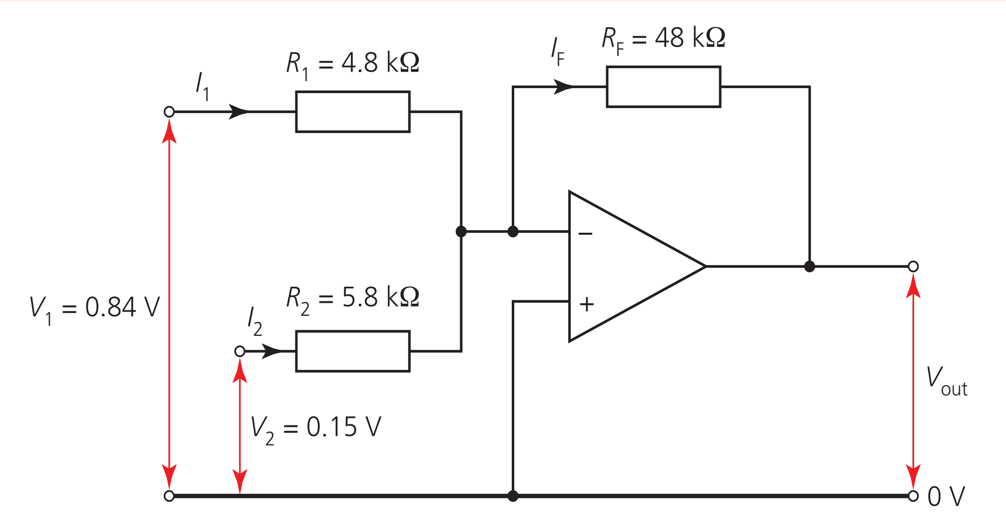


Figure 17.61 Inverting amplifier with two inputs

Applying Ohm's law:

$$I_1 = \frac{V_1}{R_1} \qquad I_2 = \frac{V_2}{R_2} \qquad I_F = \frac{-V_{out}}{R_F}$$

So:

$$\frac{-V_{out}}{R_F} = \frac{V_1}{R_1} + \frac{V_2}{R_2}$$

That is:

$$-\frac{V_{out}}{4.8 \times 10^4} = \frac{0.84}{4.8 \times 10^3} + \frac{0.15}{5.8 \times 10^3} = 2.0 \times 10^{-4}\,\text{A}$$

Then $V_{out} = -(2.0 \times 10^{-4}) \times (4.8 \times 10^4) = -9.6\,\text{V}$

The output is the same as the sum of the outputs that would be obtained if the inputs had been applied individually (each with their own gain, because they have been applied through different input resistances – which could be variable). In this way, the op-amp can be described as a *weighted* summing amplifier.

For example, sound signals of different frequency can be added together with the amplitude of each frequency individually controlled. This type of op-amp circuit can also be used in recording studios (Figure 17.62) to mix sound signals from different sources.

Figure 17.62 A recording studio

Question

1 Confirm that the outputs from each of the two inputs shown in Figure 17.61 would be –8.4 V and –1.2 V if they were connected individually.

C4 The mobile phone system

Figure 17.63 SingTel Ayer Rajah Telecommunications Tower, Republic of Singapore

Mobile phones have become a very important part of life in all parts of the world. Nearly all countries have extensive mobile phone networks (Figure 17.63) and globally the number of users of mobile phones continues to increase. Mobile phone services are relatively cheap and quick to install since they require less fixed infrastructure than traditional telephone systems, which are based on cables and optical fibres.

Base stations and cells

Every mobile phone is capable of emitting and receiving *radio waves*. In order to make a call the phone must be inter-connected with a local **base station** using waves of a particular frequency, although the emitted frequencies are different from the received frequencies. A range of frequencies (*channels*) are allocated to particular mobile phone

C.4.1 **State** that any area is divided into a number of cells (each with its own base station) to which is allocated a range of frequencies.

companies, but there are more mobile phones than radio wave frequencies available. This means that each mobile phone cannot have its own frequency, so that the same frequency must be shared with other phones at the same time (using techniques known as *multiplexing*).

It is desirable to keep the transmitting power of mobile phones as low as possible and therefore they cannot be too far away from a base station. This means that a large number of base stations are needed, and typically they may be about 5 km apart, although this distance varies considerably with the particular circumstances. Having more base stations also has the benefit of reducing the number of users for each one. The area controlled by each base station is called a **cell**. (This is the reason why mobile phones are sometimes called *cell phones*.)

Radio waves between the mobile phone and the base station are in the UHF (Ultra-High radio wave Frequencies) range. Such waves are also widely called *microwaves*. A typical frequency would be about 1 GHz, which corresponds to a wavelength of 30 cm, and this means that the aerial in the phone can be small and still efficiently send and receive signals. Because modern mobile phones are very sensitive to receiving radio waves, and also because the base stations are not required to send signals large distances, the base stations can operate at relatively low power (typically less than 100 W). Note that the words **aerial** and **antenna** are generally accepted to have the same meaning: a device which transmits a radio wave from an oscillating electric current, or vice versa.

■ Additional Perspectives

Aerials and carrier waves

While making a call, a mobile phone continuously produces radio waves as a result of a high-frequency alternating current flowing through its aerial. The waves produced are known as *carrier waves*. Many simple broadcasting transmitters use half-wave (dipole) aerials, which have an overall length equal to one half a wavelength of the radio carrier waves (Figure 17.64).

A standing wave pattern is set up on the aerial, greatly increasing the efficiency of the transmission of radio waves. Figure 17.65 is a 'polar' diagram that represents approximately how the intensity of the radio wave varies with direction from the aerial.

The carrier waves (with, for example, a frequency of 1 GHz) need to *carry* the information about the sound waves made by the caller, which have *much* lower frequencies. This means that some property(s) of the carrier waves must be modified in a way which represents the speech pattern. This could be the amplitude, frequency or phase of the carrier waves, and the process is known as *modulation*.

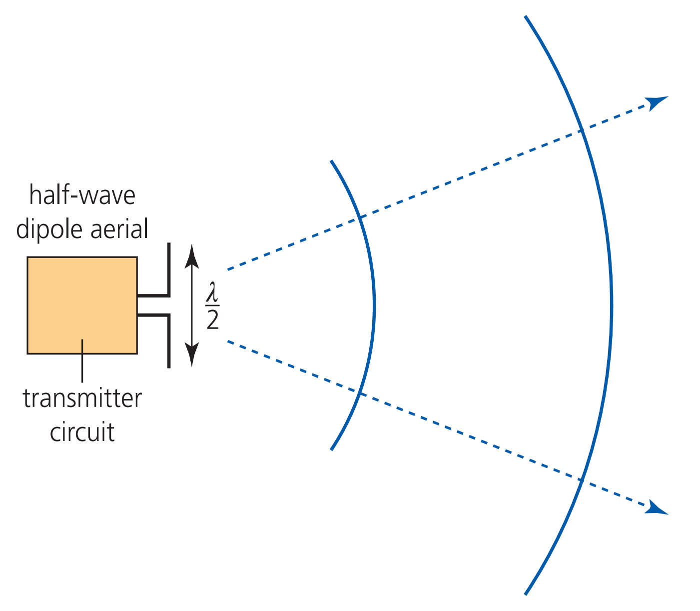


Figure 17.64 A half-wave dipole aerial

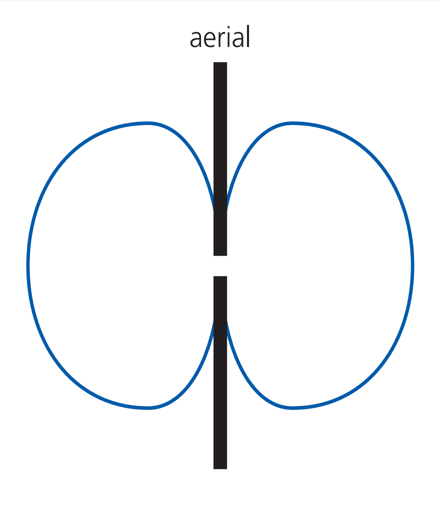


Figure 17.65 Intensity variation with direction from an aerial

Questions

1 Assuming that a simple radio transmitter has an aerial which is about a quarter of a wavelength in length, estimate the frequency of the carrier waves used in a mobile phone.
2 Find out how carrier waves can be modulated in order to carry digital phone signals.

If a base station was near the centre of its cell, with a transmitter which emitted radio waves in all directions, it would provide coverage to an approximately circular area.

The size of a cell around a base station varies, depending upon a number of factors including:

- the power of the transmitter at the base station
- the height of the transmitting aerial above the ground
- the sensitivity of the mobile phones to which it is transmitting
- the shape of the surrounding land and whether there are any hills or large buildings, etc. in the cell. Because like all other waves, radio waves will refract, reflect, diffract, interfere and be absorbed under suitable circumstances. All of these properties can affect how far a signal can be transmitted efficiently
- the number of users within a cell – there is a limit to the number of calls a base station can handle, so that in towns and cities the cells need to be smaller
- the transmitting power of the mobile phones.

A patchwork of overlapping cells of different sizes should be able to cover any given area. Such cellular networks are usually represented in simplified drawings by equally sized hexagons (see Figure 17.66).

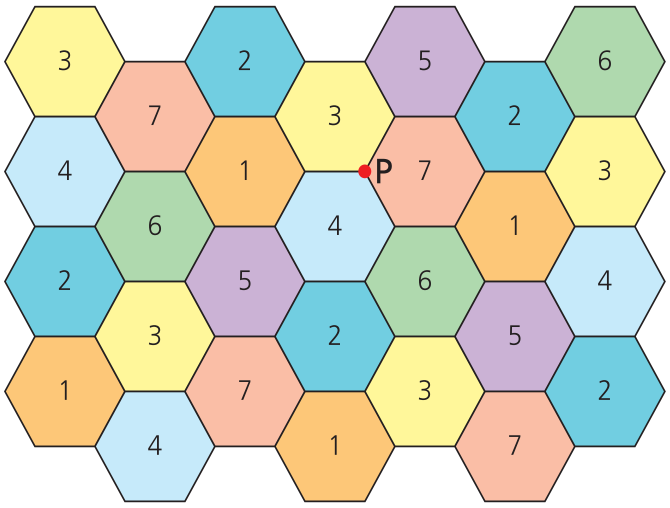


Figure 17.66 An arrangement of cells. Each number indicates that a different range (band) of frequencies is used

After a mobile phone has been switched on it continues to send out radio signals to the nearby base stations of the particular phone company network that it uses. The base station which receives the strongest signal (usually the closest) responds, a link is made and the phone can be used. If the phone is moved to another location where the signal from a different base station is stronger, it will automatically and quickly be switched to that base station, without the user knowing anything about the change. (This is sometimes called *handoff/handover*.) By comparing the strength of the signals received at several different base stations from any particular phone, the location of the mobile phone can also be determined (tracked).

In order for the cellular system to work efficiently, all the cells around any particular individual cell must operate using a different range (band) of frequencies, so that there is no possibility of interference between the waves to or from different base stations. With a hexagonal system this requires seven different frequency ranges. Consider, for example, a cell marked 1 in Figure 17.66: it is surrounded by cells 2 to 7.

In practice however, base stations are often positioned at an *intersection* of three cells, such as point P in Figure 17.66. The base station has three separate aerials, using three different ranges of frequency, pointing to three separate cells (see Figure 17.67). In this way three cells are covered from *one* place. This approach means that fewer locations are needed for base stations (than if one base station with one aerial was at the centre of each cell).

Figure 17.67 A base station with a set of three aerials

■ Additional Perspectives

GSM and UMTS

In the United Kingdom, digital mobile phones use radio frequencies in the ranges of 872–900 MHz and 1710–1875 MHz. Most European mobile phones use GSM (Global System of Mobile communication), which allocates channels to users on a time division multiple access (TDMA) basis. However, GSM is not recognized across the world, so that a phone operating that system cannot be used in all countries. An increasing number of mobile phones will be using UMTS (Universal Mobile Telecommunications System), which is internationally recognized. This uses the higher frequency 2 GHz band, and offers video and multimedia options to mobile users. It uses coding mechanisms to accommodate a larger number of users in a cell than GSM. UMTS supports 3G (*third generation*) mobile phone technology.

Question

1 Find out which system your country uses and draw up a table comparing and contrasting GSM and UMTS.

Additional Perspectives

The mobile phone handset

A block diagram of a mobile phone handset is shown in Figure 17.68. It contains a radio transmitter and receiver. The caller speaks into the microphone. The analogue-to-digital converter (ADC) in the microphone converts the analogue voice sound waves to a digital electrical signal. The parallel-to-series converter takes the whole of each digital number and emits it as a series of binary digits (bits).

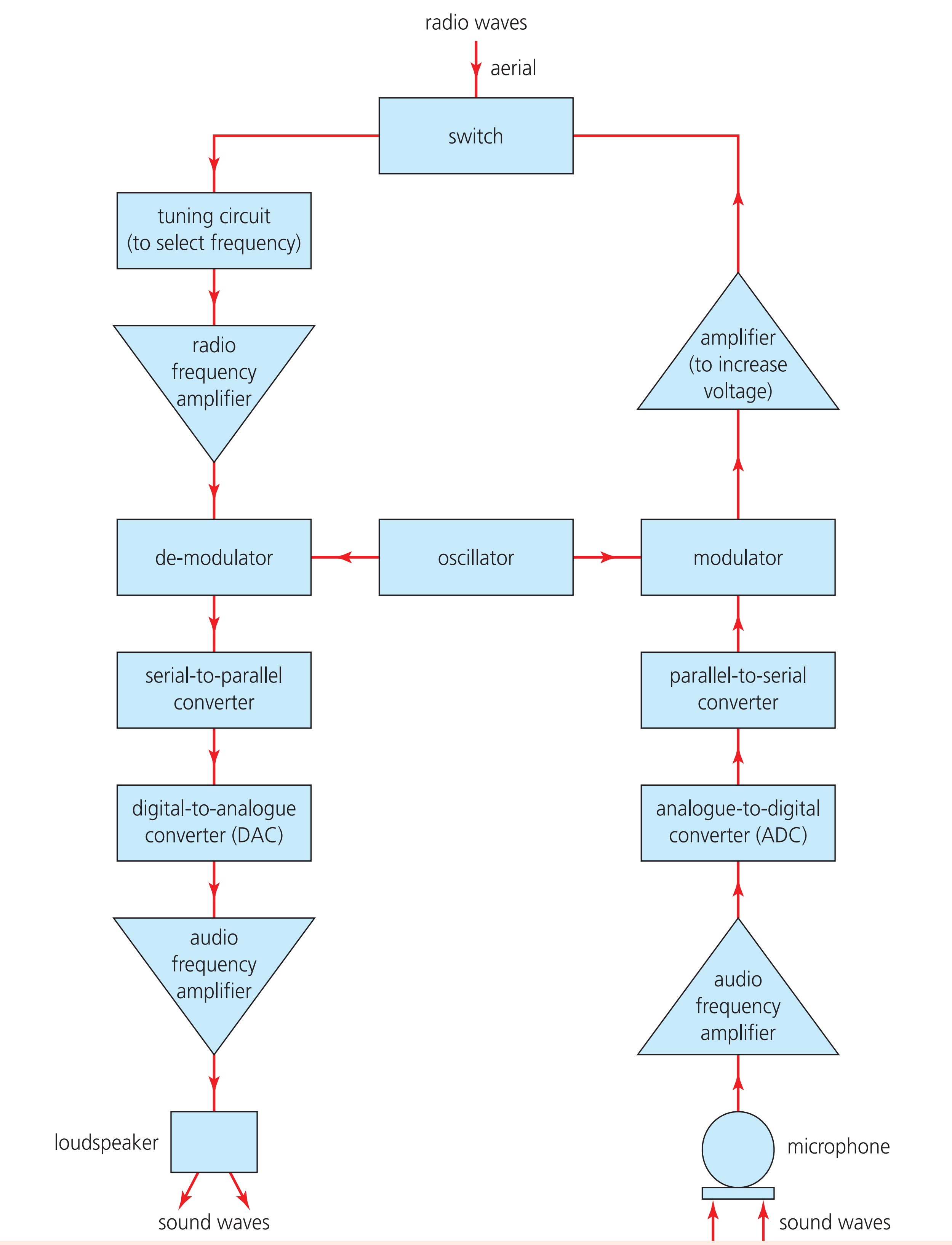


Figure 17.68 Block diagram of a mobile phone handset

The frequency of the oscillator is located by the computer at the cellular exchange. This carrier wave frequency is modulated by a sequence of bits from the parallel-to-series converter. The modulated carrier wave is then amplified and directed to the aerial where it is transmitted as a radio wave.

A signal received at the aerial is switched to a tuning circuit which selects the carrier wave frequency that has been allocated to the handset by the computer at the cellular exchange. The selected signal is amplified by the radio-frequency amplifier. It is demodulated so that the information is in digital form. The series-to-parallel converter allows each sampled digital voltage to be separated and these digital numbers are then converted into an analogue waveform in the digital-to-analogue converter (DAC). Finally, the analogue signal is amplified before the sound is produced in the loudspeaker.

Question

1 Find out how frequency key shifting is used to send data over a mobile phone system.

C.4.2 Describe the role of the cellular exchange and the public switched telephone network (PSTN) in communications using mobile phones.

The cellular exchange and the public switched telephone network (PSTN)

Public switched telephone network

When telephones were first developed towards the end of the 19th century, the connections were made using wires directly between the caller and the receiver. This was only possible because there were not many phones and all the calls were local, for example within a hotel or an office. As the number of phones increased and distances became greater, it soon became necessary to switch (exchange) connections between different phones, so that a *telephone exchange* was needed, as shown in Figure 17.69. The caller would contact the local exchange and the *telephone operator* would then make the electrical connections necessary. For calls over longer distances it would be necessary for a telephone operator at the local exchange to contact another exchange.

In modern systems telephone operators have been replaced by enormous numbers of electronic **relays** (electrically operated switches) that carry out all the switching operations automatically (see Figure 17.70), and it is now possible for (almost) any phone in the world to be automatically connected with any other phone. This vast interconnection of the world's automatic telephone systems is known as the **public switched telephone network (PSTN)**. The PSTN connects phones (fixed line or mobile) using a range of different technologies including copper wires, undersea cables, optical fibres, radio waves and satellites. In order for the PSTN to work efficiently it has been necessary for different countries to agree on common standards and systems including, for example, the way in which telephone numbers are allocated.

Figure 17.69 Telephone operators working at an early exchange

Figure 17.70 A modern automated telephone exchange

Cellular exchange

A cellular exchange controls a group of base stations (using cables) and links them to the PSTN (see Figure 17.71).

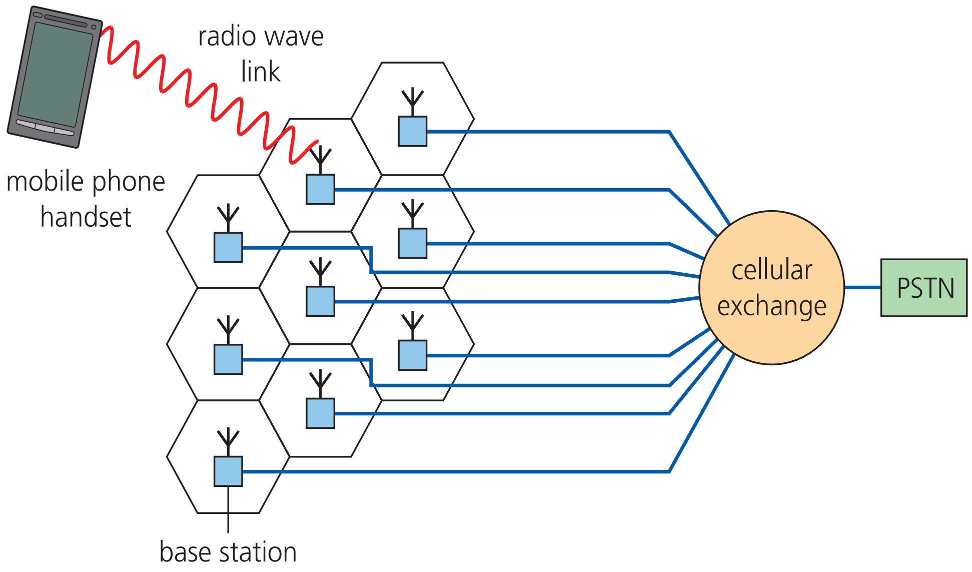


Figure 17.71 Role of the cellular exchange

The cellular exchange performs various functions including;

- allocating an appropriate range of frequencies to each cell, so that neighbouring cells do not share the same frequencies
- selecting the best base station for each mobile phone and, if necessary, rerouting the phone to a different base station with a better connection
- connecting phones to the PSTN.

66 Give reasons why it is desirable to keep the transmitting power of mobile phones as low as possible.

67 Explain why two cells which are next to each other cannot use the same radio frequencies.

68 Make a list of the possible circumstances under which a mobile phone cell could be 20+ km in width.

69 **a** If a base station emits radio waves with a total power of 24 W, estimate the intensity of radiation received at a mobile phone 5.0 km from the station.
b What assumptions did you make in your calculation?
c Explain why it is possible for the transmitting power of the mobile phone back to the base station to be much lower (for example 0.5 W).

70 Look closely at the shape of the individual aerials in Figure 17.67. Suggest a reason why aerials are much taller than they are wide.

71 Suggest possible reasons why **a** the reception of mobile phone signals can be poor, and **b** why the strength of the signal at a particular location can change from minute to minute.

72 Explain how it is possible to keep talking on a mobile phone when travelling on a fast-moving train.

C.4.3 Discuss the use of mobile phones in multimedia communication.

Multimedia communication on mobile phones

As mobile phones have increased in computing power and memory, it has become possible for *smartphones* to continuously access the Internet and have all the facilities of a laptop computer, including the facility to take, send and receive good quality pictures and videos. Such *multimedia* phones require improved bandwidth (high data transfer rates) from the cellular network, which is then generally known as 3G (*third generation*), and with even greater bandwidth, 4G.

Multimedia Messaging Service (MMS) (Figure 17.72) is a standard protocol to send messages with multimedia content between mobile phones. The most popular use is to send photographs from camera-equipped handsets, although it is used to deliver news and entertainment content including videos, pictures, and ringtones.

Some issues arising from the widespread use of mobile phones

C.4.4 Discuss the moral, ethical, economic, environmental and international issues arising from the use of mobile phones.

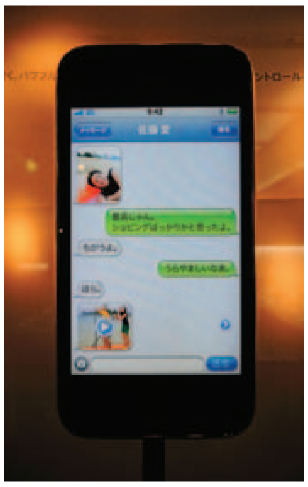


Figure 17.72 MMS on an Apple iPhone®

The use of mobile phones has been a social revolution which has enormously changed the communication habits of the people of the whole world. The benefits are considerable and obvious, but there are some less encouraging aspects which need to be considered as well.

- **Tracking**. Mobile phone tracking refers to determining the location of a mobile phone. This can be done easily using either GPS (if the phone has this installed) or by comparing the strengths of radio signals sent by the phone which arrive at different base stations. This raises privacy issues because many people do not like to think that their movements are being 'tracked'. However, in many countries the police can obtain permission to locate phones in emergency cases where people (including criminals) are missing. On the positive side, users of mobile phones may have the option to upload their phone's location to a common website where friends and family can view their location.
- **Other privacy issues**. It is possible for organizations/governments to monitor any, or all, private calls, texts and access to websites. This may, or may not, be legal under different circumstances in different countries.
- **Communication for everyone**. Because cell phone networks can be set up relatively cheaply in remote areas, it has been possible to introduce mobile phones into areas that have not had communication before, such as rural Africa and India. In developed countries nearly everyone owns a mobile phone, and some people have two, or more.
- **Communication across borders**. International communication has become much easier and it is a simple matter for a person to take their phone and use it in other countries.
- **Safety and security**. If accidents or emergencies happen away from home, a mobile phone can be used to call for help. Children can be contacted by their parents.
- **Expensive new phones**. Mobile phone manufacturers and the providers of mobile phone services have become major international companies, with their products advertised and sold everywhere. Because of this, and because they are used in public places, mobile phones have become status symbols and fashion accessories. Many people want to own the latest, expensive new models, which the companies release and heavily promote every few months.
- **Wasted resources and pollution**. Although mobile phones can still be usable for much longer, in many countries they are used for only an average of less than 18 months before being replaced with a newer model. (This is encouraged by the phone companies offering two-year contracts.) With over 6 billion phones in use around the world, probably at least a billion phones are discarded every year, and only a small percentage of these are recycled. This is a considerable waste of valuable resources, as well as a threat to the environment because of the toxic chemicals they contain.

  Base stations have become a common sight in both town and countryside. Many people do not like their appearance.
- **Health risks**. The latest, extensive scientific research into the health risks from the radio waves used by mobile phones indicates that there is no convincing evidence that they are harmful to human beings. However, such research is limited to the last 20 years or so, and it remains possible that there may be some issues, yet to be discovered, related to the use of mobile phones over very long periods of time. For this reason, some experts still advise that their use should be limited, especially for young children. In fact the most dangerous aspect of mobile phones is probably their use when driving a vehicle. Although this risk is reduced by using hands-free devices, it is not totally eliminated.
- **Inconsiderate use**. There are a large numbers of places where the use of phones is generally agreed to be unacceptable. But in other public locations the inconsiderate use of mobile phones can cause annoyance to other people nearby.
- **Reduced face-to-face communication**. For example, it is a common sight to see a group of people sitting together, but each communicating with someone else using their phone. A received call or text can often appear more important than a face-to-face conversation. The use of text messaging on mobile phones may reduce the amount of time that people spend talking to each other. However, it is also true that text messaging is important in maintaining friendships when direct contact is not possible.

- **Never out of contact**. An expectation has developed that people should be almost instantly contactable, at any time of the day or night. Many people enjoy this, but there are others who would prefer not to be involved and they are increasingly seen as outsiders, possibly leading to social isolation.

73 Choose any one of the mobile phone issues outlined in the previous section and write a few sentences explaining your thoughts on the subject. Do your fellow students share your opinions?

74 Find about the use of microcells, picocells, femtocells and sectored cells. Write a short summary of their uses.

75 Find out how mobile phones and other mobile devices have contributed to the area of mobile health (m-health). Describe some applications of devices in this developing area.

76 What is the role of a SIM card in a mobile phone?

77 Research into the specific environmental problems related to the incorrect disposal of unwanted mobile phones.

78 Find out about the competing network technologies: GSM and CDMA.

SUMMARY OF KNOWLEDGE

C1 Analogue and digital signals

- Information can be stored and transmitted in digital and analogue forms.
- An analogue signal has a large number of different values (between given limits), and hence varies continuously with time.
- A digital signal repeatedly changes between two possible values of 0 (off) and 1 (on).
- Analogue storage devices include the LP and cassette tape.
- In an LP the sound variations are stored as physical variations in a track (groove) on the LP surface.
- In a cassette tape the data is stored as variations in the orientations of magnetic particles on the tape surface.
- Digital storage devices include floppy and hard disks, and optical devices such as CDs and DVDs.
- Floppy and hard disks store data in a series of magnetic variations on the disk surface.
- Digital techniques involve signals consisting of a large number of binary digits, or bits.
- Each binary digit or bit can only take one of two possible values: binary 1 (a specific voltage) or binary 0 (zero voltage).
- Decimal (base 10) numbers (0–9) are represented by counting in powers of 10, so each digit (from right to left) represents 1, 10, 100, 1000, etc.
- Binary (base 2) numbers (0 and 1) are represented by counting in powers of 2, so each digit (from right to left) represents 1, 2, 4, 8, 16, 32, etc.
- In binary notation, the largest power of a series of binary digits is known as the most-significant bit (MSB), and the smallest power is known as the least-significant bit (LSB).
- When the number of bits used is n, the number of different possible values is 2^n.
- One byte is eight bits. 1 KB (one kilobyte) is 2^{10}, or 1024 bytes; 1 MB (one megabyte) is 2^{20}, or 1 048 576 bytes. 1 GB (one gigabyte) is 2^{30} or 1 073 741 824 bytes.
- American Standard Code for Information Interchange (ASCII): eight-bit code representing 256 or 2^8 different possible characters and formatting codes.
- To convert analogue data to digital data the analogue data is sampled at intervals. Each sampled signal is converted into one binary value among a fixed range of possible values/quantum levels.
- To improve the accuracy of the digital data the sampling frequency and the number of available quantum levels can be increased.
- A CD contains a single very long spiral-shaped track that starts in the centre. It is composed of a large number of 'pits' and spaces known as 'lands'.
- The digital information in the CD is read by sensing the amplitude of the reflection of a laser beam (of visible light) reflecting off the 'lands' and 'pits'.

- The speed of rotation of the CD is controlled by an electric motor so that a constant length of the track is scanned by the laser beam in a given period of time.
- The CD has a higher speed of revolution when the laser is reading the track near the centre compared with the outer edge.
- When the laser beam reflects from a 'pit', a strong signal (binary 1) is received and detected.
- When the laser beam reflects from the edge between a 'pit' and a 'land' destructive interference occurs and a suppressed (weak) signal (binary 0) is received and detected.
- An appropriate depth of a 'pit' for the wavelength, λ, of laser light is $\frac{\lambda}{4}$.
- Number of turns on CD = change in radius/distance between spirals on track; length of track = number of turns $\times 2\pi \times$ average radius.
- CD play time = length of track/scanning velocity, and the average length of track per bit of information = length of track/number of bits.
- Pits and lands on a DVD are closer together than on a CD, hence it can store more data. Also, DVDs allows multilayered data storage.
- Characteristics of information stored in digital format: high quality (if sampling rate is high), exact reproducibility, large quantities can be stored in a small device and data can be readily manipulated (changed and copied) without corruption.
- Characteristics of information stored in analogue format: variable quality, reproducing the data introduces noise, analogue storage devices may be large and data may be corrupted if manipulated (changed).

C2 Data capture

- Capacitors are electrical components that can store charge (and energy). The charge stored is proportional to the potential difference applied across the capacitor.
- The simplest capacitor consists of a pair of flat metal plates (separated by an insulating material).
- Capacitance is the charge stored per unit of potential difference, $C = q/V$
- The unit of capacitance is the farad, $1\,\text{F} = 1\,\text{C}\,\text{V}^{-1}$.
- The charge-coupled device (CCD) is a silicon-based device that is used to record an image focused on to its surface. The incident light releases electrons because of the photoelectric effect.
- The surface of the CCD is divided into a large number of small pixels (photodiodes) which behave as tiny capacitors.
- When light is incident on the CCD the photons cause electrons to be stored in each pixel (photodiode); the number of electrons stored depends on the intensity of light (number of photons incident per second).
- The image data is stored as charge on the two-dimensional grid of pixels on the surface of the CCD.
- A potential difference is applied across the CCD to move all the charge down one row. The end row forms a serial register, which is a row of pixels whose potential differences can be measured and recorded. The measured potential difference is proportional to the charge stored.
- An analogue-to-digital converter converts the potential differences into binary data for processing. After measurement, the charge is removed from the rows of pixels and the CCD array is ready to record another image.
- Quantum efficiency is defined as the ratio of the number of photoelectrons emitted to the number of photons incident on the pixel of the CCD.
- The magnification is defined as the ratio of the length of the image on the surface of the CCD to the actual length of the object.
- Two points on an object may be just resolved on a CCD if the images of the points are two pixels apart.
- If the quantum efficiency is high then a high quality image is generated; if the quantum efficiency is low then a poor quality image is generated and some less bright parts of the image will be lost.

- A large CCD means that the magnification can be greater, resulting in a higher resolution and better quality image.
- A higher resolution means a better quality image with a greater amount of detail.
- CCDs are used for image capturing in digital cameras (including mobile phone cameras), digital video cameras, telescopes (including the orbiting Hubble Telescope), scanners, laser printers and imaging of X-rays.
- Advantages of a CCD (in a digital camera) compared to photographic film: lower cost, much higher quantum efficiency, the image can be readily copied, deleted or processed; and storage, viewing and archiving of a large number of images is easy.
- CCDs have been developed that detect photons from the infrared, X-ray and ultraviolet regions of the electromagnetic spectrum.

C3 Electronics

- Electrical signals of low amplitude can be amplified using operational amplifiers (op-amps).
- An operational amplifier has two inputs: the inverting input (V_-) and the non-inverting input (V_+), and one output, V_{out}.
- A voltage applied to the non-inverting input will produce an output of the same sign. A varying signal applied to the non-inverting input will produce an output which is in phase with the input.
- A voltage applied to the inverting input will produce an output of the opposite sign. A varying signal applied to the inverting input will produce an output which is out of phase with the input.
- The output voltage cannot exceed the voltage of the power supply to the op-amp. If the op-amp attempts to produce a greater output, the output will remain constant, and is then said to be saturated.
- An op-amp is described as being in open loop mode if the there is no (feedback) connection made from the output back to either of the inputs.
- Op-amps detect the difference in voltage between the signals applied to their inputs, and then multiply it by some pre-determined open loop gain: $V_{out} = G_0(V_+ - V_-)$
- An ideal op-amp has the following characteristics:
  - infinite gain
  - infinite input resistance, so that the input current is zero
  - zero output resistance.
- The action of connecting the output of an amplifier to one of its inputs is known as using feedback. If this reduces the gain, then the feedback is described as negative. The gain with feedback is called the closed loop gain.
- Although negative feedback reduces the gain in an amplifier, stability is improved and distortion of the output signal is reduced.
- Op-amps can be connected as amplifiers in two basic configurations: inverting and non-inverting.
- The closed loop gain, G, of an ideal inverting amplifier can be shown to be: $-\frac{R_F}{R}$.
- The closed loop gain, G, of an ideal non-inverting amplifier can be shown to be: $1 + \frac{R_F}{R}$.
- A comparator circuit compares the two voltages applied to the inputs of an op-amp. It will give a saturated (high) positive or saturated (low) negative voltage, depending on which voltage is higher. In this way, the comparator gives a digital output from analogue inputs.
- If the non-inverting input is higher, the output voltage will have the same sign as the input. If the inverting input is higher, the output voltage will have the opposite sign to the input.
- Comparator circuits usually have a transducer of some kind (for example an LDR or thermistor) joined in series with a resistor as a potential divider connected to one input. The voltage across the transducer is compared with the voltage provided by another potential divider connected to the other input. The digital output is used to switch a device such as an LED or buzzer on or off.

- Data that is transmitted in digital form can be corrupted and distorted due to noise and dispersion.
- Op-amps are used in circuits called Schmitt triggers to restore distorted pulses to their original shapes.
- The Schmitt trigger switches at different voltages depending upon whether the input is moving from low to high or from high to low.

C4 The mobile phone system

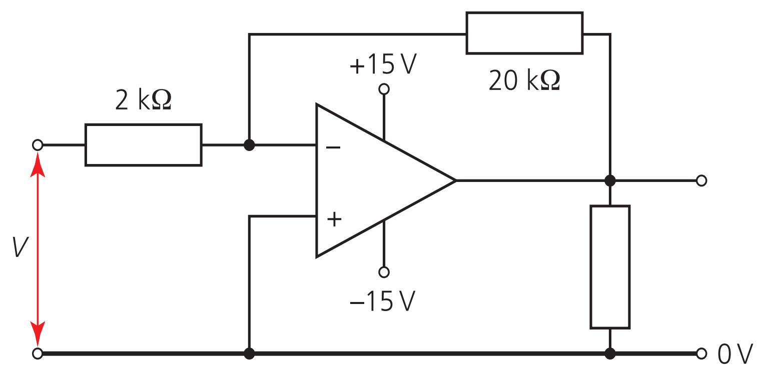
- Mobile phones are low-powered radio receivers and transmitters with a maximum range of about 10 km.
- In a mobile phone network, every region is divided into a number of cells, each with a base station.
- The size of cells in a network depends mainly on geographical features and the population density within each cell.
- The cells overlap so that the mobile phone should always be within range of a base station.
- The mobile phone transmitter sends a radio signal to the base station at one frequency; the base station sends the return signal to the mobile phone receiver at a slightly different frequency.
- Each mobile phone is connected to the public switched telephone network (PSTN) through a base station and a cellular exchange.
- A computer at the cellular exchange selects the carrier wave frequency for each mobile phone and also the best base station to connect with each phone.
- If the phone moves, the cellular exchange can automatically switch the connection to another base station.
- With increased bandwidth, it is possible for mobile phones to be connected to the Internet and to download and play multimedia files (music, picture and video files).
- Despite their enormous advantages, there are a range of social and environmental issues associated with the widespread use of mobile phones.

Examination questions – a selection

All of the IB questions and IB-style questions from Papers 1 and 2 which are to be found at the end of Chapter 14 are suitable for the revision of Option C, although the actual option examination paper (Paper 3) does not contain any multiple-choice type questions.

Paper 3 IB questions and IB-style questions

Q1 The diagram shows an inverting amplifier circuit. Assume that the op-amp is ideal.



a State what is meant by an inverting amplifier circuit. [1]

b **i** State two characteristics of an ideal op-amp. [2]

ii State one advantage of negative feedback in analogue circuits. [1]

c Calculate the voltage gain of this amplifier. [1]

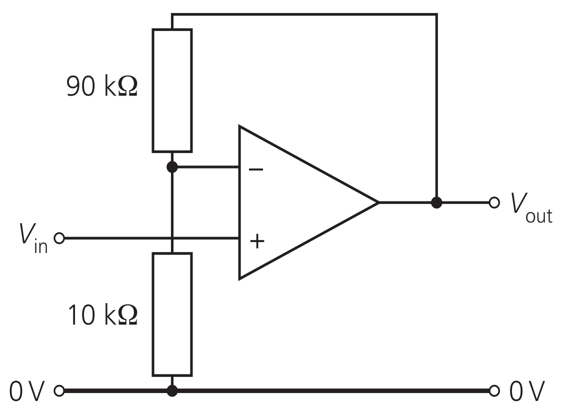
d Determine the current (in μA) which flows through the 2 kΩ resistor at the instant when the input signal voltage is +40 mV. [1]

Q2 **a** Explain the role of the public switched telephone network (PSTN) in making a telephone call. [2]

b List two functions of a cellular exchange in the operation of a mobile phone network. [2]

Q3 This question is about the op-amp.
Diagram 1 shows a non-inverting amplifier circuit.

Diagram 1

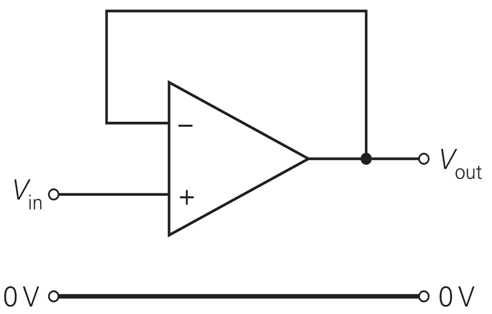


a Suggest why the amplifier is referred to as non-inverting. [1]

b The input voltage for the amplifier in **a** is V_{in} = 2.0 mV.
Calculate:
i the gain G of the amplifier [1]
ii the output voltage V_{out}. [1]

c Diagram 2 shows a particular non-inverting amplifier.

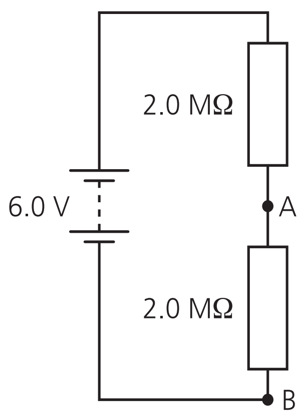
Diagram 2



Explain, in terms of the properties of an op-amp, why the gain of this non-inverting amplifier is equal to 1. [3]

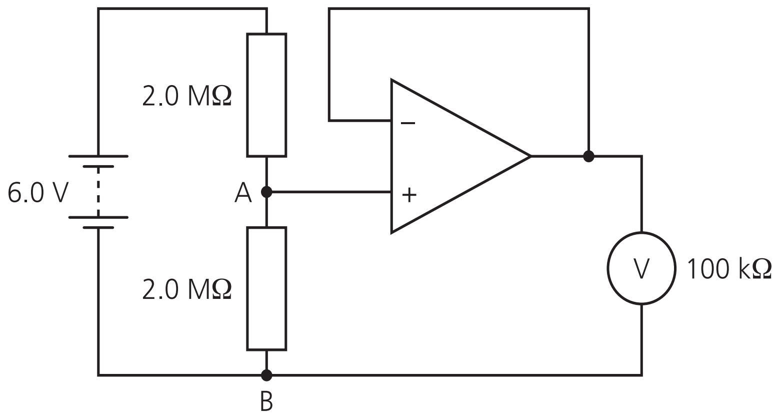
d Diagram 3 shows a circuit in which the battery has an emf of 6.0 V and negligible internal resistance. Two 2.0 MΩ resistors are connected in series to the battery.

Diagram 3



i State the value of the potential difference between points A and B. [1]
ii A voltmeter of resistance 100 kΩ is used to measure the potential difference across points A and B.
State why the reading on the voltmeter is not equal to the value stated in **d i**. [1]
iii The circuit in diagram 3 is modified to include the circuit shown in diagram 2 (diagram 4).

Diagram 4

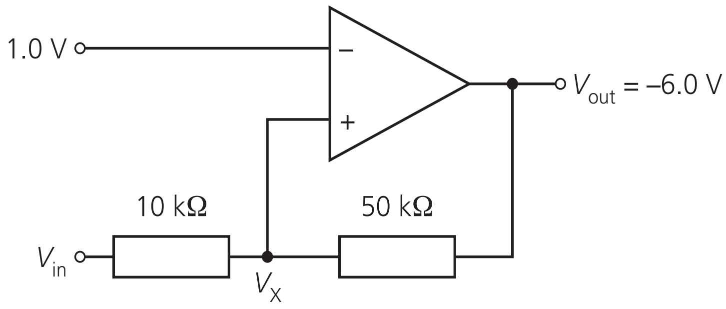


Explain why the voltmeter reads the value of the potential difference as stated in **d i**. [3]

Standard Level Paper 3, May 09 TZ1, QC2

Q4 As a result of noise in electric circuits, digital pulses can often lose their shape and hence distort the information that they carry. The pulses can be re-shaped using a circuit called a Schmitt trigger.

In the situation shown, the output voltage V_{out} of the amplifier is at its minimum value of –6.0 V. The voltage at the non-inverting input to the amplifier is equal to 1.0 V and at the inverting input it is V_X. The output voltage will switch to its maximum value +6.0 V if the voltage V_X just exceeds +1.0 V.



Determine the minimum voltage V_{in} that will result in an output voltage of +6.0 V. [4]

Standard Level Paper 3, Nov 09, QC1 (Part c)

Option
E

Chapter 18 Astrophysics

STARTING POINTS

- Gravity is the only significant force that we need to consider when trying to understand and explain the motion of galaxies, stars, planets, etc.
- When masses are pulled closer together by the force of gravity, there is a transfer from potential energy to kinetic energy as the masses gain speed.
- Temperature is a measure of the average kinetic energy of particles.
- Density = mass/volume.
- A gravitational force which continually acts perpendicular to the direction of motion of a mass can result in a circular orbit if the mass has exactly the right speed, consider $F = mv^2/r$. In practice, orbits will be elliptical (oval) rather than circular. The time for one complete orbit is called the period.
- The force of gravity that acts between two masses reduces with the square of the distance between them. Another example of an inverse square law is the relationship between intensity and distance for radiation spreading out (without absorption) equally in all directions from a point source.
- Charge-coupled devices (CCDs) can be used to measure the intensity of radiation.
- Nuclear fusion is the dominant energy source in stars. When two light nuclei fuse together to make a heavier nucleus a large amount of energy is released.
- An object which would absorb all radiation falling on its surface is called a perfect 'black body'. A perfect black body also emits the maximum amount of radiation possible, although the emitted intensity is different at different wavelengths. Much more radiation is emitted at higher temperatures and the peak value of intensity also moves to shorter wavelengths for hotter objects.
- The Stefan–Boltzmann law for a black body relates the total emitted power to its temperature (K) and surface area: $P = \sigma AT^4$.
- When electrons move between energy levels of atoms they emit or absorb definite sized quanta of energy which are related to the size of the energy transitions involved ($E = hf$). In this way, atoms of each element emit or absorb a range of wavelengths which is unique to that element. Each element therefore can be identified from its emission or absorption spectrum.
- The volume of a sphere can be calculated from $V = (\frac{4}{3})\pi r^3$. The surface area of a sphere can be calculated from $A = 4\pi r^2$.
- Higher level students may also be familiar with the Doppler effect: the shift in wavelength/frequency when there is relative motion between a source and observer.

E1 Introduction to the universe

In the first section of this chapter we will describe in outline what we can see in the night sky and begin to develop an appreciation of the enormous size of the universe that we live in. We will begin with a review of the solar system.

Additional Perspectives

Astrophysics Internet sites

For many students, astrophysics is a fascinating subject, but the opportunities for practical work are obviously limited. However, a considerable amount of very interesting information and stunning images are available on the Internet and, without doubt, it will greatly enhance the study of this topic if the reader has easy and frequent access to the websites of prominent space organizations, such as the European Space Agency (ESA), NASA, Hubble, etc.

The solar system and beyond

E.1.1 Outline the general structure of the solar system.

The **Sun** and all the objects orbiting it are collectively known as the **solar system**. Our Sun is a star and it is very similar to billions of other stars in the universe. It has many objects orbiting around it which are held in their orbits by gravity. The largest of these objects are called **planets**. Most of the planets have one or more objects orbiting around them. These are called **moons**. Mercury and Venus do not have any moons, Mars has two and the Earth has only one, but Jupiter and Saturn have many. The Sun is the only large-scale object in the solar system which emits visible light; the others are only visible because they reflect the Sun's radiation towards Earth.

The Sun was formed about 4.6 billion years ago from the collapse of an enormous cloud of gas and dust. Evidence from radio-isotopes in the Earth's surface suggests that the Earth was formed about 4.5 billion years ago.

Table 18.1 shows some basic data on the eight planets that orbit the Sun. Table 18.2 expresses the same information but by comparison with the Earth.

Table 18.1 Planetary data (all data is correct to two significant figures)

Planet	Mass/10^{24} kg	Radius of planet/10^6 m	Mean distance from Sun/10^{11} m	Period/y
Mercury	0.33	2.4	0.58	0.24
Venus	4.9	6.1	1.1	0.62
Earth	6.0	6.4	1.5	1.0
Mars	0.64	3.4	2.3	1.9
Jupiter	1900	69	7.8	12
Saturn	570	57	14	29
Uranus	87	25	29	84
Neptune	100	25	45	160

Table 18.2 Comparing data on planets with Earth

Planet	Relative mass	Relative radius	Relative mean distance from Sun	Relative period
Mercury	0.055	0.38	0.39	0.24
Venus	0.82	0.95	0.72	0.62
Earth	1.0	1.0	1.0	1.0
Mars	0.11	0.53	1.5	1.9
Jupiter	320	11	5.2	12
Saturn	95	9.0	9.6	29
Uranus	15	4.0	19	84
Neptune	17	3.9	30	160

The distances given in Table 18.1 are only averages because the planets are not perfect spheres and because their orbits are **elliptical** (oval), rather than circular. The Earth's orbit, however, is very close to being circular so the Earth is always about the same distance from the Sun. (The Earth is closest to the Sun in January but there is only about a 3% difference between the smallest and largest separations). An ellipse has two **focuses** (foci) and the Sun is located at one of those two points. The **period** of the Earth's orbit is, of course, one year, but note that the further a planet is from the Sun, the longer its period.

The inner planets (Mercury, Venus, Earth and Mars) are solid, but the outer planets (Jupiter, Saturn, Uranus and Neptune) are mostly gaseous. Mercury, Venus, Mars, Jupiter and Saturn may be seen from Earth with the unaided ('naked') eye, depending on where they are in their orbits around the Sun. A telescope is needed to observe the other planets.

1 a Calculate the average density of Earth and Jupiter.
b Why are they so different?

2 a What is the average orbital speed of the Earth?
b Compare the Earth's speed to that of Mercury.

3 a If there was a planet located at 35×10^{11} m from the Sun, suggest how long it might take to complete its orbit.
b Would such a planet be visible to the unaided eye? Explain your answer.

4 a What is the smallest planet and what is its mass?
b Why is Pluto not considered to be a planet?

5 What is the largest planet and what is its diameter?

Additional Perspectives

Kepler's laws of planetary motion

Johannes Kepler first stated his three famous laws early in the 17th century. He was a German mathematician working with the detailed records of the famous Danish astronomer, Tycho Brahe. As such, his laws were completely **empirical** (based only on observation) and at that time there was no known explanation of them.

- **First law**: all planets move in elliptical paths with the Sun at one focus. (Before that, it was generally assumed that planetary orbits were circular.)
- **Second law**: a line drawn from a planet to the Sun 'sweeps out' equal areas in equal times during the orbit. (This was a precise way of expressing the more general statement that planets travel more quickly when they are closer to the Sun, see Figure 18.1.)
- **Third law**: the square of the time period of a planet is proportional to its average radius cubed ($T^2 \propto R^3$).

About 100 years later, Isaac Newton was able to use his theory of universal gravitation to explain *why* these empirical laws were true.
(This has been covered in Chapter 9 for Higher Level students.)

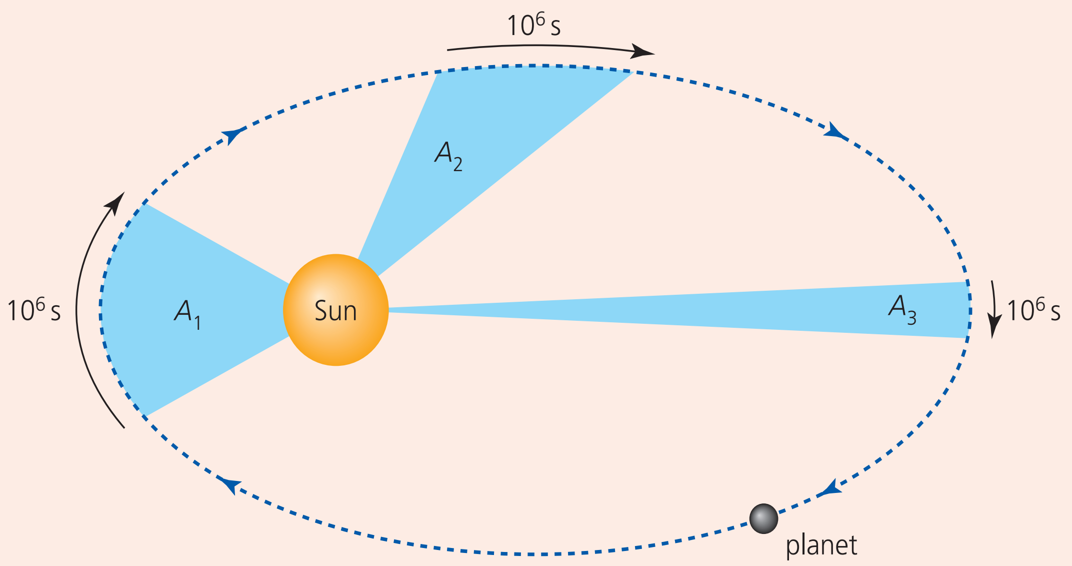


Figure 18.1 A planet sweeps out equal areas in equal times as it orbits the Sun ($A_1 = A_2 = A_3$)

Questions

1 How far from the Sun would a planet have to be in order that the time to complete one orbit was exactly 20 years?
2 Suggest why the time period of a planet in orbit around the Sun does not seem to depend on its mass.
3 Use the data in Table 18.2 to verify Kepler's third law **a** by calculation and **b** graphically.

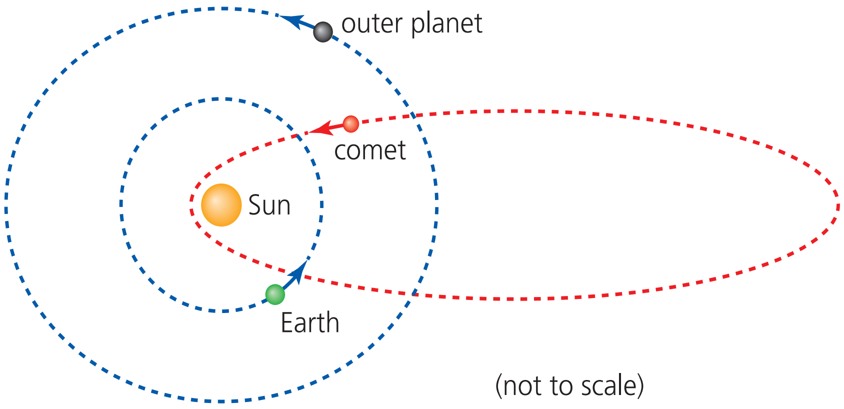


Figure 18.2 The eccentric ('flattened') path of a comet

Comets are relatively small lumps of rock and ice which also orbit the Sun, but typically with very long periods and very elliptical paths (see Figure 18.2). They therefore spend relatively little of their time in the inner solar system close to the Sun and the inner planets, such as Earth. When they approach the Sun, radiation and the outflow of particles (the solar wind) often cause a comet to develop a diffuse *tail* of dust and gas, which always points away from the Sun. This, together with the rarity of seeing them, has made comets a matter of great curiosity for many of the world's civilizations. Probably the most famous comet is named after the British astronomer and mathematician Edmund Halley (1656–1742). Halley correctly predicted this comet would next be seen in 1758 (which was 16 years after his death). Halley's' comet has a period of 75 years; it was last seen in 1986 and will be seen next in the year 2061.

Asteroids are large rocks which are generally bigger than comets but much smaller than the planets. They do not have 'tails' and most orbit the Sun in approximately circular orbits between Mars and Jupiter, in a zone called the **asteroid belt**.

Because they are relatively small, the *trajectories* (paths) of asteroids and comets may be significantly altered if they pass 'close' to a planet (especially Jupiter), when they are subject to large gravitational forces.

Figure 18.3 A comet and its tail

Additional Perspectives

Asteroids colliding with the Earth

Science-fiction authors and movie makers enjoy frightening us all with stories about asteroids or comets colliding with the Earth, but it is only in recent years that scientists have come to realize that such a major collision is not as unlikely as they had previously thought. In 1994 a large comet (Shoemaker–Levy 9) collided with Jupiter and the effect of the impact was easily seen with telescopes and was broadcast around the world on television (Figure 18.4). If a similar comet collided with Earth, the results would be catastrophic, although not quite on a scale comparable to the asteroid collision with Earth about 65 million years ago, which is thought to have led to the extinction of many species, including the dinosaurs.

Figure 18.4 Astronomers watch the impact of comet Shoemaker–Levy 9 with Jupiter

We only need to look at the crater-covered surface of the Moon to become aware of the effects of collisions with asteroids and comets, but similar evidence is not so easy to find on the Earth's surface. Rocks of diameter 10 m or less usually break up in the Earth's atmosphere before impacting, so an asteroid would need to have a diameter of about 50 m or more before its impact would leave a noticeable and long-lasting crater. The effects of friction with the air might also cause an asteroid to explode before it impacted the Earth's surface. Of course, most of the Earth is covered with water and no craters would be formed after an impact with the oceans. Also, old craters may well have been eroded, weathered or just covered with vegetation over long periods of time.

Actual estimates about the size of possible asteroids which could collide with Earth and the probability of such events occurring are continually being refined. But, in general terms, we know that the probability of the Earth being struck by an asteroid is inversely related to its size. An asteroid 50 m in diameter may impact the Earth about every 1000 years; a 1 km asteroid about every 500 000 years and a 10 km asteroid once every 100 000 000 years. The chance of a catastrophic impact in an average human lifetime may be about 1 in 10 000.

There may be up to a million asteroids in our solar system capable of destroying civilization if they impacted with Earth, but it is not easy to observe many of them, nor track their movements. Much effort is now going into Near Earth Asteroid Tracking and researching what might be done if a dangerous impact was expected.

Questions

1 Calculate the kinetic energy of an asteroid of diameter 1 km and average density 4000 $kg\,m^{-3}$ travelling at a speed of 20 $km\,s^{-1}$. Compare your answer with 25 megatonnes of TNT, the energy that would be released from a 'large' nuclear bomb. (1 tonne of TNT is equivalent to 4.2×10^9 J.)

2 Use the Internet to find out when the next large asteroid is expected to pass near to Earth. How close will it come and what how dangerous would it be if it hit us?

Stars and galaxies

All stars appear to the unaided eye as points of whitish light, but there are very large differences in brightness between them. Indeed, most stars are so dim that we cannot see them without the help of telescopes, if we can see them at all.

Ancient astronomers believed that the stars were all equal distances away from Earth on the surface of one or more vast *celestial spheres* around the Earth. We now know that they are distributed throughout three-dimensional space.

It is very important to realize that the brightness of any star that we observe depends on how far away it is *and* the amount of light it emits. So, we cannot just assume that dimmer stars are further away. By simple direct observation we cannot know the distance between us and any particular star, or the distance between the stars. Stars which seem to be close together in the sky may indeed have a much greater distance between them than stars that appear to human eyes on Earth to be much further apart.

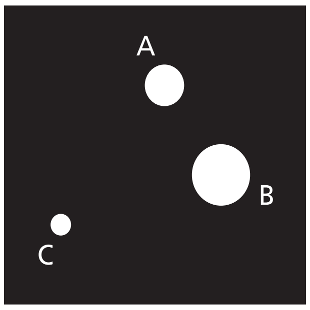


Figure 18.5 The relative brightnesses of three stars (indicated by the diameters of the dots)

In Figure 18.5, stars A and B appear to be close together, but in reality, in three dimensional space, star A could be much closer to star C than star B. The situation may be further confused by differences in the brightness of the three stars. For example, it is perfectly possible that star B could be the furthest away of these three stars and only appears brightest because it emits much more light than the other two.

When we look at the stars in the night sky, they seem to be distributed almost randomly but we are only looking at a tiny part (within our own galaxy) of an enormous universe. The force of gravity causes billions of stars to collect into groups, all orbiting the same centre of mass. These groups are known as **galaxies** (Figure 18.6). Some of the spots of light we see in the night sky are distant galaxies (rather than individual stars). Billions of galaxies have been observed using astronomical telescopes.

We, the Sun and all the other stars that we can see with the unaided eye are in a galaxy called the **Milky Way**.

Figure 18.6 Billions of stars in an elliptical galaxy

Clusters and constellations

E.1.2 Distinguish between a stellar cluster and a constellation.

Within a galaxy, groups of stars may be identified which are relatively close to each other and which move together because of the gravitational forces between them. These stars are all about the same age and origin. They are known as **stellar clusters** and the number of stars they contain can vary widely.

Clusters should not be confused with **constellations**. Ancient societies, such as Chinese, Indian and Greek civilizations, attempted to see some order in the apparent random scattering of the stars that we can see from Earth. They identified different parts of the night sky by distinguishing patterns of stars representing some aspect of their culture, such as the Greek hunter, Orion (see Figure 18.7). These *patterns* of visible stars are called constellations, although it should be clearly understood that the stars within any given constellation do not necessarily have anything in common and they are probably not even 'close together', despite the impression we have by viewing them from Earth. Although many constellations were first named thousands of years ago, their names are still widely used today to identify parts of the night sky.

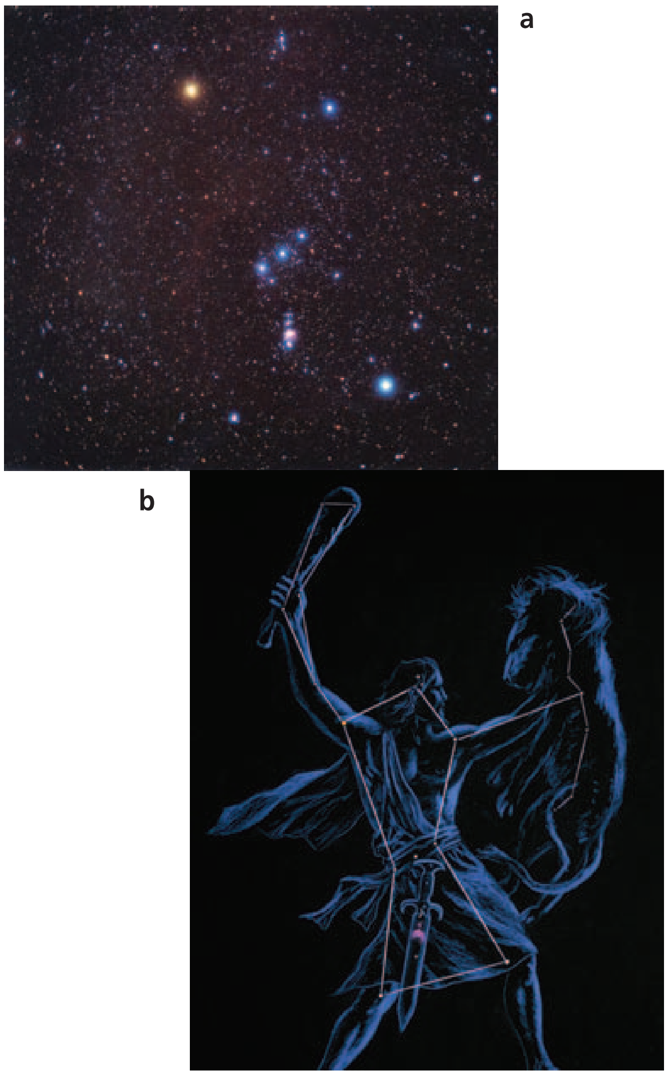


Figure 18.7 The constellation of Orion: **a** the stars in the sky **b** a representation from mythology

Units for distance used in astronomy

E.1.3 Define the *light year*.

The universe is enormous! Rather than use metres to measure distance, astronomers usually prefer to deal with smaller numbers and have introduced alternative units for distance.

- The **light year, ly**, is defined as the *distance* travelled by light in vacuum in one year at a speed of $3.00 \times 10^8\,\text{m}\,\text{s}^{-1}$.

- A light year is easily shown to be 9.46×10^{15} m; this value is provided in the IB *Physics data booklet*.
- The **astronomical unit**, **AU**, is equal to the mean distance between the Earth and the Sun, 1.50×10^{11} m.
- 1 **parsec**, **pc**, is equal to 3.26 ly (defined later in section E.3).

Worked example

1 a A star is 5.4 ly from Earth. How far away is that in kilometres?

b What is the distance from the Earth to the Sun in light years?

a $5.4 \times (9.46 \times 10^{15}) = 5.1 \times 10^{16}\,\text{m} = 5.1 \times 10^{13}\,\text{km}$

b $1.5 \times 10^{11} / 9.46 \times 10^{15} = 1.6 \times 10^{-5}\,\text{ly}$

The scale of the universe

E.1.4 Compare the relative distances between stars within a galaxy and between galaxies, in terms of order of magnitude.

The diameter of the **observable universe** is about 9×10^{10} ly. The speed of light limits the amount of the universe which we can, in principle, 'observe'. The distance to the edge of the observable universe is equal to the speed of light multiplied by the age of the universe (but the expansion of space itself must be considered, which will be discussed later in the chapter).

Distances between stars and between galaxies vary considerably. As a *very approximate* guide there might be 10^{12} stars in a big galaxy and a typical separation of stars within it may be about 1 ly, with a typical total diameter of a galaxy being about 10^4 ly (Figure 18.8). The billions of galaxies are separated from each other by vast distances, maybe 10^7 ly or more.

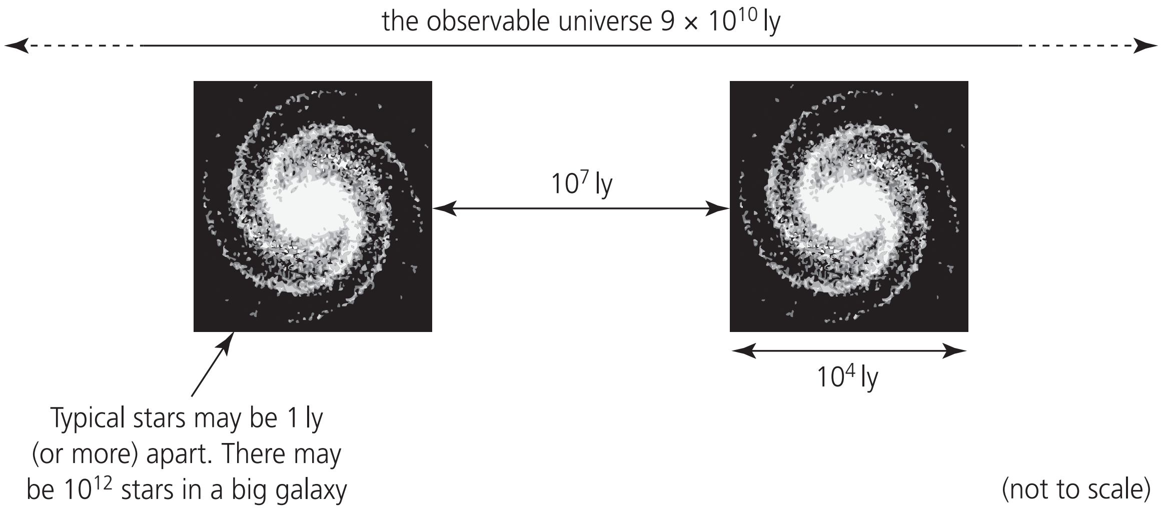


Figure 18.8 Approximate dimensions of galaxies

6 What is the approximate size of the observable universe in:
a km
b pc?

7 Proxima Centauri is the nearest star to Earth at a distance of 4.0×10^{16} m.
a How many light years is this?
b If the Earth was scaled down from a diameter of 1.3×10^7 m to the size of a pin head (1 mm diameter), how far away would this star be on the same scale?

8 Our solar system has an approximate size of at least 10^{11} km.
a How many light years is that?
b If you were making a model of our solar system using a ball of diameter 10 cm to represent the Sun, how far away would the 'edge' of the solar system be? (Sun's diameter = 1.4×10^6 km.)
c Research into how the edge of the solar system can be defined and what objects in the solar system are the most distant from the Sun.

9 Calculate the time for light to reach Earth from the Sun.

10 **a** Estimate how long would it take a spacecraft travelling away from Earth at an average speed of 4 km s^{-1} to reach:
i Mars **ii** Proxima Centauri.
b Find out the highest recorded speed of a spacecraft.

11 Use the data in Figure 18.8 to make a very rough estimate of the number of stars in the observable universe.

Observing the night sky

E.1.5 **Describe** the apparent motion of stars/constellations over a period of a night and over the period of a year, and **explain** these observations in terms of the rotation and revolution of the Earth.

On a clear night, far away from the light pollution of towns, it may be possible to see a few thousand stars in the night sky with the unaided eye. A total of about 5000 stars are visible from Earth with the human eye, but not all can be seen at the same time or from the same place. What we can see depends on our location, the time of the night and the time of the year. This variation is because of the Earth's motion – its spin on its axis and its orbit around the Sun. At any one time in any one place we might be able to see about half of the observable stars.

Stars seem to stay in exactly the same positions/patterns (relative to other stars) over thousands of years and therefore we can locate the stars precisely on a *star map*, such as shown in Figure 18.9 (overleaf). Although stars are moving very fast, their motion is not noticeable from Earth, even over very long periods of time (in human terms), because they are such enormous distances away from us.

If we observe the stars over a period of hours on any one night we will notice that they appear to move across the sky from east to west – in exactly the same way as the Sun appears to move during the day. These apparent motions are actually produced because the Earth spins in the opposite direction. Time-lapse photography can be used to show the paths of stars across the sky during the night. Such photographs can even show the complete circular path of stars which are close to the Earth's extended axis (Figure 18.10).

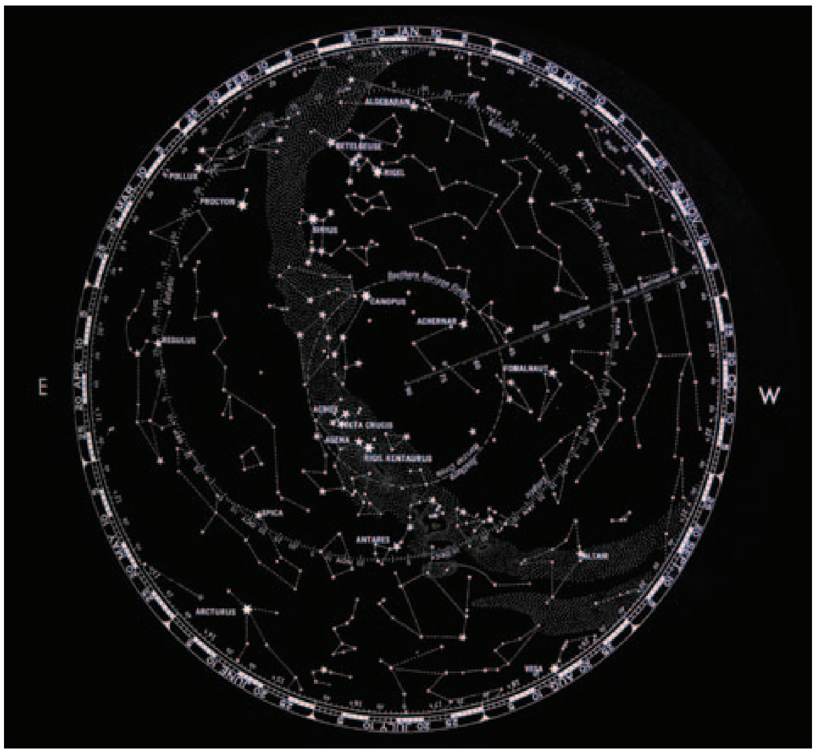


Figure 18.9 A star map for the Southern hemisphere

Figure 18.10 The apparent rotation of the stars as the Earth spins

In the course of one day, the Earth's rotation causes our 'view' of the stars to revolve through 360°, but, of course, during day time we are not able to see the stars because of the light from the Sun. (Radio astronomers do not have this problem.) Our night-time view changes slightly from one night to the next, and after six months we are looking in exactly the opposite direction, as shown in Figure 18.11.

The Sun, the Moon and the five planets which are visible with the unaided eye are all much, much closer to Earth than the stars. Their movements as seen from Earth can seem more complicated and they cannot be located in fixed positions on a star map. The Sun, the Earth, the Moon and the planets all move in approximately the same plane. This means that they follow similar paths across the sky as seen by us as the Earth rotates.

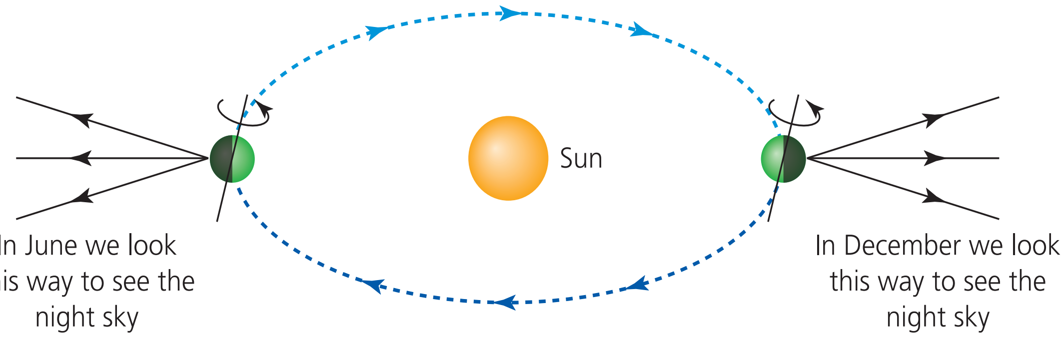


Figure 18.11 How our view of the night sky changes during the year

The Sun and the Moon are the biggest and brightest objects seen in the sky. In comparison, the stars appear only as points of light. The closest planets may just appear as discs (rather than points) of light, especially Venus – which is the brightest natural object in the night sky (other than the Moon).

There are a few other things we might see in the night sky. At certain times, if we are lucky, we may also be able to see a comet, an *artificial satellite*, or a *meteor* – which causes the streak of light seen in the sky when a rock fragment enters the Earth's atmosphere and burns up due to friction. Occasionally, parts of meteors are not completely vaporized and they reach the Earth's surface. They are then called *meteorites* and are extremely valuable for scientific research, being a source of extra-terrestrial material.

E2 Stellar radiation and stellar types

We will now consider stars in more detail, the nature of the radiation they emit and the source of that energy.

All stars, including our Sun, have a great deal in common, although there can be significant differences in size, power and age. There can also be slight differences in their colour. Understanding these differences and how they are interconnected enables astronomers to sort all the billions of the stars into a few basic types.

Energy source

E.2.1 **State** that fusion is the main energy source of stars.
E.2.2 **Explain** that, in a stable star (for example, our Sun), there is an equilibrium between radiation pressure and gravitational pressure.

All the radiated energy received on Earth from the Sun has been released originally from the Sun's interior by the fusion of hydrogen to form helium.

Nuclear fusion happens in all stars (until near the end of their 'lifetimes') and is the dominant energy transformation in stars.

This process can be simplified to the following nuclear equation:

$$4\,^1_1\text{H} \rightarrow \,^4_2\text{He} + 2\,^0_1\text{e} + \text{other particles/radiations (neutrinos and photons)}$$

Each completed nuclear fusion of helium from four hydrogen nuclei (protons) is accompanied by a decrease in mass and an equivalent release of energy amounting to 27 MeV (for more detail, see the Additional Perspective on page 690). The fusion of heavier elements occurs later in the lifetime of stars.

Over a very long period of time, gravity has pulled the protons closer together until they have gained very high kinetic energies (that is, the temperature is extremely high – millions of kelvin). The protons have so much kinetic energy that they can overcome the very high *electrostatic forces of repulsion* between them and fuse together to make helium. When this happens on the large scale it is commonly described as the *birth* of a star.

The high temperatures create a **thermal gas pressure** and the emitted radiation also creates a **radiation pressure** outwards in opposition to the **gravitational pressure** inwards. These pressures will remain equal and opposite for a very long time, during which the star will remain the same size, stable and unchanging – that is, it will be in **stellar equilibrium** (Figure 18.12). There is also a balance between energy transferred from fusions and energy radiated from the surface. It may be helpful to compare this to a balloon in equilibrium under the action of the gas pressure outwards and the pull of the elastic inwards. During this period the star is known as a **main sequence** star. Eventually the supply of hydrogen will be used up and the star will no longer be in equilibrium. This will be the beginning of the end of its 'lifetime' and the star will eventually 'die'. Our Sun is approximately halfway through its lifetime as a main sequence star.

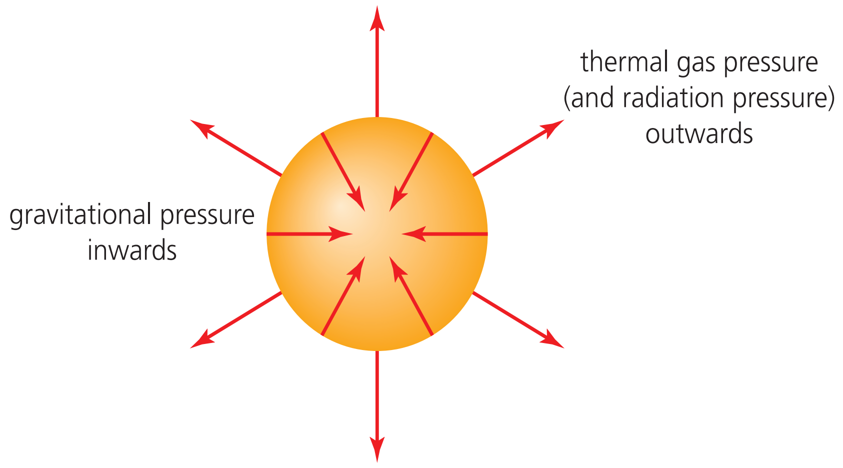


Figure 18.12 A stable main sequence star in equilibrium

Additional Perspectives

More about fusion in stars

Figure 18.13 shows in more detail the so called 'proton–proton' chain in which six hydrogen nuclei (protons) combine in three steps to form one helium nucleus and two new hydrogen nuclei (protons).

The nuclear equations for these three steps can be written as follows:

1 ${}^1_1H + {}^1_1H \rightarrow {}^2_1H + {}^0_1e + \nu$

2 ${}^2_1H + {}^1_1H \rightarrow {}^3_2H + \gamma$

3 ${}^3_2H + {}^3_2H \rightarrow {}^4_2He + 2{}^1_1H$

Each of the three fusions transfers energy, as discussed in Chapter 7, with the third step releasing about half of the total energy. Comparison of total nuclear binding energies before and after these nuclear reactions shows that the energy transferred in the three-step reaction is 27 MeV.

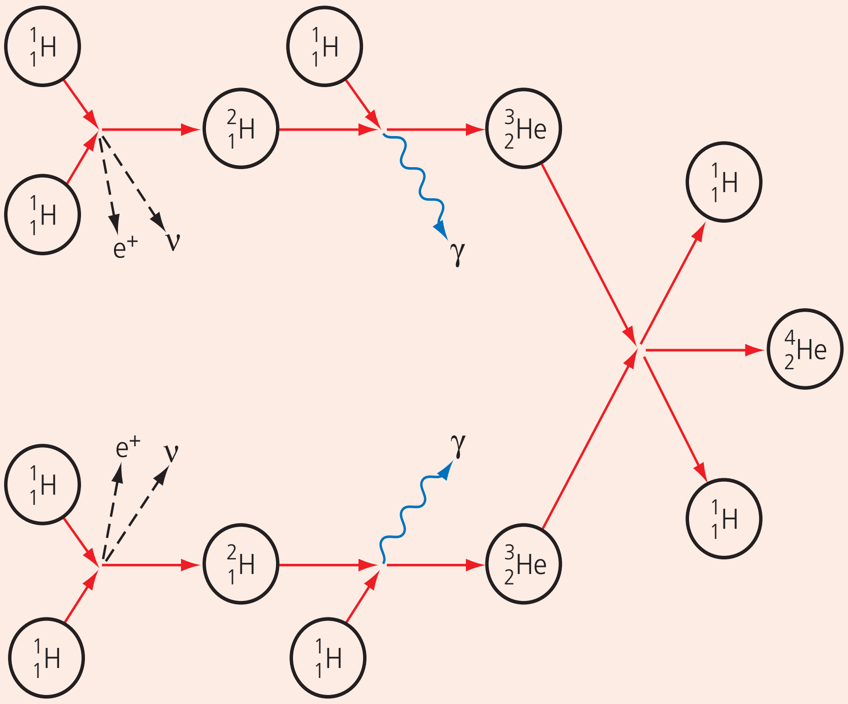


Figure 18.13 Nuclear fusion of hydrogen to form helium

Photons and neutrinos are emitted from the Sun, and the positrons produced in the reactions annihilate with electrons to produce more gamma ray photons.

Question

1 Convert 27 MeV to joules. Then, using $E = mc^2$, show that each nuclear reaction involves a decrease in mass of about 5×10^{-29} kg. Given that the power of our Sun is 4×10^{26} W, calculate the loss of mass every second.

Luminosity

E.2.3 Define the *luminosity* of a star.
E.2.4 Define *apparent brightness* and state how it is measured.

The **luminosity**, ***L***, of a star is defined as the total power it radiates (in the form of electromagnetic waves). It is measured in watts, W.

For example, the luminosity of the Sun is 3.8×10^{26} W.

We would reasonably expect that the energy from any star spreads out equally in all directions, so the power arriving at a distant observer on Earth will be very considerably less than the power emitted.

The **apparent brightness**, ***b***, of a star (including the Sun) is defined as the *intensity* (power/receiving area) on Earth. The units are $W\,m^{-2}$.

The apparent brightness of the Sun is approximately 1400 $W\,m^{-2}$ above the Earth's atmosphere. This is also called the solar constant, which was discussed in Chapter 8.

The apparent brightness of a star will depend on its luminosity *and* its distance from Earth.

The apparent brightness of stars is a vital piece of information about them. Although the amount of intensity received from the stars will often be very low (a typical value could be $10^{-13}\,W\,m^{-2}$), astronomers have developed very accurate means of measuring apparent brightnesses using charge-coupled devices (CCDs), in which the charge produced in a semiconductor is proportional to the number of photons received and hence the apparent brightness. (CCDs are covered for Higher Level students in Chapter 14.)

Assuming that none of the emitted energy is absorbed or scattered as it travels across space, the power received per square metre anywhere on a sphere of radius d will be equal to the emitted power (luminosity) divided by the 'surface' area of the sphere, as shown in Figure 18.14.

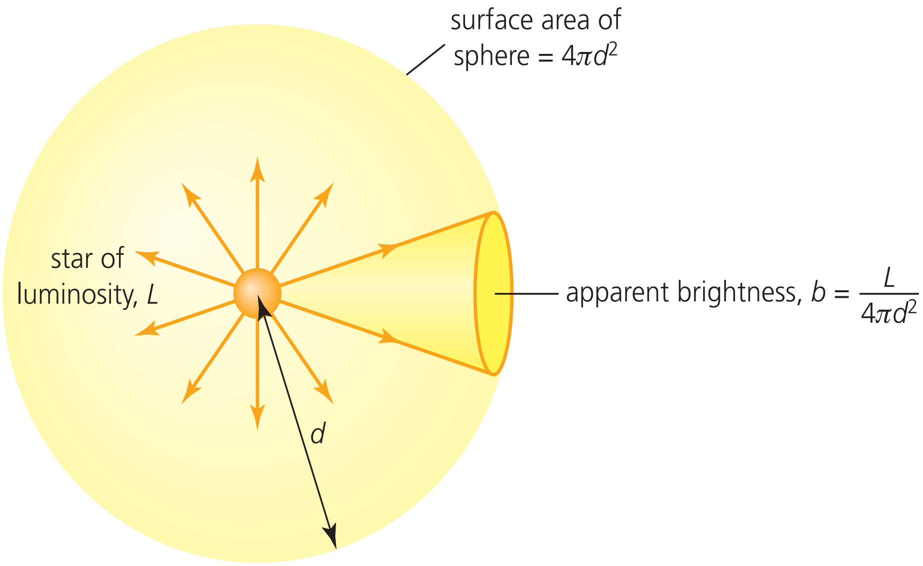


Figure 18.14 Relating apparent brightness to luminosity

$$\text{apparent brightness, } b = \frac{L}{4\pi d^2}$$

This equation is given in the IB *Physics data booklet*.

This is another example of an *inverse square relationship*. If the distance from a star is multiplied by 2, then the apparent brightness is divided by 2^2; if the distance is multiplied by, for example, 37, then the apparent brightness will be divided by 37^2 (= 1369), etc. This is illustrated by Figure 18.15, which shows that, for example, at three times the distance, the same power is spread over nine (3^2) times the area.

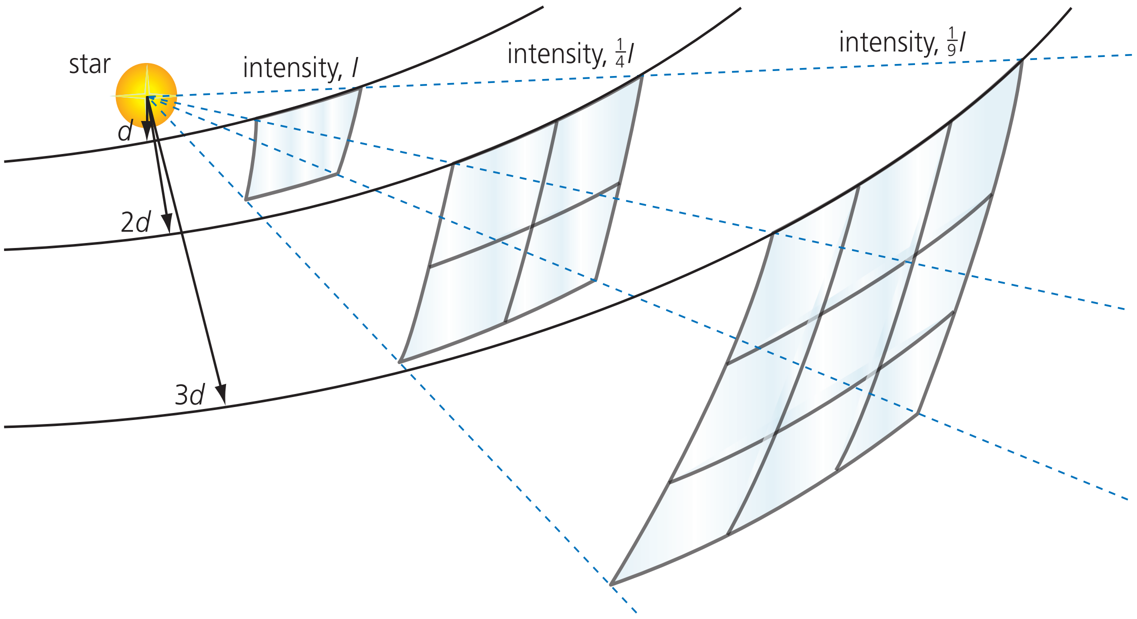


Figure 18.15 How intensity changes with the inverse square law

The importance of this equation lies in the fact that once we have measured the apparent brightness of a star and we know its distance from Earth, then it is a simple matter to calculate the luminosity of the star. However, accurately determining the distance from Earth to a star is not easy.

Not surprisingly, very little radiation is absorbed or scattered as it travels billions of kilometres through almost empty space, although the effects of the journey should be considered when studying the most distant galaxies. However, 100 km of the Earth's atmosphere does have a very significant effect, reducing brightness and resolution in many parts of the spectrum. That is why astronomers often prefer to use telescopes placed on mountain tops, or on satellites above the Earth's atmosphere to gather data.

Worked example

2 A star of luminosity 6.3×10^{27} W is 7.9×10^{13} km from Earth. What is its apparent brightness?

$$b = \frac{L}{4\pi d^2}$$

$$b = \frac{6.3 \times 10^{27}}{4\pi \times (7.9 \times 10^{13} \times 10^3)^2}$$

$$b = 8.0 \times 10^{-8}\,\mathrm{W\,m^{-2}}$$

12 How far away from Earth is a star which has a luminosity of 2.1×10^{28} W, and an apparent brightness of $1.4 \times 10^{-8}\,\mathrm{W\,m^{-2}}$?

13 A star that is 12.4 ly from Earth has an apparent brightness of $2.2 \times 10^{-8}\,\mathrm{W\,m^{-2}}$. What is its luminosity?

14 Calculate the distance to the Sun using values for its luminosity and apparent brightness.

15 Star A is 14 ly away from Earth and star B 70 ly away. If the apparent brightness of A is 3200 times greater than that of star B, calculate the ratio of their luminosities.

16 If the radiation from the star in question **12** has an average visible wavelength of 5.5×10^{-7} m, estimate how many visible photons arrive every second at a human eye of pupil diameter 0.50 cm.

Additional Perspectives

Telescopes on the ground and telescopes in orbit

Waves from all parts of the electromagnetic spectrum arrive at the Earth from outer space and it is truly impressive to consider just how much scientists have learned about the universe from studying these various radiations. Most of this option is about how that information is interpreted, but little has been included about how waves from the various parts of the electromagnetic spectrum provide different information about their sources. Figure 18.16 shows a telescope designed to focus and detect radio waves from outer space.

When radiation passes through the Earth's atmosphere some of it may be absorbed, refracted or scattered, and these effects will often depend on the wavelengths involved. For example, in visible light, the blue end of the spectrum is scattered more than red light, and that helps to explain blue skies and red sunsets. We only have to look through the shifting haze above a hot surface to appreciate just how much the convection currents in the air affect what we see.

Astronomers have long understood the advantages of placing optical telescopes on the tops of mountains to reduce the adverse effects of the atmosphere on the images seen (Figure 18.17). The highest mountains are, of course, much lower than the height of the atmosphere, which is usually assumed to be approximately 100 km, although there is no distinct 'edge'.

Figure 18.16 A telescope at the Very Large Array, New Mexico, USA receiving radio waves from space

Figure 18.17 The telescopes at the Paranal Observatory on the top of Cerro Paranal, a mountain in the Atacama desert in Chile

Figure 18.18 The Hubble telescope

The use of telescopes on orbiting satellites has greatly increased the *resolution* of images from space (the resolution of images was discussed in detail in Chapter 11, for Higher Level students only). The Hubble telescope (Figure 18.18) has been at the centre of attention, with many of its spectacular images well known around the world. The telescope was launched in 1990 and named after the famous American astronomer, Edwin Hubble. It has a mass of about 11 tonnes and orbits approximately 560 km above the Earth's surface, taking 96 min for one complete orbit. One of the greatest achievements of astronomers using the Hubble telescope has been accurately determining the distances to very distant stars, enabling a much improved estimate for the age of the universe.

The second major advantage of placing a satellite in orbit is that it can detect radiations that would otherwise be absorbed in the atmosphere before reaching any *terrestrial* telescopes (those on the Earth's surface). Figure 18.19 indicates (approximately) the effect that the Earth's atmosphere has on preventing radiations of different wavelengths from reaching the Earth's surface.

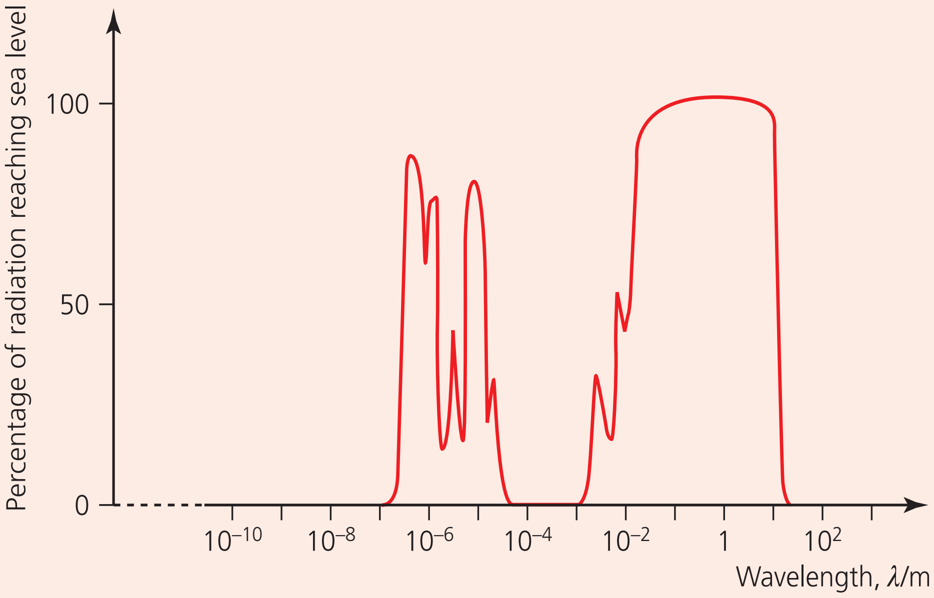


Figure 18.19 How the Earth's atmosphere affects incoming radiation

Questions

1 Make a sketch of Figure 18.19 and indicate and name the different sections of the electromagnetic spectrum.

2 Visit the Hubble web site, look at the magnificent images from space and make a list of the important characteristics of the telescope.

Additional Perspectives

A different kind of observatory

Before the invention of the telescope in about 1608, astronomers throughout the world made impressively accurate observations with their unaided eyes, often by also using a range of different devices to measure small angles.

More than 100 years after the discovery of the telescope, between 1727 and 1734 Maharaja Jai Singh II built an impressive observatory at Jaipur in India, which consisted of 14 large geometrical structures to assist astronomy with the unaided eye (Figure 18.20). The biggest of these is 27 m tall and is the largest sundial in the world. Its shadow can easily be seen to move at a rate of up to 6 cm every minute.

Figure 18.20 Jantar Mantar of Jaipur, India

The purpose of the structures was to measure time and the apparent motions of the planets and stars, but also to be impressive structures in themselves and to stimulate interest in the newly developing science of astronomy. In India at that time astronomy and astrology were closely connected, as they had been throughout the world in nearly all civilizations (and even today for many people).

Question

1 Many people believe that the positions of the Moon, stars and planets can influence our individual lives and our futures. Do you think that this is possible? Explain your answer.

Wien's law and the Stefan–Boltzmann law

Stefan–Boltzmann law

E.2.5 Apply the Stefan–Boltzmann law to compare the luminosities of different stars.

Stars can be considered to be 'perfect' emitters of radiation. That is, to a good approximation, they are *perfect black bodies*, emitting the maximum possible radiation over a range of wavelengths – a concept introduced in Chapter 8. Of course, it may be difficult to imagine our Sun, or any other star, as a 'black body'. The term 'black body' seems misleading when discussing the *emission* of radiation from an object, because it arose originally in respect of a body's ability to *absorb* radiation, although a good absorber will also always be a good emitter.

Because stars are almost perfect black bodies, we can apply the **Stefan–Boltzmann law** to them, remembering, from Chapter 8, that the emitted luminosity (power), L, depends only on the surface area, A, and the absolute (Kelvin) temperature, T, of the emitting surface according to the equation:

$$L = \sigma A T^4$$

This equation is given in the IB *Physics data booklet*.

(σ is the **Stefan–Boltzmann constant** = $5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$, given in the IB *Physics data booklet*). Remember also that the surface area of a sphere = $4\pi r^2$.

Worked example

3 What is the luminosity of a star of radius 2.70×10^6 km and surface temperature 7120 K?

$$L = \sigma A T^4$$
$$L = (5.67 \times 10^{-8}) \times 4\pi \times (2.7 \times 10^6 \times 10^3)^2 \times (7120)^4$$
$$L = 1.33 \times 10^{28}\,\mathrm{W}$$

17 A star has a surface area of $1.8 \times 10^{19}\,\mathrm{m^2}$ and a surface temperature of 4200 K. What is its luminosity?

18 If a star has a luminosity of 2.4×10^{28} W and a surface temperature of 8500 K, what is:
a its surface area
b its radius?

19 What is the surface temperature of a star which has an area of $6.0 \times 10^{20}\,\mathrm{m^2}$ and a luminosity of 3.6×10^{30} W?

20 If the star in question **17** above is 17.3 ly away, what will its apparent brightness be when seen from Earth?

21 If the star in question **18** above has an apparent brightness of $2.5 \times 10^{-8}\,\mathrm{W\,m^{-2}}$, how many kilometres is it from Earth?

22 Compare the luminosities of these two stars: star A has a surface temperature half that of star B, but its radius is forty times greater.

23 A star has eighty times the luminosity of our Sun and its surface temperature is twice that of the Sun. How much bigger is the star than our Sun?

Wien's law

E.2.6 **State** Wien's (displacement) law and apply it to explain the connection between the colour and temperature of stars.

Using the Stefan–Boltzmann law, if the luminosity of a star is known, we can then calculate its surface area, A, and hence its radius, r, but only if we know its surface temperature. This temperature can also be determined if the spectrum of radiation from the star has been analysed.

When objects emit radiation from their surfaces they do so over a range of different wavelengths (or frequencies) and this distribution will change depending upon the surface temperature. For approximate perfect black-body emitters, like stars, the distribution of emitted wavelengths at different temperature is well understood and best represented graphically, as shown in Figure 18.21 (this graph is similar to one previously seen in Chapter 8 on page 299).

Notice again from this graph that the overall intensity is *much* greater at higher temperatures and also that the wavelength at which the radiation is greatest becomes lower as the surface gets hotter. **Wien's (displacement) law** describes this behaviour:

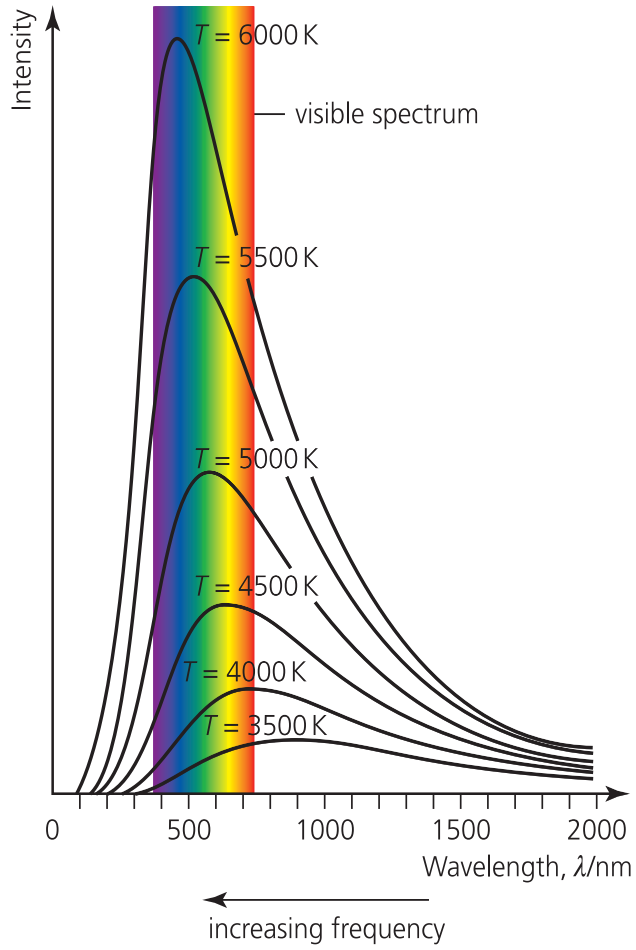


Figure 18.21 The distribution of wavelengths emitted from stars at different temperatures

$$\lambda_{max}T = \text{constant}$$

This equation and a value for the constant are given in the IB *Physics data booklet*.

This is an *empirical* law and the constant was determined experimentally to be 2.90×10^{-3} K m (remember that the temperature must be in kelvins). Wien's law can be used to determine the surface temperatures of the stars, but remember that in doing this we are assuming that they are *perfect black bodies*.

Worked example

4 What is the surface temperature of a star which emits radiation with a peak of intensity at 1.04×10^{-7} m?

$$\lambda_{max}T = \text{constant}$$

$$(1.04 \times 10^{-7})T = 2.90 \times 10^{-3}$$

$$T = \frac{2.90 \times 10^{-3}}{1.04 \times 10^{-7}}$$

$$T = 27\,900\,\text{K}$$

24 If the surface temperature of the Sun is 5700 K, at what wavelength is the emitted radiation maximized? In what part of the visible spectrum is this wavelength?

25 A star emits radiation which has its maximum intensity at a wavelength of 6.5×10^{-7} m.
- **a** What is its surface temperature?
- **b** If it has a luminosity of 3.7×10^{29} W, what is the surface area of the star?
- **c** What is its radius?

26
- **a** At what wavelength does a star with a surface temperature of 8200 K emit radiation with maximum intensity?
- **b** If this star has a radius of 1.8×10^{6} km, what is its luminosity?
- **c** If it is 36 ly from Earth, what is its apparent brightness?

27 The star Canopus has a luminosity of 5.8×10^{30} W and a radius of 4.5×10^{10} m. Use this data to estimate the wavelength at which it emits the most radiation.

28 Sketch graphs comparing the emission spectra from the stars Betelgeuse (3600 K) and Alkaid (20 000 K).

Stellar spectra

E.2.7 **Explain** how atomic spectra may be used to deduce chemical and physical data for stars.

Using spectra to determine the chemical composition of stars

As the *continuous spectrum* emitted from a star passes through its cooler outer layers, some wavelengths will be absorbed by the atoms present. When the radiation is detected on Earth, an *absorption spectrum* (previously discussed in Chapter 7) will be observed.

Since we know that every chemical element has its own unique spectrum, this information can be used to identify the elements present in the outer layers of a star. The element helium is the second most common in the universe (after hydrogen), but it was not detected on Earth until 1882. Fourteen years earlier, however, it had been identified as a new element in the Sun from its spectrum (see Figure 18.22).

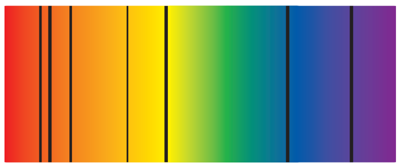


Figure 18.22 The absorption spectrum of helium

Using spectra to determine the velocity of stars and galaxies

If a source of light is not stationary but moving (very quickly) towards or away from an observer there will be a *shift* (slight change) in the wavelengths (or frequencies) of the spectral lines observed. The pattern of the absorption lines on the spectrum is the same, but all the lines are very slightly shifted from the positions they would occupy if there were no motion. This is called the **Doppler effect/shift**. (This is discussed in Chapter 11 for Higher Level students only.) We are all familiar with the Doppler effect in the sound received from moving vehicles –as a car approaches we hear a higher-pitched sound (smaller wavelength) than when it is moving away from us (Figure 18.23).

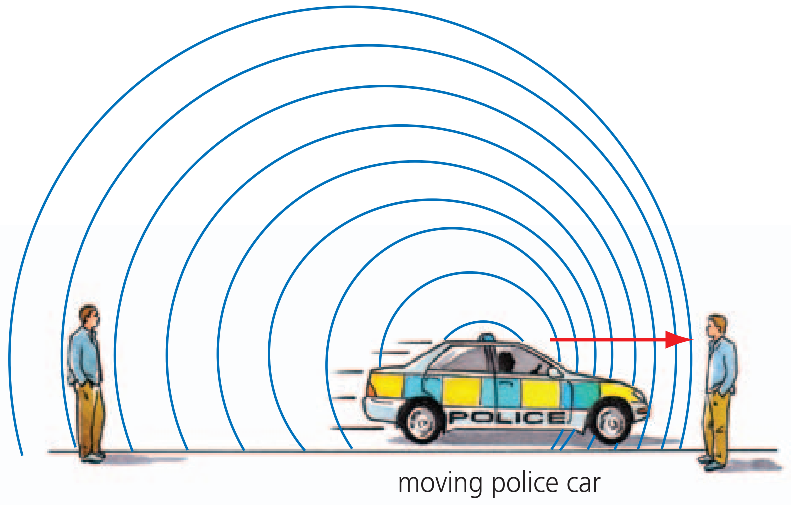


Figure 18.23 The Doppler effect for sound

In the case of light waves, the shift is very small and usually undetectable unless a source, such as a star (or galaxy), is moving very quickly. Careful observation of the line spectrum received from a star (see Figure 18.24) can be used to calculate the velocity of the star.

In example A in Figure 18.24, all the absorption lines have been shifted towards lower frequencies and this is commonly described as a **red-shift**. A red-shift occurs in the

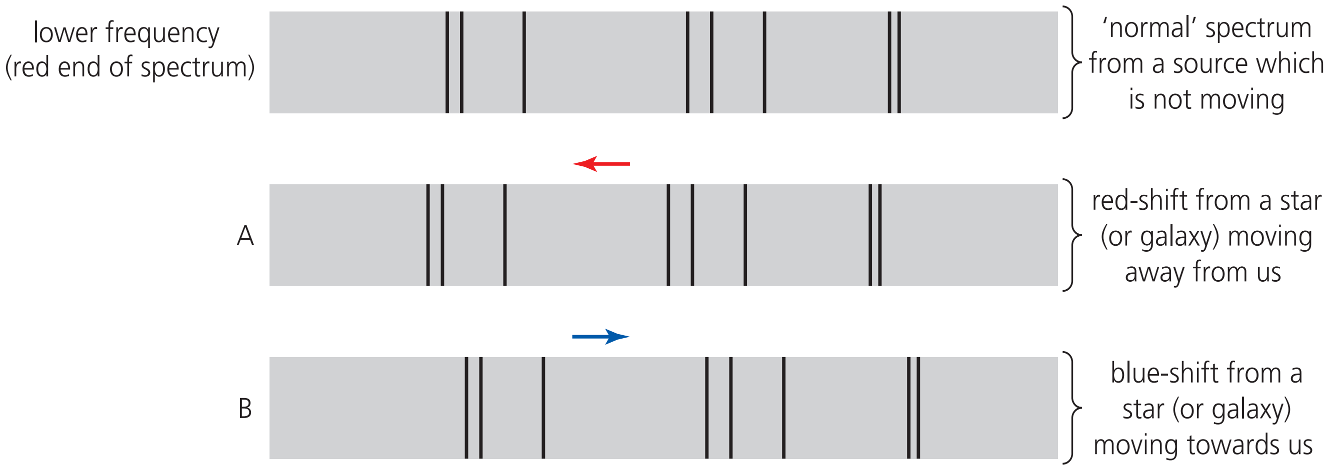


Figure 18.24 Red- and blue-shifts

radiation received from a star or galaxy that is moving away (*receding*) from the Earth. If a star or galaxy is moving towards Earth then the shift will be towards higher frequencies and is called a **blue-shift**, as shown in example B.

In this way it is possible for astronomers to determine the direction of movement of many stars and galaxies. Since a faster speed results in a larger shift, the galaxy's or star's speed can also be calculated from precise measurements of the shift. The light from most galaxies is red-shifted because the universe is expanding.

The classification of stars by the colours that they emit

E.2.8 Describe the overall classification system of spectral classes.

Figure 18.21 shows that the *continuous emission spectra* produced by various stars show slight differences depending on their surface temperatures. To observers on Earth this will be noticed as slight differences in colour, and this has long been the way in which astronomers group and classify different stars. In general, as shown in Figure 18.21, cooler stars are slightly redder and hotter stars are slightly bluer. Table 18.3 lists the eight **spectral classes** into which all visible stars are placed.

Table 18.3 Spectral classes, temperatures and colours

Spectral class	Surface temperature/K	Colour
O	30 000–50 000	Blue
B	10 000–30 000	Blue–white
A	7500–10 000	White
F	6000–7500	Yellow–white
G	5000–6000	Yellow
K	3500–5000	Yellow–red (orange)
M	2000–3500	Red

This apparently haphazard system of lettering stars according to their colour is an adaptation of an earlier alphabetical classification. A widely quoted *mnemonic* for remembering the order (from the hottest) is 'Only Bad Astronomers Forget Generally Known Mnemonics', or you may like to make up your own.

29 **a** What is the spectral class of our Sun?
b We often refer to the light from our Sun as 'white'. Discuss whether this is an accurate description.

30 Two common types of star are called red giants and white dwarfs. What spectral class would you expect them to be?

31 What is the spectral class and colour of the star Alkaid (referred to in question 28)?

32 The spectra from most stars and galaxies are *red-shifted*. Explain what this means and explain the reason for the shift.

33 Explain in detail how the absorption spectrum of helium shown in Figure 18.22 was formed.

Summary of information that we can deduce about a star from the radiation received on Earth

- We can calculate its *luminosity* if we know its apparent brightness and distance away ($b = L/4\pi d^2$).

- We can calculate its *surface temperature* if we analyse its continuous spectrum ($\lambda_{max}T$ = constant).
- We can calculate its *surface area* (and radius) if we know its luminosity and surface temperature ($L = \sigma AT^4$).
- We can determine *elements present in its outer layers* if we observe its absorption spectra.
- We can calculate its *speed and relative direction of movement* by measuring the Doppler shift of its spectrum.
- (For most stars it is also possible to estimate their *mass* from their luminosity; this will be explained later in the Higher Level section.)

Once astronomers had collected all this information about a large number of stars, they clearly wanted to identify similarities and differences to see if there were any obvious patterns. It turned out that all stars have a great amount in common, and that the differences were much less than expected.

Types of star

E.2.9 **Describe** the different types of star.

Although there are differences in size, temperature and luminosity between stars, they all behave in very similar and predictable ways during their 'lifetimes'. The stars that we can see have very different *ages*. During most of their lifetimes they change very little and during this time we call them 'main sequence' stars. It is only later, when the supply of hydrogen begins to get used up, that there are significant differences between them because of their masses, and then they change to different types of star.

- **White dwarf stars** are relatively hot and therefore blue/white in colour, but their luminosity is relatively low because they are small in size and we need telescopes to see them.
- **Red giant stars** are relatively cool and therefore yellow/red in colour. However, they have a higher luminosity than many stars because they are large. Similar comments can be used to describe the rarer **red supergiant stars**, except they are even bigger.
- **Cepheid variables** are an unusual type of star because their luminosities vary. They have a very important role in astronomy because their regular variations in brightness can be used to determine the distance to distant galaxies.
- It is common for two stars to orbit each other (or, more precisely, orbit their common centre of mass). This is called a **binary star system**. Observations of this kind of system are very important in determining the mass of stars using calculations which involve the stars' separation and their orbital period.

Different types of binary stars

E.2.10 **Discuss** the characteristics of spectroscopic and eclipsing binary stars.

- **Visual binary stars** are a pair of binary stars which can be seen (with a telescope) to be separate (Figure 18.25).

  But it is often not possible using even the very best telescopes to see that the object being observed is in fact two stars rather than one (in other words, they cannot be visually *resolved* as two separate sources), so this must be determined in other ways by trying to detect slight variations in the radiation received. Because of this, many binary stars are classified as 'spectroscopic' or 'eclipsing', depending on how they have been detected.

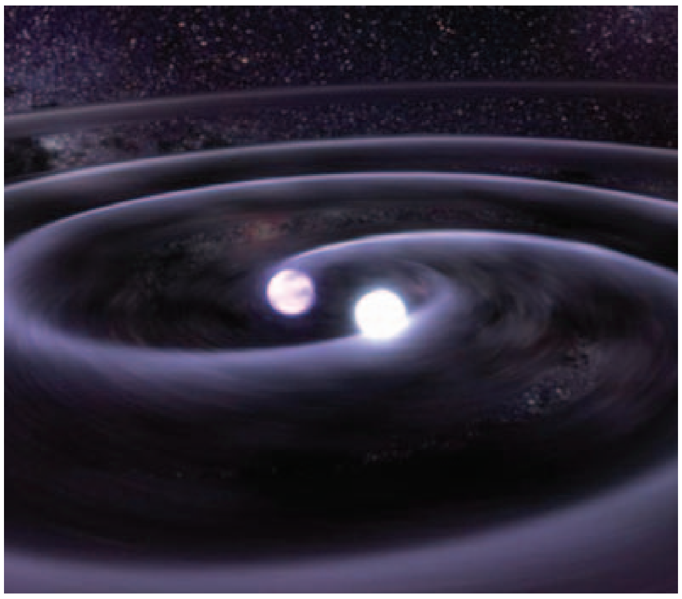


Figure 18.25 In a visual binary star system the stars can be seen to be separate

- **Spectroscopic binary stars** are those binary systems that can be identified using the Doppler effect. Because at all times the two stars are moving in opposite directions to each other, when the radiation from one is red-shifted, radiation from the other is blue-shifted (Figure 18.26).

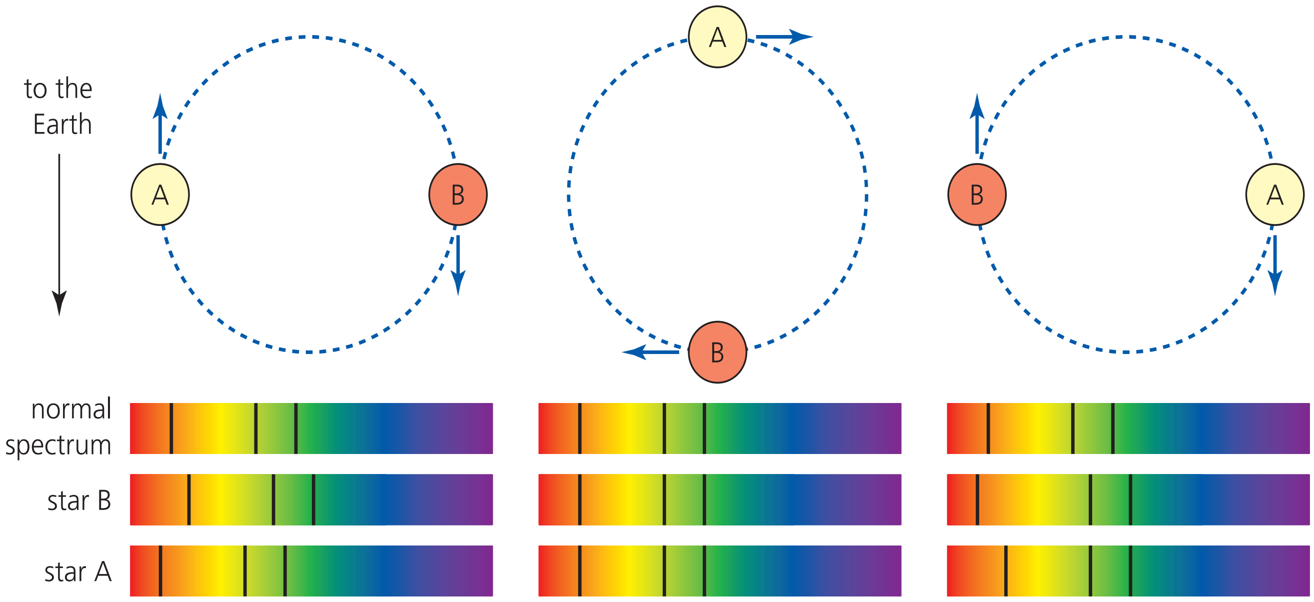


Figure 18.26 The Doppler effect in spectroscopic binary stars

- **Eclipsing binary stars** are those rarer systems in which the orbits are in the same plane as a line from the stars to the observer, so that in each orbit of the system one star will pass in front of the other and block (*eclipse*) radiation which would otherwise travel to the observer.

  A typical sequence is shown in principle in Figure 18.27, although it has been simplified by assuming that one star is much larger and does not move. The letters on the intensity–time graph refer to the position of the smaller star, as shown in the diagram. The size of the drop in brightness will depend on the comparative sizes and brightnesses of the two stars.

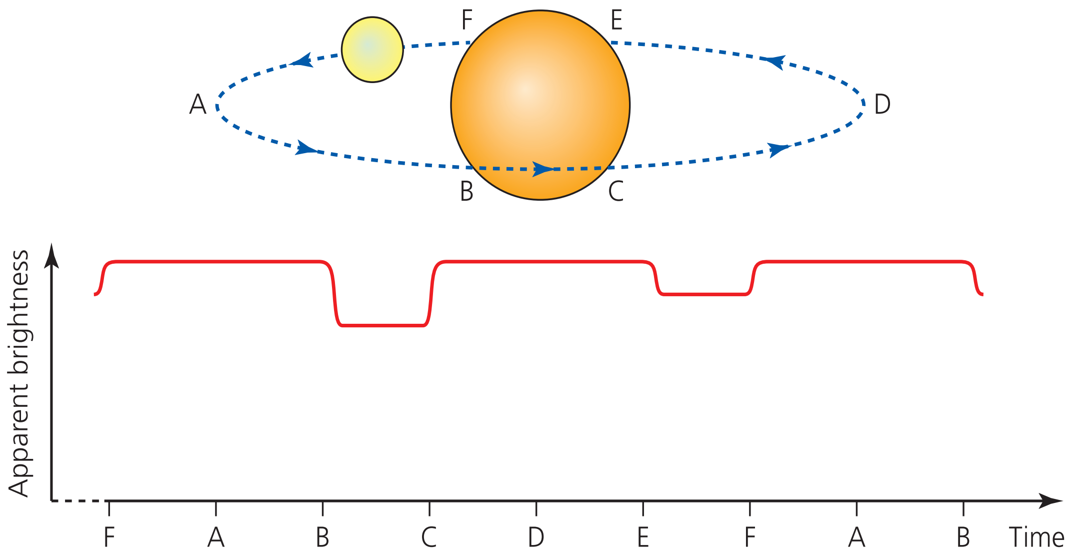


Figure 18.27 Variation in brightness received from eclipsing binary stars

34 **a** Make a sketch of the graph in Figure 18.27. Mark on your sketch:
  i the *transit time* of the small star across the larger star
  ii the time period for one orbit of the system.
b Add to your sketch a graph to represent the apparent brightness of a binary star system with two identical stars orbiting each other.

35 Explain why non-visual binary systems are more likely to be detected using the Doppler effect than an eclipsing effect.

36 The Moon and the Earth are a binary system, but the mass of the Earth is about 80 times that of the Moon. We commonly say that the Moon orbits the Earth, but could we also say that the Earth orbits the Moon? Find out where the centre of mass of the system is in relation to the Earth's surface.

37 Figure 18.26 shows the Doppler shifts of two similar spectral lines received from both stars.
- **a** Is it reasonable to assume that the same lines will be present in the spectrum of both stars? Explain your answer.
- **b** Make a copy of the three spectra shown for the first diagram, and add two more spectra to show how A and B would appear if the system had a shorter orbital period.

Additional Perspectives

Determining the mass of stars

Binary star systems play an important role in astronomy because they provide a good way of calculating stellar masses. Figure 18.28 shows a simplified example in which the two stars are shown moving in orbits around their common centre of mass. Note that they are always diametrically opposite each other and moving in opposite directions.

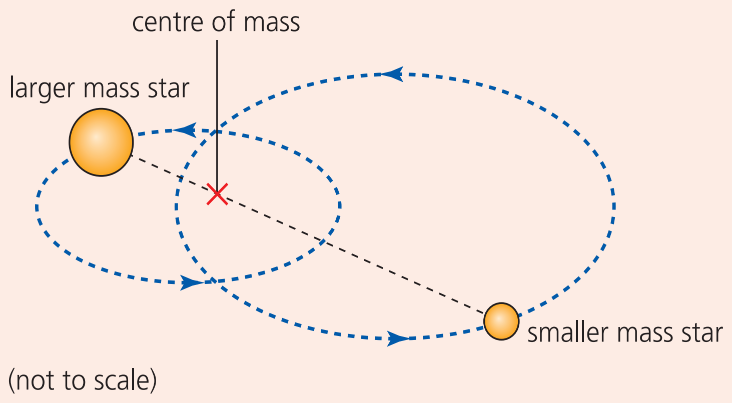


Figure 18.28 A simplified binary star system

Higher Level students saw in Chapter 9 that Newton's law of universal gravitation could be used to express Kepler's third law in the following form:

$$\frac{R^3}{T^2} = \frac{GM}{4\pi^2}$$

When applied to a binary system, M is the combined mass of the two stars and R is their average separation. If the period, T, can be measured, then the total mass can be calculated. More detailed calculations are needed to determine the mass of individual stars in the system, but the two masses are often similar to each other.

Similar reasoning can be applied to the masses of rotating galaxies and this plays an important role in estimating the mass of the universe.

Question

1 A binary star system is observed to have a period of 18.4 years. If the two stars have an average separation of 4×10^{12} m, what is their combined mass?

The Hertzsprung–Russell diagram

E.2.11 Identify the general regions of star types on a Hertzsprung–Russell (HR) diagram.

The luminosity of a star depends on its surface temperature and surface area ($L = \sigma AT^4$), so a particular star may be luminous because it is hot or because it is big, or both. In the early 20th century two scientists, Hertzsprung and Russell, separately plotted similar charts (diagrams) of luminosity against temperature in order to determine if there was any pattern in the way that the stars were distributed. Figure 18.29 shows a large number of individual stars plotted on a **Hertzsprung–Russell (HR) diagram**, with all luminosities compared to the luminosity of our Sun ($L_\odot$).

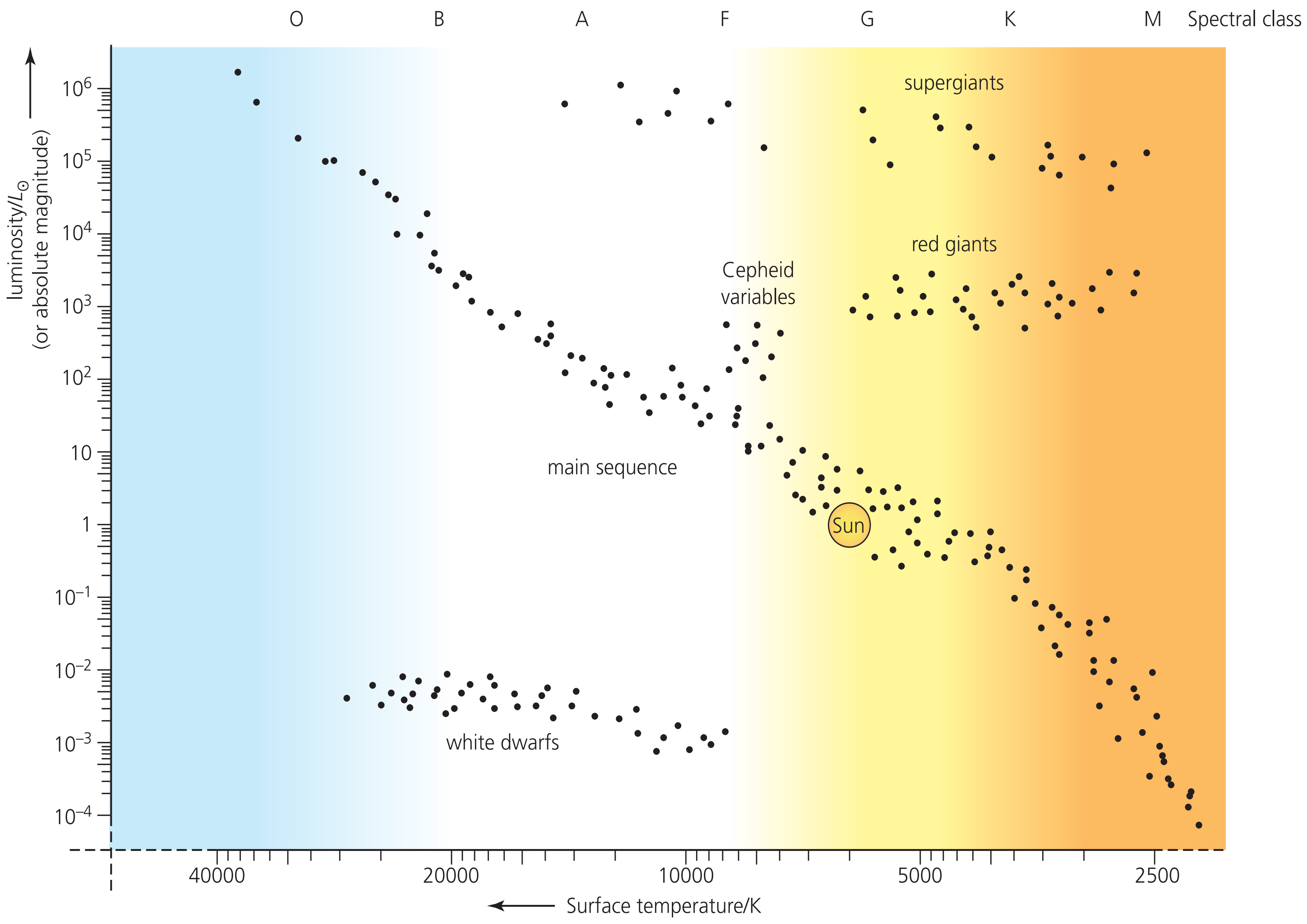


Figure 18.29 The Hertzsprung–Russell diagram

The Hertzsprung–Russell diagram seeks to find order in all the apparent diversity of the stars and has some important features.

- The vertical axis representing luminosity is also commonly labelled in terms of *absolute magnitude* (see Section E3).
- The variation of stars' luminosities (or absolute magnitudes) is enormous, so that the scale used is *logarithmic* rather than *linear*.
- The horizontal scale representing surface temperature can alternatively be labelled as *spectral class* – see Table 18.3 on page 697.
- By common convention, the horizontal temperature scale is reversed in order to retain the traditional order of the spectral classes. (This scale is also non-linear.)

Regardless of which quantities are plotted, or which scales are used, it should be apparent that the stars are *not* distributed at random in this diagram and that it can be used as the basis for classifying stars into certain groups.

Most stars (about 90%) can be located in the central band which runs from top left to bottom right in the diagram. These stars are stable and they are known as '**main sequence**' stars. The only significant structural difference between each of them is their mass/size (which is *not* indicated on the diagram). A more massive main sequence star will have a higher rate of fusion and therefore a higher temperature, giving it a greater luminosity. Such stars are therefore found towards the top left of the diagram. (The mathematical relationship is discussed in Section E5.) Our Sun is a main sequence star located near the middle of the diagram.

The equation $L = \sigma A T^4$ is used to interpret data from this diagram. It reminds us, for example, that if two stars have about the same luminosity, L, the one with the lower surface temperature, T, must be bigger and have the greater surface area, A.

It is important to realize that stable, main sequence stars stay in the *same* position on this diagram. It is only towards the end of their 'lifetimes' that changes occur which result in them moving to new positions on the HR diagram. For example, a main sequence star will become a red giant (or red supergiant), then later it may become a white dwarf (depending on its mass).

38 Using the HR diagram, how is it possible to deduce that white dwarfs are small stars?

39 Would it be correct to describe many Cepheid variables stars as 'yellow giants'? Explain your answer.

40 In a few billion years our Sun will begin to change into a red giant star. Outline the changes that will take place.

41 What is a typical surface temperature and luminosity (in W) of a class A star?

E3 Stellar distances

Accurate measurement of the distances to stars provides the essential data that is the foundation for most other stellar calculations. We will discuss three methods for estimating stellar distances. These are known as **parallax**, **spectroscopic parallax** and by the use of **Cepheid variables**.

Perhaps the most obvious method is similar to that in which we might determine the distance to an inaccessible object, such as a boat or a plane, on Earth. If the object can be observed from two different places, then its distance away can be calculated using trigonometry. This *triangulation* method is shown in Figure 18.30.

An observer on land sees the boat from position P and then moves to position Q. If the angles α and β are measured and the distance PQ is known, then the other distances can be calculated. When astronomers want to locate a star they can try to observe it from two different places, but the distance between two different locations on Earth is far too small compared with the distance between the Earth and the star. Therefore, astronomers observe the star from the same telescope at the same location, but at two different places in the Earth's orbit, that is, at different times of the year. To get the greatest difference they usually take two measurements separated in time by six months.

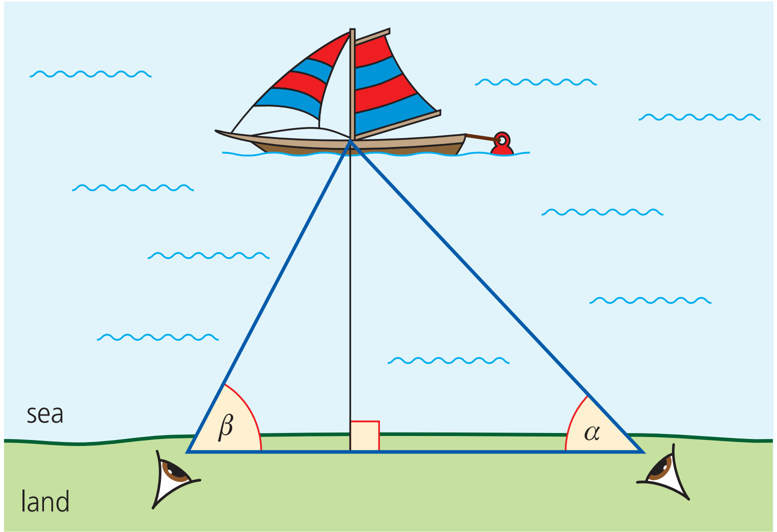


Figure 18.30 Determining the distance to a ship at sea using triangulation

The triangulation method described above to locate a boat would be much more difficult if the observer was in a *moving* boat at sea, and this is similar to the difficulty faced by astronomers locating stars from Earth. The problem can be overcome by comparing the position of the star to other stars much further away (in the 'background'). This is known as the parallax method.

Parallax method

E.3.2 **Describe** the stellar parallax method of determining the distance to a star.
E.3.3 **Explain** why the method of stellar parallax is limited to measuring stellar distances less than several hundred parsecs.

Parallax is the visual effect of a nearby object appearing to move its position, as compared to more distant objects (behind it), when viewed from different positions. A simple example is easily observed by looking at your finger held in front of your face and the background behind it, first with one eye and then the other. In the same way, a 'nearby' star can appear to *very slightly* change its position during the year compared to other stars much further away (although, as we have noted before, stars generally appear to remain in fixed patterns over very long periods of times).

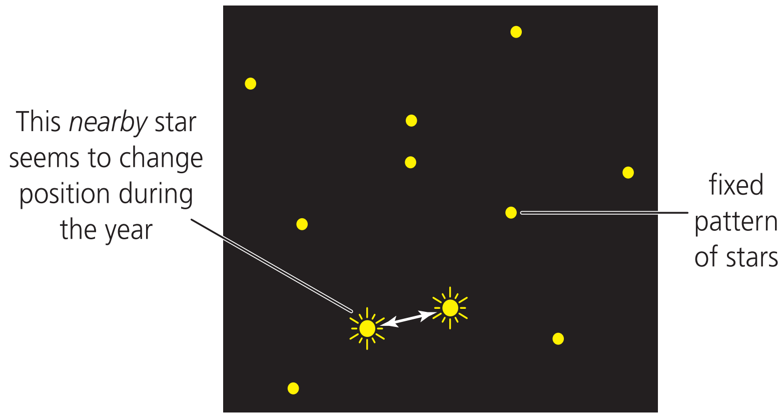


Figure 18.31 A nearby star's apparent movement due to parallax

Stellar parallax (Figure 18.31) was first confirmed in 1838. Many astronomers had tried to detect it before (without success) because the existence of stellar parallax provides evidence for the motion of the Earth around the Sun.

Using telescopes, astronomers measure the **parallax angle** between, for example, observations of the star made in December and June. Figure 18.32 shows the angular positions of a nearby star in December and June marked as P_1 and P_2. (In Figures 18.32 and 18.33 the size of the parallax angle has been much exaggerated for the sake of clarity.)

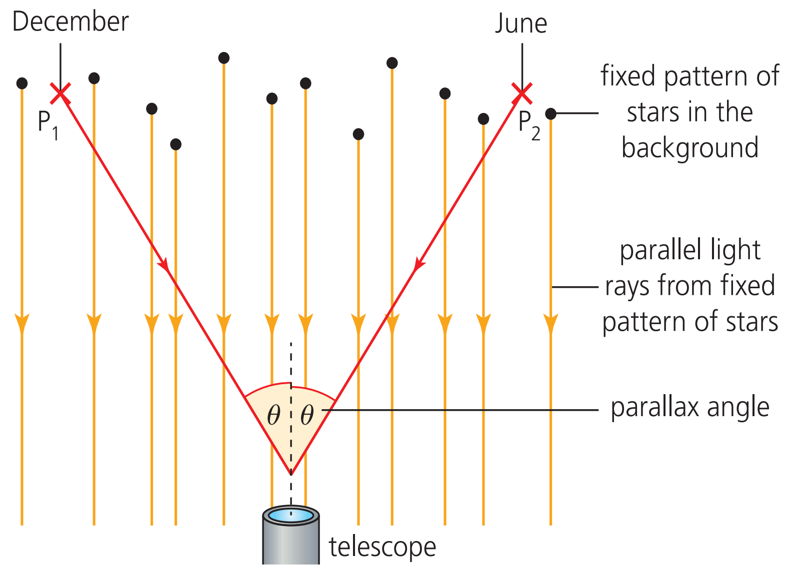


Figure 18.32 Measuring the parallax angle six months apart

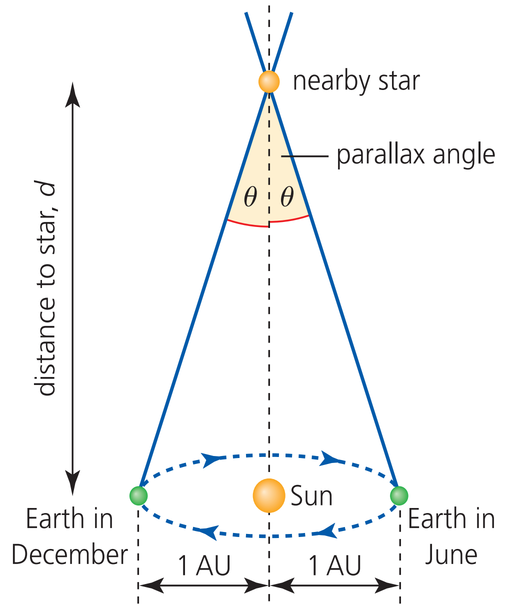


Figure 18.33 The geometry of the parallax angle

If the measurements are made exactly six months apart, the distance between the locations where the two measurements are taken is the diameter of the Earth's orbit around the Sun. We assume that the orbit is circular, so that the radius is constant. The Earth–Sun distance (1.50×10^{11} m) is used frequently in calculations and is known as one **astronomical unit** (AU).

The parallax of even the closest stars is very small because of the great distances involved and this means that the parallax angles are so tiny that they are measured in **arc-seconds**, also called seconds. (There are 3600 arc-seconds in a degree.)

Once the parallax angle has been measured, simple geometry can be used to calculate the distance to the star (Figure 18.33):

$$\text{parallax angle, } \theta \text{ (rad)} = \frac{1.50 \times 10^{11}}{d \text{ (in m)}}$$

(Note that the distance from the Earth to the star and the distance from the Sun to the star can be considered equal for such small angles, so that θ (rad) = $\sin\theta = \tan\theta$.)

Worked example

5 Calculate the distance, d, to a star if its parallax angle, θ, is 0.240 arc-seconds and 0.240 arc-seconds = 1.16×10^{-6} radians.

$$\theta = \frac{1.50 \times 10^{11}}{d}$$

$$\text{Therefore, } d = \frac{1.50 \times 10^{11}}{1.16 \times 10^{-6}} = 1.29 \times 10^{17}\text{ m } (= 13.7\text{ ly})$$

Parsec – a unit of distance

In order to measure the distances to nearby stars, astronomers measure parallax angles. The larger the parallax angle, the closer the star is to Earth. The angle and the distance

E.3.1 Define the *parsec*.

are inversely proportional. This simple relationship is the basis for an alternative unit for astronomical distances, the **parsec**, **pc**.

One parsec is defined as the distance to a star which has a parallax angle of one arc-second.

A similar calculation to that shown above will show that one parsec, the distance to a star with a parallax angle of one arc-second, is 3.09×10^{16} m = 3.26 ly. (This figure is given in the IB *Physics data booklet*.)

$$d\ (\text{parsec}) = \frac{1}{p\ (\text{arc-second})}$$

This equation is given in the IB *Physics data booklet*.

A star with a parallax angle, *p*, of two arc-seconds will be at a distance of 1/2 = 0.5 pc. A parallax angle of 0.25 arc-seconds will be from a star which is 1/0.25 = 4 pc away, etc. Table 18.4 shows the relationship between parallax angle and distance. Table 18.5 provides a summary of the conversions between units of distance used in astronomy.

Table 18.4 Parallax angles in arc-seconds and distances in parsecs

Parallax angle/arc-seconds	Distance away/pc
0.10	10.00
0.25	4.00
0.50	2.00
1.00	**1.00**
2.00	0.50
4.00	0.25

Table 18.5 Summary of distance units commonly used in astronomy

Unit	Metres, m	Astronomical units, AU	Light years, ly
1 AU =	1.50×10^{11}		
1 ly =	9.46×10^{15}	6.30×10^{4}	
1 pc =	3.09×10^{16}	2.06×10^{5}	3.26

E.3.4 Solve problems involving stellar parallax.

42 The parallax angle for Barnard's star is measured to be 0.55 arc-seconds. How far away is it from Earth:
- **a** in pc
- **b** in m
- **c** in ly?

43 What are the parallax angles for three stars which are at the following distances from Earth?
- **a** 2.47×10^{15} km
- **b** 7.9 ly
- **c** 2.67 pc

44 Convert an angle of 1 arc-second to:
- **a** degrees
- **b** radians.

The stellar parallax method is limited by the inability of telescopes on Earth to accurately measure very small angles below 0.01 arc-seconds. This means that this method is usually limited to those stars which are relatively close to Earth, within about 100 pc (= 1/0.01) and well within our own galaxy. The use of telescopes on satellites above the turbulence and distortions of the Earth's atmosphere can extend the range considerably, but it is still not suitable for the majority of stars, which are much further away.

Absolute and apparent magnitudes

Before describing the other two methods for determining the distances to stars which are further way than 100 pc, we need to introduce the idea of the *magnitude* of stars.

Apparent magnitude

E.3.5 Describe the apparent magnitude scale.

The people of ancient civilizations were much more familiar with the stars than most of us in the modern world. (They didn't have all the benefits and distractions of the modern world; nor the atmospheric pollution, especially light pollution.) Apart from accurately observing and recording the positions and apparent motions of the stars, they also compared their brightness. More than 2000 years ago, the Greeks devised a 'magnitude' scale for the brightness of stars and the same scale still provides the basis for recording present-day astronomers' observations of apparent brightness, despite its somewhat confusing nature. That is, astronomers usually prefer to discuss and compare stars' **apparent magnitudes**, m (no units) rather than their equivalent *apparent brightnesses* (W m^{-2}).

The original apparent magnitude scale was intended to cover the range between the brightest and dimmest stars that were visible with the unaided eye. (Of course, there were no telescopes at that time.) The brightest star was said to have a magnitude of one and the dimmest star a magnitude of six. All the others were assigned a whole number between two and five. A star of apparent magnitude five was considered to be two times brighter than a star of magnitude six, a star of apparent magnitude four was two times brighter than five, and so on. This meant that the scale was *logarithmic* (rather than linear) with each step corresponding to multiplying by two, so that the five steps in the scale corresponded to a total difference in brightness of $2 \times 2 \times 2 \times 2 \times 2 = 2^5 = 32$.

We now know that the maximum difference in apparent magnitude (brightness) of the stars visible without a telescope is closer to ×100 (rather than ×32), and therefore the scale was adapted in the 1850s so that each of the five steps between one and six then corresponded to a factor 2.512 (because $2.512^5 = 100$).

With very precise measurement of apparent brightness it is now possible to give each star a precise apparent magnitude correct to two or more decimal places, rather than using only whole numbers. Furthermore, the scale has been extended to vary between about +30 and −30 to include every object in the sky, from the very dimmest, which can only be seen with the very best telescopes, to the brightest planet (Venus) or even the Sun (−27). Between these extremes there is an enormous difference in apparent brightness, with a ratio of 2.515^{60} or about 10^{24}. It would be impossible to plot such an enormous range of possible values on a linear scale and this is why a logarithmic scale is used. Figure 18.34 shows the scale, with a few examples.

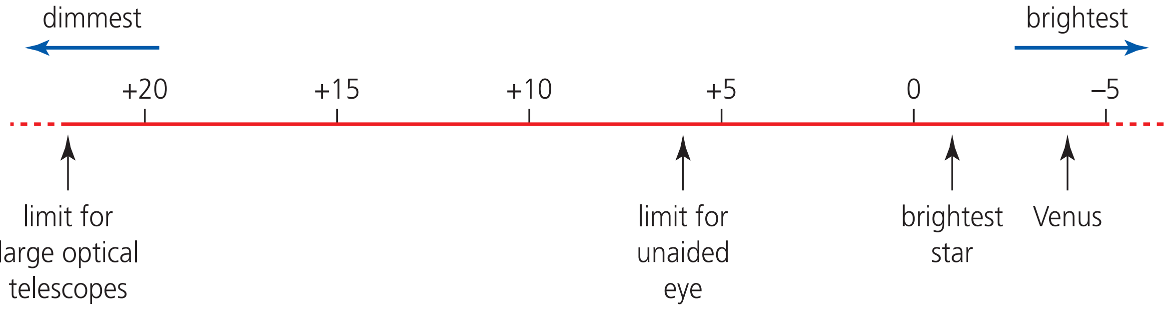


Figure 18.34 Apparent magnitude scale

The apparent magnitude scale is widely used by astronomers. Owing to its historical origins, it may seem confusing at first, because:

1 *brighter* stars are said to have *lower* magnitudes
2 the scale is logarithmic
3 and therefore it is possible for the brightest objects to have negative apparent magnitudes.

Directly converting apparent brightnesses (in $W\,m^{-2}$) to apparent magnitudes is *not* a requirement of the IB Diploma Physics course. However, students need to know how to *compare* the apparent brightnesses of two (or more) stars when given their apparent magnitudes. Consider the following examples.

- If the apparent magnitudes of two stars differ by 1, their apparent brightnesses differ by a factor of 2.512.
- If the apparent magnitudes of two stars differ by 2, their apparent brightnesses differ by a factor of 2.512^2 (= 6.31).
- If the apparent magnitudes of two stars differ by 1.73, their apparent brightnesses differ by a factor of $2.512^{1.73}$.
- If the apparent magnitudes of two stars differ by Δm, their apparent brightnesses differ by a factor of $2.512^{\Delta m}$.

E.3.8 **Solve** problems involving apparent brightness and apparent magnitude.

Worked example

6 Two stars have apparent magnitudes of –0.97 (star A) and +3.41 (star B).

a Which star is brighter as seen from Earth?
b Why is it impossible to know which star has the greater luminosity?
c What is the ratio of their apparent brightnesses received on Earth?
d If the apparent brightness of star A is $1.39 \times 10^{-12}\,W\,m^{-2}$, calculate the apparent brightness of star B.

a Star A, because brighter stars have smaller apparent magnitudes.
b Because there is no information about how far away the two stars are from Earth.
c Difference in apparent magnitudes, $\Delta m = 3.41 - (-0.97) = 4.38$
Ratio of apparent brightness = $2.512^{\Delta m} = 2.512^{4.38} = 56.5$
d $\dfrac{1.39 \times 10^{-12}}{56.5} = 2.46 \times 10^{-14}\,W\,m^{-2}$

45 Star S has an apparent magnitude of 4.8 and star T has an apparent magnitude of 3.2. How much brighter is star T than star S as viewed from Earth?

46 Is it possible for a star to have an apparent magnitude of 0? Explain your answer.

47 The intensity of radiation received from star X is 10 times greater than the intensity received from star Y. If the apparent magnitude of star X is –0.22, what is the apparent magnitude of star Y?

48 Star M has an apparent magnitude of +0.53 and star N has an apparent magnitude of –0.47 and apparent brightness of $4.0 \times 10^{-8}\,W\,m^{-2}$.
a Which of these stars can be seen with the unaided eye?
b Calculate the apparent brightness of star M.

Absolute magnitude

E.3.6 **Define** absolute magnitude.

Like apparent brightness, apparent magnitude cannot be used to compare the actual luminosities of stars because there is no consideration given to the distance between the star and Earth. To get over this problem, astronomers calculate the magnitudes that stars would have if they were all the same distance away (a distance of 10 pc is always used).

The **absolute magnitude, M**, of a star is defined as the apparent magnitude it would have if it was observed from a standard distance of 10 pc.

Using the distance to the star in parsecs, the conversion between apparent and absolute magnitudes is performed with the following equation.

$$m - M = 5 \lg\left(\frac{d}{10}\right),\ \text{where } d \text{ is the distance to the star in parsec}$$

This equation is in the IB *Physics data booklet*, but there is no need to be able to understand or to explain its origin.

Absolute magnitude can be considered as a widely used indication of a star's luminosity, using an arbitrary scale. Table 18.6 gives some examples (for a variety of reasons, the values quoted are not definitive).

Table 18.6 Magnitudes of the brightest stars as seen from Earth (Alpha Centauri is a binary system)

Star	Apparent magnitude/*m*	Distance/pc	Absolute magnitude/*M*
Sirius	−1.44	2.7	1.4
Canopus	−0.72	70	−4.9
Alpha Centauri	−0.27	1.3	4.2
Arcturus	−0.043	11	−0.25
Vega	−0.032	7.7	0.60

Worked example

E.3.7 Solve problems involving apparent magnitude, absolute magnitude and distance.

7 A star is 3.4 pc from Earth and has an apparent magnitude of +0.38. Show that its absolute magnitude is 2.7.

$$m - M = 5 \lg\left(\frac{d}{10}\right)$$

$$0.38 - M = 5 \lg\left(\frac{3.4}{10}\right)$$

$$M = 0.38 - 5 \lg(0.34)$$

$$M = 2.7$$

49 A star Sirius has an apparent magnitude of +1.7 and an absolute magnitude of +4.3. How far away is it (in pc)?

50 The star Deneb has an absolute magnitude of about −8.4 and is one of the most luminous visible stars. If its distance away is 2700 ly, what is its apparent magnitude?

51 The star Antares is 5.2×10^{15} km from Earth and has an apparent magnitude of 0.96. Calculate its absolute magnitude.

E.3.9 State that the luminosity of a star may be estimated from its spectrum.
E.3.10 Explain how stellar distance may be determined using apparent brightness and luminosity.
E.3.11 State that the method of spectroscopic parallax is limited to measuring stellar distances less than about 10 kpc.

Spectroscopic parallax

Spectroscopic parallax is a method of determining stellar distances by using the HR diagram to determine the luminosity of a star from its spectrum.

But, note that the name *spectroscopic parallax* is misleading because it has nothing to do with using parallax techniques.

Using Wien's law the surface temperature of a star can be determined from its spectrum and, assuming that it is a main sequence star, it is a relatively simple matter to use the HR diagram (Figure 18.29 on page 701) to estimate its luminosity, L, and hence calculate its distance, d, away from Earth by using $b = L/4\pi d^2$ and a measurement of its apparent brightness, b.

This method assumes that the extra distance travelled by radiation from more distant stars has not been altered in any way by the journey. For example, if any radiation is absorbed or scattered during the journey, the value of apparent brightness used in calculations will be less than it would have been without the absorption or scattering, leading to an over-estimate of the distance to the star.

Because the exact position of the star on the HR diagram may not be known with accuracy and because of unknown amounts of scattering/absorption, there is a significant uncertainty in this method of determining stellar distances.

For distances greater than about 10 000 pc (10 kpc) determining stellar distance using spectroscopic parallax becomes very unreliable.

In fact, the use of spectroscopic parallax is mostly confined to our galaxy. The majority of stars are obviously further away in other galaxies, so to determine the distances to those galaxies we need other methods. Table 18.7 summarizes the methods used for different distances.

Table 18.7 Methods for measuring astronomical distances

Distances up to about...	Method
Several hundred parsec (0.5 kpc)	Stellar parallax
Several thousand parsec (10 kpc)	Spectroscopic parallax
Several million parsec (20 Mpc)	Cepheid variables
Greater distances	Supernovae and/or red-shift measurements

E.3.12 **Solve** problems involving stellar distances, apparent brightness and luminosity.

Worked example

8 Estimate the distance from Earth (in pc) of a main sequence star which has a surface temperature of 7500 K and an apparent brightness of $4.6 \times 10^{-13}\,\text{W m}^{-2}$.

Looking at the HR diagram we can determine that the luminosity of a main sequence star of this luminosity is about thirty times the luminosity of the Sun (3.8×10^{26} W), or approximately 1.1×10^{28} W.

The equation $b = \frac{L}{4\pi d^2}$ can then be used to calculate the distance d:

$$4.6 \times 10^{-13} = \frac{1.1 \times 10^{28}}{4\pi d^2}$$

$$d = 4.4 \times 10^{19}\,\text{m}$$

$$d = \frac{4.4 \times 10^{19}}{3.26 \times 9.46 \times 10^{15}} = 1400\,\text{pc}$$

52 A main sequence star emits radiation with a maximum intensity of 3.7×10^{-7} m. Determine:
a its surface temperature
b its spectral class
c its luminosity
d its distance away, if its apparent brightness is $5.2 \times 10^{-8}\,\text{W m}^{-2}$.

53 Give two reasons why spectroscopic parallax gives unreliable results for distant stars.

54 A main sequence star of luminosity 6.3×10^{30} W is a distance of 24 000 pc away from Earth.
a Where would you expect to locate it on a HR diagram?
b What is its approximate surface temperature and stellar class?
c Calculate its apparent brightness.

E.3.13 Outline the nature of a Cepheid variable.
E.3.14 State the relationship between period and absolute magnitude for Cepheid variables.
E.3.15 Explain how Cepheid variables may be used as 'standard candles'.

Cepheid variables

The closest galaxy to the Milky Way is Andromeda, at a distance of about 0.8 Mpc. But most galaxies are much, much further away and the spectroscopic parallax methods described cannot be used for determining stellar distances greater than about 10 kpc. Observations of a certain kind of relatively large class of star, called a **Cepheid variable**, can greatly extend the range, to about 20 Mpc. Cepheid variables are named after the star Delta Cephei, which was the first one discovered.

The vital property of Cepheid variables that makes them so useful is that their luminosity (absolute magnitude) can be determined from observing *variations* in their apparent brightness (magnitude). Figure 18.35 shows the changing brightness of a Cepheid variable which has a period of about 5 days. Students in the Northern Hemisphere may be familiar with the North Pole Star (Polaris), which is a Cepheid variable with a period of about 4 days.

Cepheid variables are unstable stars in which the outer layers regularly expand and contract (typically by 30%) with surprisingly short periods (in astronomical terms), resulting in very regular and precise variations in luminosity. A typical period is a few weeks. If the surface temperature remains approximately constant then the increasing luminosity is explained by the greater surface area when the star expands.

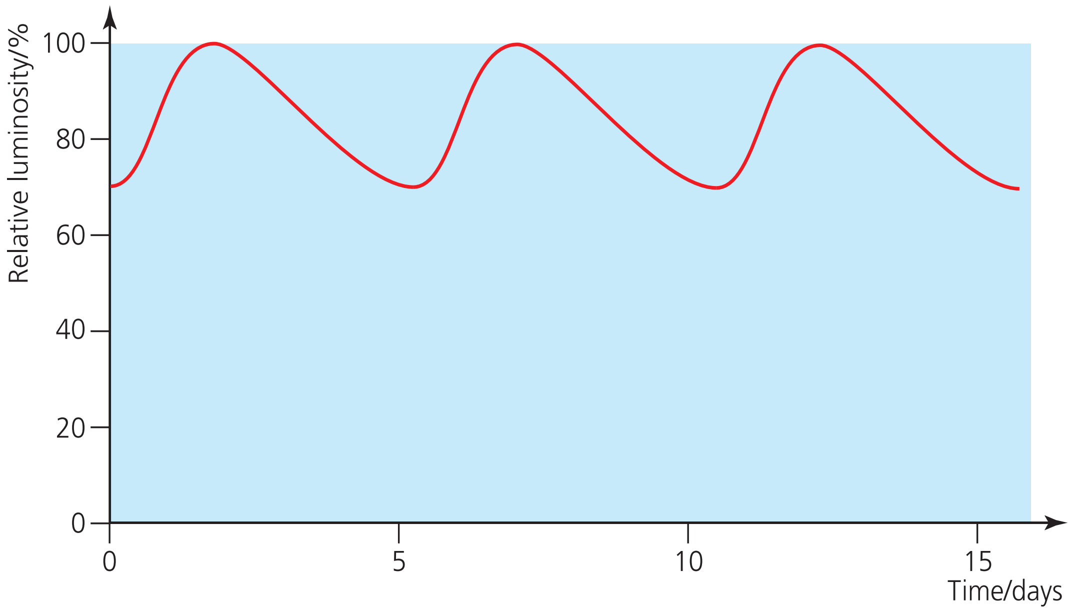


Figure 18.35 Variation in luminosity (or absolute magnitude) of a Cepheid variable

Although Cepheid variables are not common stars, they are important and their behaviour has been studied in great depth. From observations on those Cepheid variables which are relatively close to Earth (so that their distances can be confirmed using stellar parallax methods), it is known that there is a precise relationship between the time period of their pulses of luminosity (and hence their received apparent brightness on Earth) and the peak value of that luminosity. This was first discovered by Henrietta Leavitt (Figure 18.36) in 1908. This is called the **period–luminosity relationship** and it is commonly presented in graphical form as shown in Figure 18.37.

The greater the periodic variation in luminosity, the greater the absolute magnitude (luminosity) of the Cepheid variable.

Figure 18.36 Henrietta Leavitt discovered the periodicity of Cepheid variables in 1908

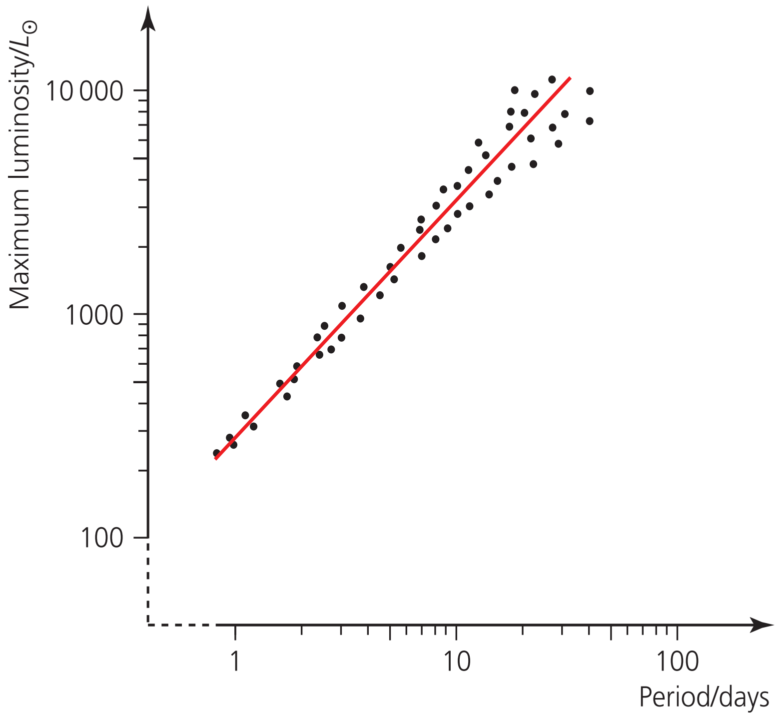


Figure 18.37 Period–luminosity (absolute magnitude) relationship for a Cepheid variable

Note the logarithmic nature of both the scales on this graph, which is necessary to produce the linear graph. Luminosity may be replaced with brightness or magnitude on the vertical axis.

Once again, if the luminosity and apparent brightness are known, the equation $b = L/4\pi d^2$ can then be used to calculate the distance, d. However, the inaccuracies in the data involved mean that these estimates of distance, especially to the furthest galaxies, are uncertain. This uncertainty is a significant problem when estimating the age of the universe.

Astronomers often describe Cepheid variables as '**standard candles**' because if their distance from Earth is determined, it can then be taken as a good indication of the distance of the whole galaxy from Earth (since that distance is very much greater than the distances between stars within the galaxy, see Figure 18.8).

Worked example

E.3.16 Determine the distance to a Cepheid variable using the luminosity–period relationship.

9 A Cepheid variable in a distant galaxy is observed to vary in apparent brightness with a period of 9.0 days. If its maximum apparent brightness is $1.92 \times 10^{-9}\,\mathrm{W\,m^{-2}}$, how far away is the galaxy?

From a graph of luminosity–period (similar to Figure 18.37), the maximum luminosity can be determined to be 3100 times the luminosity of the Sun.

$$\text{luminosity} = 3100 \times (3.8 \times 10^{26}\,\mathrm{W}) = 1.2 \times 10^{30}\,\mathrm{W}$$

$$b = \frac{L}{4\pi d^2}$$

$$1.92 \times 10^{-9} = \frac{1.2 \times 10^{30}}{4\pi d^2}$$

$$d = 7.0 \times 10^{18}\,\mathrm{m} = 740\,\mathrm{ly} = 230\,\mathrm{pc}$$

55 **a** If a Cepheid variable has a period of 15 days, what is its approximate maximum luminosity?
b If the star is 3.3 Mpc from Earth, what is the maximum observed apparent brightness?

56 A Cepheid variable is 15 kpc from Earth and is observed to have an apparent brightness of $8.7 \times 10^{-13}\,\mathrm{W\,m^{-2}}$.
a Calculate the maximum luminosity of this star.
b Use Figure 18.37 to estimate the time period of the variation in the star's luminosity.

57 For very large distances astronomers may use supernovae (rather than Cepheids) as 'standard candles'. Suggest a property of supernovae which might be necessary for this.

E4 Cosmology

Cosmology is the study of the universe. It has always been the nature of many individuals, societies and cultures to wonder what lies beyond the Earth. The fact that the Sun and the stars appear to revolve around the Earth led early civilizations, understandably but wrongly, to believe that a stationary Earth was the centre of everything, and this belief was often fundamental to their religions. Indeed, even today some people still believe from their everyday observations, or their religious beliefs, that the Sun orbits around the Earth rather than the other way around.

Additional Perspectives

'The shoulders of giants'

Nicolas Copernicus, a Polish astronomer and cleric (Figure 18.38), is considered by many to be the founder of modern astronomy. In 1530 he published a famous paper stating that the Sun was the centre of the universe and that the Earth, stars and planets orbited around it (a helio-centric model). At the time, and for many years afterwards, these views directly challenged 'scientific', philosophical and religious beliefs. It was then generally believed that the Earth was at the centre of everything (a 'geo-centric' model). That profound and widespread belief dated all the way back to Ptolemy, Aristotle and others nearly 2000 years before. It should be noted, however, that Aristarchus, in ancient Greece, is generally credited with being the first well-known person to propose a helio-centric model.

Figure 18.38 Copernicus

More than 100 years after Copernicus, and still before the invention of the telescope, an eccentric Danish nobleman, Tycho Brahe, became famous for the vast number of very accurate observations he made on the motions of the five visible planets. He worked mostly at an elaborate observatory on an island in his own country, but went to Prague a few years before his death in 1601.

Johannes Kepler was Brahe's assistant before he died and he later worked on Brahe's considerable, but unexplained, data to produce his three famous laws of planetary motion.

At about the same time in Italy, the astronomer Giordano Bruno had taken the helio-centric model further with revolutionary suggestions that the universe was infinite and that the Sun was *not* at the centre. The Sun was, Bruno suggested, similar in nature to the other stars. He was burned at the stake in 1600 for these beliefs – killed for his, so-called, heresy. About 30 years later, one of the greatest scientific thinkers of all time, Galileo Galilei, was placed on trial by the Roman Catholic Church under similar charges. Many years earlier he had used the newly invented telescope to observe the moons of Jupiter and had reasoned that the Earth orbited the Sun in a similar way, as had been proposed by Copernicus. Under pressure, he publicly renounced these beliefs and was allowed to live the rest of his life under house arrest. All this has provided the subject of many books, plays and movies.

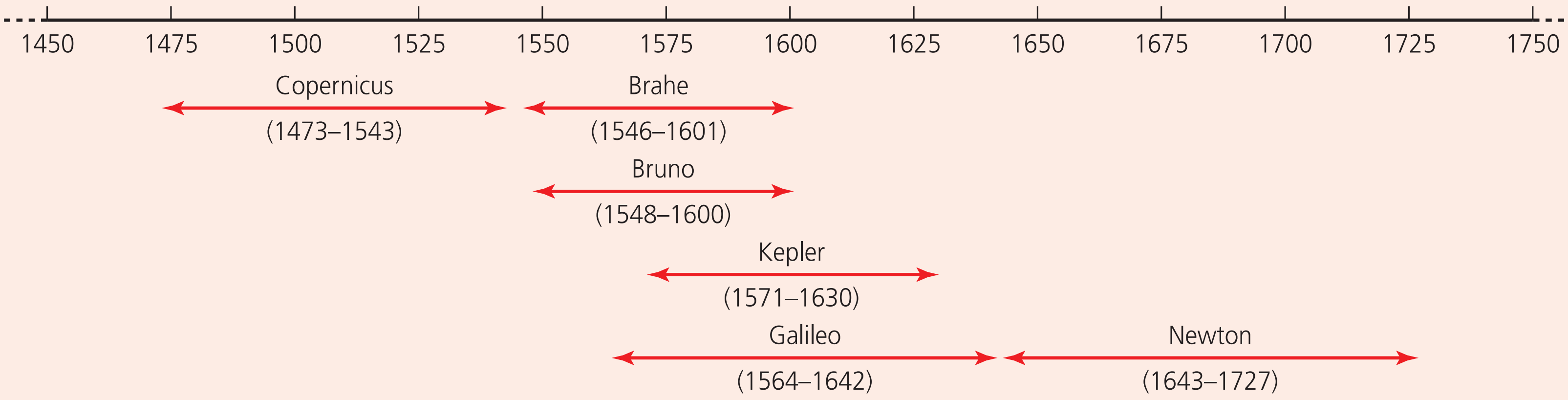


Figure 18.39 Time line of some famous early astronomers

Although Kepler had found an accurate way to mathematically describe the motion of the planets, an explanation was not produced until about eighty years later, when Newton was able to use the motion of the planets and the Moon as evidence for his newly developed theory of universal gravitation (Chapter 6).

Questions

1 Many people would place Newton and Galileo in a list of the top five scientists of all time but, to a certain extent, that is just a matter of opinion.
  a Why do you think Newton and Galileo are so highly respected?
  b What criteria would you consider when trying to decide if a scientist was 'great'?
2 Research the origins of the quotation which is the title of this Additional Perspective.

Olbers' paradox

E.4.1 Describe Newton's model of the universe.
E.4.2 Explain Olbers' paradox.

In the Newtonian model of the universe, the Earth, the Sun and the planets were just tiny parts of an *infinitely large* and *unchanging* (*static*) universe that had always been the way it is and always would be the same. In this model, the universe, on the large scale, is more or less the same everywhere, that is, it is *uniform* with stars approximately evenly distributed. Newton reasoned that unless all of these assumptions (sometimes called *postulates*) were valid, then gravitational forces would be unbalanced, resulting in the movement of stars (which were thought to be stationary at that time).

Of course, the idea of *infinity*, that there may be no end or edge to the universe is not easy for the human mind to comprehend, but imagining what would be beyond the edge of a finite universe is no easier.

But there is a big problem with this Newtonian model of the universe that many astronomers soon realized – if the universe is infinite and contains an infinite number of stars, there should be no such thing as a dark sky at night, because light from the stars should be arriving from all directions at all times. This is known as **Olbers' paradox**, named after one of the leading astronomers of the 19th century, Heinrich Wilhelm Olbers. (A *paradox* is an apparently true statement that seems to contradict itself. 'I always lie' is a widely quoted paradoxical statement, as is the fact that standing still can seem to be more tiring than walking.)

It is, however, good science to also provide mathematical evidence in support of the argument against the Newtonian model and for this we use the inverse square law.

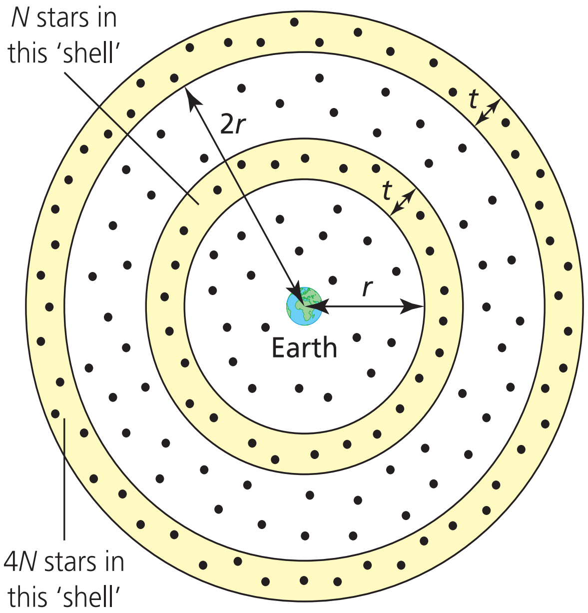


Figure 18.40 Stars in 'shells' around the Earth

Imagine stars in a *shell* of thickness, *t*, surrounding the Earth at a distance, *r* (Figure 18.40). Suppose that there are *N* stars in that shell. Now compare that with another shell of the same thickness, but at twice the distance from Earth (2*r*). The volume of the second shell will be four times the first because it has four times the area (surface area of a sphere $= 4\pi r^2$).

Assuming that stars are approximately evenly distributed (as in the Newtonian model) there will be 4*N* stars in the second shell. Using the inverse square law for luminosity, we know that the average brightness of light received on Earth from individual stars in the second shell will be a quarter of that from stars in the first shell because they are two times further away.

Therefore, the total light received on Earth from the two shells will be the same, because although the average light from individual stars in the second shell is a quarter of the light from the stars in the first shell, there are four times as many stars. Similar arguments can be made for any or all shells of equal thickness around the Earth and this means that stars from all distances should contribute equal intensities to the light received on Earth (assuming that nearby stars do not block the light from distant stars). With an infinite number of stars and shells there should be no such thing as a dark night sky – in fact the sky should be infinitely bright, which it definitely is not.

It was clear that either the reasoning given above and/or the Newtonian model of the universe needed changing or rejecting. Since the mid to late 1960s the **Big Bang model of the universe** has been widely accepted by astronomers and that has solved Olbers' paradox.

The Big Bang model

E.4.3 **Suggest** that the red-shift of light from galaxies indicates that the universe is expanding.
E.4.4 **Describe** both space and time as originating with the Big Bang.

When the absorption spectra from galaxies are examined it is found that they are mostly red-shifted. As explained earlier in this chapter, these red-shifts (Figure 18.24) are evidence that most galaxies are moving away (receding) from Earth.

It is important to realize that this is true for galaxies observed in all directions and would also be true for any observer viewing galaxies from *any* other location in the universe. Almost all galaxies are moving apart from all other galaxies. Our position on Earth is not unique, nor special, and we are not at the 'centre' of the universe. The universe does not have a centre.

The conclusion from these observations can only be that the universe is expanding.

Calculations using the measurements of the amount of red-shift also confirmed that the further away a galaxy is, the faster it is receding from us and other galaxies. This simple conclusion had very important implications. The faster a galaxy is moving, the further it has travelled. Observations first carried out by Edwin Hubble in the late 1920s on Cepheid variables confirmed that the motion of galaxies is consistent with all of them originating at the same place and the same time. An expanding model of the universe had been proposed a few years earlier by Georges Lemaître, followed by his hypothesis of a 'primeval atom', and was developed in the 1940s into what is now called the Big Bang model.

If radiation from a star or galaxy is observed to have a blue-shift, it is because it is moving towards Earth. This is not evidence against the Big Bang model because such an object is moving within a gravitationally bound system (a galaxy, a cluster of galaxies or a binary stars system) and, at the time of observation, it was moving towards the Earth faster than the system as a whole was moving away. For example, our neighbouring galaxy, Andromeda, exhibits a small blue-shift – it is moving towards us as part of its motion within our local group of galaxies, which is a gravitationally bound system.

In the Big Bang model, the universe was created at *a point* about 13.7 billion (1.37×10^{10}) years ago. At that time it was incredibly dense and hot, and ever since it has been expanding and cooling down.

The expansion of the universe is the *expansion of space itself* and it should *not* be imagined as similar to an explosion, with fragments flying into an existing space (void), like a bomb exploding. It may be helpful to visualize the expanding galaxies and space using marks on a very large rubber sheet to represent galaxies (so large that the edges cannot be seen) – if the sheet is stretched equally in all directions, all the marks move apart from each other. Of course, a model like this is limited to only two dimensions (Figure 18.41).

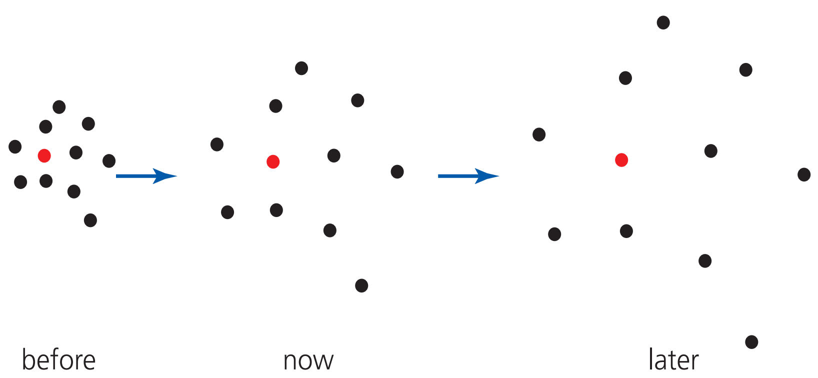


Figure 18.41 An expanding universe

It is very tempting to ask 'what happened before the Big Bang?' In one sense, this question may have no answer because the human concept of time is all about change and before the Big Bang there was nothing to change. The Big Bang should be considered as the creation of everything in our universe … matter, space *and* time.

Cosmic microwave background radiation (CMB)

E.4.5 **Describe** the discovery of cosmic microwave background (CMB) radiation by Penzias and Wilson.
E.4.6 **Explain** how cosmic radiation in the microwave region is consistent with the Big Bang model.

When it was first seriously proposed in the late 1940s, many astronomers were not convinced by the Big Bang model (they mostly preferred what was then known as the *Steady State Theory* of an unchanging universe). However, the discovery in 1964 by Penzias and Wilson of **cosmic microwave background (CMB)** radiation provided the evidence that confirmed the Big Bang model for most astronomers. Penzias and Wilson discovered that low-level microwave radiation can be detected coming (almost) equally from all directions, rather than from a specific source. (Later, important tiny variations were discovered in the CMB, a discovery which has vital implications for understanding the non-uniform structure of the universe and the formation of galaxies, as shown in Figure 18.42.)

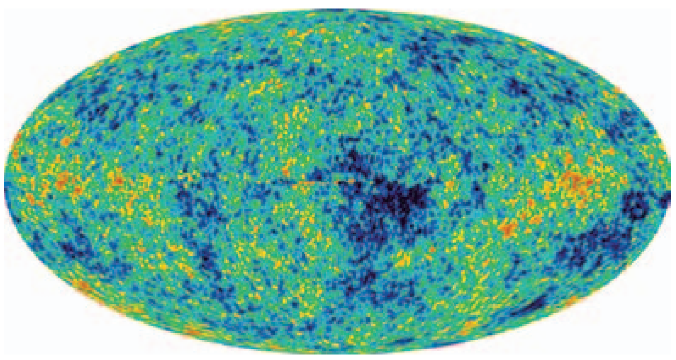


Figure 18.42 'Ripples in space' – a map of the whole sky at microwave wavelengths, showing the very small variations (1 part in 100 000) in the CMB radiation

We have seen before (Chapter 8) that everything emits electromagnetic radiation and that the range of wavelengths emitted depends on temperature. The Big Bang model predicted that the universe cools down as it expands and that the average temperature of the universe would now be about 2.7 K. Wien's law (for black bodies) can be used to calculate the peak wavelength associated with this temperature:

$$\lambda_{max} T = \text{constant}$$

$$\lambda_{max} = \frac{\text{constant}}{T} = \frac{2.9 \times 10^{-3}}{2.7} = 1.1 \times 10^{-3}\,\text{m}$$

Figure 18.43 shows the black-body radiation spectrum emitted from matter at 2.7 K. When this radiation was discovered coming almost equally from all directions and with a perfect black-body spectrum by Penzias and Wilson, the Big Bang model was confirmed.

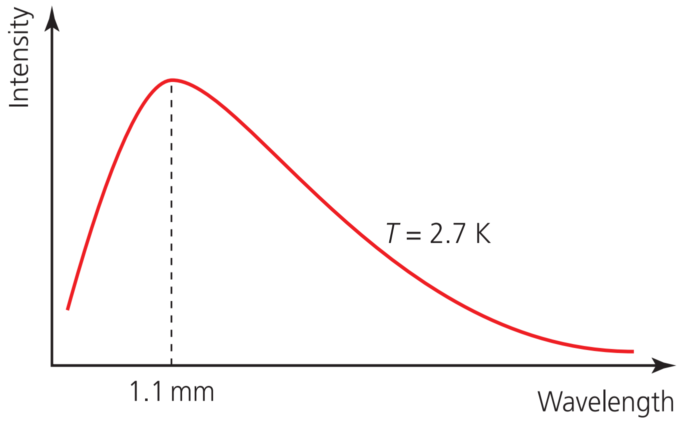


Figure 18.43 Spectral distribution for a temperature of 2.7 K

Resolving Olbers' paradox

E.4.7 Suggest how the Big Bang model provides a resolution to Olbers' paradox.

Olbers' paradox was finally resolved because Newton's postulates about a static, unchanging universe were no longer accepted. It now seemed that the universe could be finite, with a finite number of stars, each having a finite lifetime, thus limiting the amount of radiation that could reach Earth. More importantly, even if the universe was infinite, it was now known to have a definite age, which meant that the universe that was observable to us was limited to the distance light could travel in that time.

The universe that we can (in theory) observe from Earth is a sphere around us of radius 4.6×10^{10} ly. This is known as the **observable universe**, or the *visible universe*. (This distance is greater than 1.4×10^{10} ly because space has expanded since the Big Bang.) If there is anything further away, we cannot detect it because the radiation has not had enough time to reach us.

58 Summarize the two major discoveries which support the Big Bang model of the universe.

59 Astronomers look for 'shifts' in spectra as evidence for an expanding universe. The spectrum of which element is most commonly used, and why?

60 Draw a diagram to help explain why the light from some galaxies may be blue shifted.

61 A shift from 4.102×10^{-7} m to 4.237×10^{-7} m is observed from a particular galaxy.
- **a** Is this a red-shift or a blue-shift?
- **b** How big is the shift in frequency, Δf?
- **c** Use the equation $\Delta f = (v/c)f$ (from Chapter 11), where f is the frequency emitted by the galaxy, to calculate the speed, v, of the galaxy.

62 **a** How will the average temperature of the universe change in the future if:
- **i** the universe continues to expand
- **ii** the universe begins to contract?

b If a distant alien civilization near the edge of our (observable) universe detected CMB, what directions would the radiation be coming from and what temperature would they detect?

TOK Link: The development of scientific theories and paradigm shifts

A paradigm is a set of beliefs, or patterns of thought, with which individuals or societies organize their thinking about a particular issue, whether it is big or small. It is like a framework for all our thoughts and actions when, for example, we consider the place of human beings in the universe, or decide where to go on a holiday.

In scientific terms, a paradigm could be said to be a pattern of beliefs and practices that effectively define a particular branch of science at any period of time. An obvious example from this chapter would be the set of ideas associated with the, now discredited, belief that the Earth is at the centre of the universe and the various consequences of that fundamental idea.

The phrase *paradigm shift* has been used increasingly during the last 50 years since it was first popularized by Thomas Kuhn and others in the early 1960s. It is used especially with respect to developments in science. There are plenty of examples which suggest that, while scientific understanding, knowledge and practices obviously evolve

Figure 18.44 Ernest Rutherford was responsible for a significant paradigm shift in 20th century atomic physics

and, hopefully, improve over time, many of science and technology's greatest achievements have occurred following a relatively sudden (and perhaps unexpected or even seemingly unimportant) discovery or invention, or following the genius of an individual who has the insight to look at something in a completely new way. The phrase 'to think outside the box' has become very popular in recent years and it neatly summarizes an encouragement to look at a problem differently from the way others think about it (the 'box' being the paradigm).

A paradigm shift occurs when new insights, technology and discoveries have such a fundamental effect that current ideas or beliefs have to be rejected. Most individuals, organizations and societies find that a very difficult thing to do, even to the point where they strongly reject overwhelming evidence that their prevailing beliefs or actions are no longer reasonable. As discussed earlier in this chapter, the Roman Catholic Church's response to scientific evidence that the Earth revolved around the Sun was to simply ignore it and persecute those who held those beliefs.

The Big Bang model is another example of a paradigm shift in astronomical thinking and, if extra-terrestrial life were ever discovered, then most of us would look at ourselves in a completely new way – a tremendous paradigm shift! On a less profound scientific level, the technology of the Internet and the introduction of social networking sites are having such a dramatic effect on the way that many people interact, that they can be described as producing a paradigm shift in communication. Characteristically, there are many people who are unwilling to accept such changes in their lives and who believe that they are unnecessary, or even harmful.

Question

1 In what way did the work of Ernest Rutherford (Figure 18.44) radically change basic ideas in physics?

The development of the universe

Before trying to predict the future, we should consider briefly the past, and how we can estimate the age of the universe.

Worked example

10 Using red-shift observations, the speed of recession of a galaxy was measured to be $3800\,\text{km s}^{-1}$. If the galaxy is determined to be 54.1 Mpc away (using a Cepheid variable), show that an estimate of when it started moving is about 14 billion years ago. What important assumption did you make?

$$\text{time, } t = \frac{\text{distance}}{\text{speed}}$$

$$t = \frac{54.1 \times 10^6 \times 3.09 \times 10^{16}}{3.800 \times 10^6} \quad \text{(see Table 18.5 for conversion rates)}$$

$$t = 4.40 \times 10^{17}\,\text{s}$$

$$t = \frac{4.40 \times 10^{17}}{365 \times 24 \times 3600} = 1.39 \times 10^{10}\,\text{y}$$

Calculations like this will give approximately the same answer for any galaxy, and this is an *approximate* estimate for the age of the universe. To do this calculation, however, we needed to make an unjustified assumption – that the speed of the galaxy was always the same. If galaxies have always been moving with the same speed (as they are now), then the universe has always expanded at the same rate and, presumably, always will expand at the same rate. This could be represented graphically as shown in Figure 18.45.

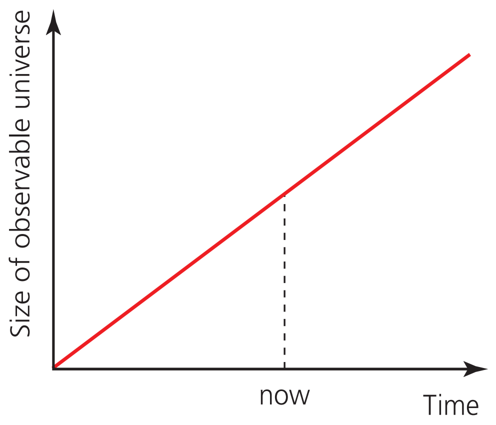


Figure 18.45 How the size of the universe would vary *if* each galaxy always moved with the same speed (which is not true!)

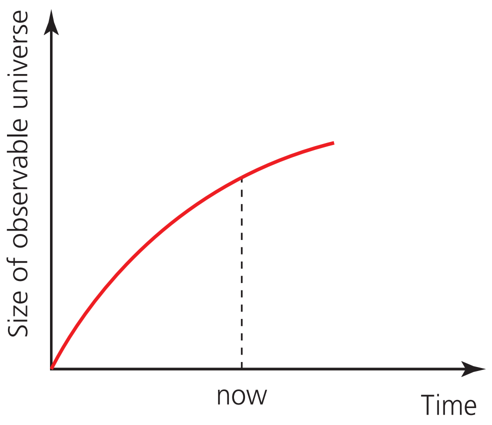


Figure 18.46 How the size of the universe would vary *if* the rate of expansion continued to decrease

This would mean that the future of the universe is predictable and understood, but that is very far from the truth! The universe has *not* been expanding at a constant rate partly because galaxies are moving against the force of gravity, so that it might seem more sensible to assume that they are slowing down – in just the same way as a stone thrown up in the air decelerates. If this were true an improved graph of the universe's expansion may be predicted to be as shown in Figure 18.46.

Is the universe open, closed or flat?

E.4.8 Distinguish between the terms open, flat and closed when used to describe the development of the universe.

It is convenient to reduce the possible future of the universe to three possibilities, which are represented in Figure 18.47, each dependent on the total mass of the universe.

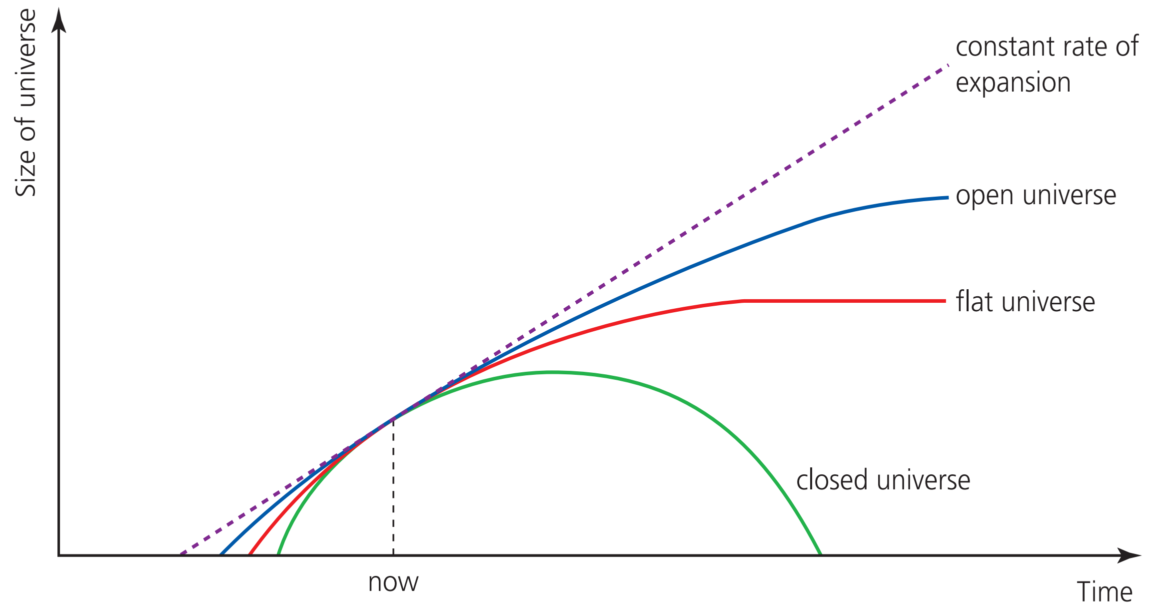


Figure 18.47 Possible futures of the universe

- In an **open universe,** expansion will continue forever because the total mass and hence the forces of gravity are not great enough to stop it, although they slow the rate of expansion.
- If the mass of the universe is large enough, the forces of gravity will be able to slow down and eventually stop expansion at some time in the future, after which the universe will then collapse back upon itself (the so-called *Big Crunch*). This would be called a **closed universe**.
- The third, mathematical, but seemingly unlikely possibility lies between the other two: in a **flat universe.** The mass of the universe would have *exactly* the right value to allow it to expand for ever, but to converge to a limit after an infinite time.

It should be clear that a key factor needed to predict the future of the universe, or understand how it developed in the past, is an estimate of its total mass. However, astronomers usually refer to the average *density* of the universe, rather than its mass (density, ρ = mass/volume).

Critical density

E.4.9 Define the term *critical density* by reference to a flat model of the development of the universe.
E.4.10 Discuss how the density of the universe determines the development of the universe.

The **critical density, ρ_c**, of the universe is the density that would result in a flat universe (as described above).

- If the average density of the universe, ρ, is greater than ρ_c ($\rho > \rho_c$) the universe will be closed.
- If $\rho = \rho_c$ the universe will be flat.
- If $\rho < \rho_c$ the universe will be open.

A value for the critical density of the universe can be obtained using Newton's law of gravitation and an estimate of the current rate of expansion. (This is *not* required by the IB Physics course, but the calculation is similar to escape speed calculations in Chapter 9.) Because the rate of expansion is not known with great accuracy, the critical density is also only an approximation.

The critical density has been estimated by astronomers to be about $10^{-26}\,\text{kg}\,\text{m}^{-3}$, which is equivalent to an average of about 6 hydrogen atoms in every cubic metre. Compare that to 10^{23}, which is the approximate number of air molecules per cubic metre near the Earth's surface. Clearly, most of the universe is just empty space!

Estimating the mass and density of the universe

E.4.11 Discuss problems associated with determining the density of the universe.

To help to predict the future of the universe, we need to determine the *actual average density* and then compare it to the *critical* density.

For a simple approximation:

average density of universe = average mass of galaxies × number of galaxies in a very large, known, volume

The mass of a galaxy can be estimated from observations of the rates of rotation of stars within the galaxy and knowledge of rotational dynamics. For example, using fairly straightforward physics, it is possible to deduce the mass of our own galaxy, the Milky Way, from the facts that it has a radius of about 50 000 ly and rotates with a period of about 2×10^8 y. Using these facts we can calculate that the Milky Way is about 10^{12} times the mass of our Sun. (This kind of calculation is *not* required by the IB Physics syllabus.)

But there are large uncertainties in these kinds of measurements, so that the average density of the observable universe is not known accurately.

Dark matter

There is also another problem – when the mass of a galaxy is estimated from the number of stars it contains multiplied by an estimate of their average mass, the result is much less than (about 17% of) the mass calculated from rotational dynamics. This can only be explained by the fact that there must be a lot of matter (about 83% of the total) that we are unable to observe. This is known collectively as **dark matter** and there have been many different suggestions about what it could be.

- **MACHOs** – Massive Astronomical Compact Halo Objects. These could be old stars or small 'failed' stars, very large planets or even black holes, any or all of which simply do not emit enough radiation for us to detect them.
- **WIMPs** – Weakly Interacting Massive Particles. There could be very large numbers of fundamental particles that we do not know about yet simply *because* they are very difficult to detect.
- **Neutrinos**. The masses of neutrinos are presumed to be very small, but they are still not known with any certainty. If neutrinos are present in the universe in large numbers they could contribute significant mass.

Current scientific evidence

E.4.12 State that current scientific evidence suggests that the universe is open.

Figure 18.48 Adam Riess, Saul Perlmutter and Brian Schmidt

Observations of distant supernovae in recent years have indicated that the rate of expansion of the universe has been *increasing* for several billion years, rather than the *decreasing* rate of expansion suggested by basic gravitational considerations. Work on this topic by three physicists, Perlmutter, Riess and Schmidt (Figure 18.48), was jointly awarded the Nobel Prize for physics in 2011.

Despite this, the very latest research suggests that the actual density of the universe is very close to the critical density.

If the universe currently has an increasing rate of expansion, a completely new explanation is needed. Much research is currently taking place into the possible existence of 'dark energy', which is a theoretical form of energy responsible for the force that must be overcoming gravity and making the universe expand. If confirmed, 'dark energy' would account for about 73% of all the energy–mass in the universe.

Dark matter, dark energy and the ultimate fate of the universe are the subject of intense research by astronomers and nuclear physicists. These are some of the biggest unsolved problems in science.

63 Use the Internet to find out what you can about 'dark energy' and how the idea can be used to help explain an increasing rate of expansion of the universe.

64 It has been estimated that there are about 8×10^{10} galaxies in the observable universe, each containing on average about 1.2×10^{11} stars and each star having an approximate average mass of 2.0×10^{30} kg. Use these figures to estimate the observable mass of the universe.

Astrophysics research

E.4.13 Discuss an example of the international nature of recent astrophysics research.
E.4.14 Evaluate arguments related to investing significant resources into researching the nature of the universe.

Astrophysics and particle physics research is very expensive. For example, the total final costs of the orbiting Hubble telescope and the Large Hadron Collider at CERN are each likely to be well over five billion US dollars. Many people say that this is not a sensible use of such large amounts of money (and other resources) and that the same money spent elsewhere could have very real benefits for humanity, for example, by improving medical care, sanitation and malnutrition in under-developed countries. Of course, similar arguments might be put forward about the resources used for many other areas of scientific research or human endeavour, but what makes astronomical research unique in this respect is its extreme costs and the absence of any obviously useful outcome. Expanding basic human knowledge about the universe is admirable and understandable, but is it a sufficient reason to spend a fortune?

For example, could society justify the expense of sending astronauts to Mars? In the 1960s and early 1970s many Americans questioned the costs of the manned trips to the Moon and they were stopped afterwards for that very reason. However, it is widely claimed that there were many very useful, but not directly intended, benefits of the lunar programme, including rapid developments in communication, computing, lasers and electronics which would otherwise not have happened (or not so quickly). Certainly, it is indisputable that we can never be sure where our scientific enquiries may lead us and, if we only ever start research projects which have definite and beneficial aims, we reduce the opportunity for unexpected discovery and developments.

One way of reducing the costs of research into astrophysics is to share ideas and information between countries and to make sure that the same expensive research is not being done in different places at the same time. Of course, modern communication makes this easy and there are many such examples of shared research, particularly within Europe and between Europe and the USA. Two planned major examples are:

- The European Space Agency (which has 19 member states) plans to send a mission in the year 2022 to explore the icy moons of Jupiter. This is known as the JUICE project and the mission will not arrive near Jupiter until 2030.
- The James Webb Space Telescope (JWST) is planned to be launched in 2018. It will be a successor to the Hubble telescope, aimed at studying the birth and evolution of galaxies and planets. It a joint project involving NASA and about 20 different countries, with an estimated cost of nearly 10 billion US dollars.

With such international collaboration and teamwork, it is now less likely that significant developments will be made by individuals working in isolation (as they have been in the past). However, issues such as self-interest, financial reward, personal and national pride, and independence will inevitably still be involved in research projects.

65 Write a short speech explaining why you think, or don't think, that your country (probably together with other countries) should spend a lot of money on astrophysics projects, for example, sending astronauts to Mars.

66 Use the Internet to obtain outline details of any one current or planned international astrophysics research project.

Additional Perspectives

Are we alone?

Many people are fascinated by the thought that there may be aliens (extra-terrestrials) in outer space. It has long been the subject of a large number of books and movies and a surprisingly large number of people seem to believe that aliens have visited Earth. It is, however, a serious subject for scientific research, although not one that many governments want to pay a lot of money for.

We might imagine that alien life-forms could be very different from life on Earth and exist in very different environments, but it does seem logical to start looking for evidence of extra-terrestrial life on planets such as our own. Planets are likely to be found around most of the countless billions of stars in the observable universe and, with recent developments in the use of telescopes, many possibilities have been found. (Planets in other star systems do not emit significant radiation and are therefore not easy to detect.)

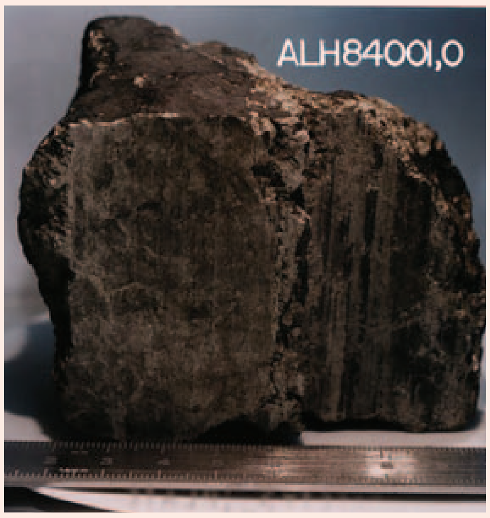


Figure 18.49 The meteorite from Mars with possible signs of life

In 1996, the whole world was excited to learn that a meteorite from Mars, discovered years before in Antarctica, may contain evidence of ancient microscopic bacteria (Figure 18.49). The President of the USA even announced it on television, but a lot of further examination has proved inconclusive.

Since the first radio broadcasts about 90 years ago, some of the electromagnetic waves that we use for various kinds of communication will have been (unintentionally) transmitted away from Earth and into space, travelling a distance of about 90 light years by now. In recent years, some attempts have been made to deliberately send signals to any extra-terrestrials that may or may not exist, but the waves will not have travelled 'far' into space yet. A signal received and then returned from a planet orbiting a 'nearby' star, for example, at a distance of 6 light years away from Earth, will take an overall time of 12 years to go and come back. A 'real-time' conversation with life on a (nearby) planet would seem to be impossible (even a radio signal to Mars takes at least four minutes to reach the planet).

Many astronomers believe that it is likely that we are not alone in the universe, simply because of the enormous number of possible planets and because the universe is old enough for life to have slowly developed elsewhere. But, if this is true, why haven't we received any evidence of their existence? (After all, there could be millions – or more – of other inhabited planets!)

Some people would say that they don't want to know about the possibility of aliens. The discovery of any extra-terrestrial life-form would change the way most of us think about ourselves, raise important issues for our religious and other beliefs and possibly bring a significant threat to our civilization, whether that would be intentional or accidental.

Questions

1 Find out how personal computers can be used to search for extra-terrestrial intelligence (SETI).
2 What is Fermi's paradox? Assuming that intelligent extra-terrestrial life exists, suggest some reasons why we have not been able to detect any evidence for it.
3 What kind of radiation could convince astronomers that they come from 'unnatural sources' (that is, alien life-forms)?
4 Do *you* believe in aliens? Why do so many people *want* to believe in aliens (when there is no scientifically convincing evidence)?

E5 Stellar processes and stellar evolution

All main sequence stars involve similar nuclear processes. We will now review those processes and then look in more detail at what happens to these stars when the supply of hydrogen in their core begins to run out, how new elements can be formed, and how the fate of any star depends on its mass.

Nucleosynthesis

E.5.1 **Describe** the conditions that initiate fusion in a star.

The average density of the universe is amazingly low and the space between stars is almost, but not completely, empty. The particles present in this 'empty' space are mostly hydrogen and helium with traces of heavier elements from earlier star systems. Where it is densest, this interstellar matter is often described simply as 'clouds' of gas and dust, often called nebulae (singular: **nebula** – see Figure 18.50).

Figure 18.50 The famous 'Pillars of Creation' in the Eagle Nebula

The gravitational forces between the individual particles are tiny, but if, for whatever reason, the total mass in a certain region increases, the accumulative gravitational effects can make the gas cloud very slowly collapse inwards. In this process gravitational potential energy is transferred to kinetic energy of the particles and, given a long enough time, the particles eventually gain enough energy (that is, if the temperature becomes high enough), that the hydrogen nuclei can overcome the electrostatic repulsion between their positive charges. This means that nuclear fusion to form helium can take place. This process was outlined at the beginning of this chapter and nuclear fusion was described in Chapter 7. In general, the formation of heavier nuclei from the fusion of lighter nuclei is called **nucleosynthesis**.

Stars contain separated ions and electrons rather than neutral atoms. This is the fourth state of matter, called **plasma**. When the temperature of a star becomes high enough, light will be emitted and we may refer to it as the 'birth' of a star. The gravitational pressure inwards will be balanced by the outwards thermal gas pressure and radiation pressure, and a star can remain stable for billions of years. These are the main sequence stars which comprise the diagonal line across the Hertzsprung–Russell diagram.

What happens to a star when the supply of hydrogen is reduced?

E.5.2 **State** the effect of a star's mass on the end product of nuclear fusion.
E.5.3 **Outline** the changes that take place in nucleosynthesis when a star leaves the main sequence and becomes a red giant.

Over a long period of time, the amount of hydrogen in the core of a star gets less so that eventually the outwards pressure is reduced and becomes less than the inwards gravitational pressure. This occurs when the mass of the core is about 12% of the star's total mass and there is plenty of hydrogen remaining in the outer layers of the star. The star begins to contract and gravitational energy is again transferred to kinetic energy of the particles (the temperature of the core rises even higher than before – to 10^8 K and higher). Now, in all but the very smallest stars, it is possible for the helium to fuse together to form carbon and possibly some larger nuclei, releasing more energy so that the star becomes more luminous. This results in the outer layers of the star expanding considerably and, therefore, cooling. So, the star has a hotter core but it has become larger and cooler on the surface, and its colour/spectral class changes. It is then known as a **red giant** (or if it is very large, a **red supergiant**) and it will leave the main sequence part of the HR diagram. Our Sun is about halfway through its time as a main sequence star and it will become a red giant in about seven billion years' time.

All main sequence stars follow predictable patterns but, the greater the mass of a star, the greater the gravitational energy and the greater the temperature when it collapses. This means that the red giants formed from more massive stars have the higher temperatures needed to fuse heavier elements.

Higher temperatures are needed for the nucleosynthesis of heavier elements because nuclei with more protons have a greater positive charge, so there are greater electric forces of repulsion between them.

- *If the initial mass of a star is less than approximately four times the mass of the Sun,* the temperature of its collapsing core will rise high enough for the fusion of helium to form carbon and oxygen (but not any heavier elements, Figure 18.51). The star will be known as a *red giant* during this phase of its lifetime. It will later lose its outer extended envelope, leaving behind a compact, dense core which will slowly cool to become a **white dwarf** star.

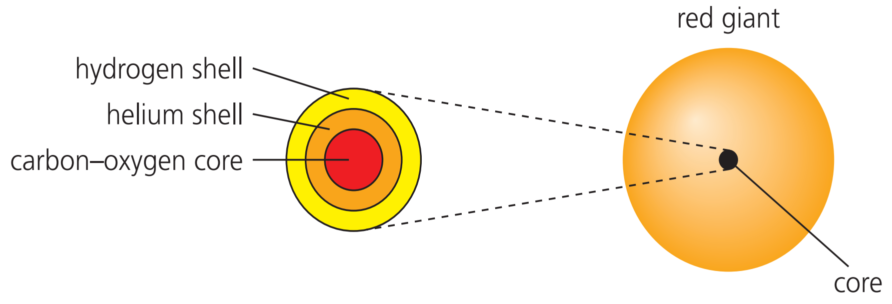


Figure 18.51 The core of a small red giant

- *If the initial mass of a star is between approximately four times and eight times the mass of the Sun,* the core of the *red giant* formed will have even greater temperatures, so that the nucleosynthesis of neon and magnesium can occur (from carbon and oxygen). This star will also later lose its envelope and become a white dwarf.
- *If the initial mass of a star is greater than approximately eight times the mass of the Sun,* the temperature will become great enough for the nucleosynthesis of even heavier elements like silicon, and finally iron (Figure 18.52). Because an iron nucleus has the highest binding energy per nucleon (see Chapter 7), it is the most stable and it is not possible to fuse even heavier elements in this way. At this stage of its life the star is known as a *red supergiant* and it will later experience a **supernova**. Elements heavier than iron are produced in supernovae.

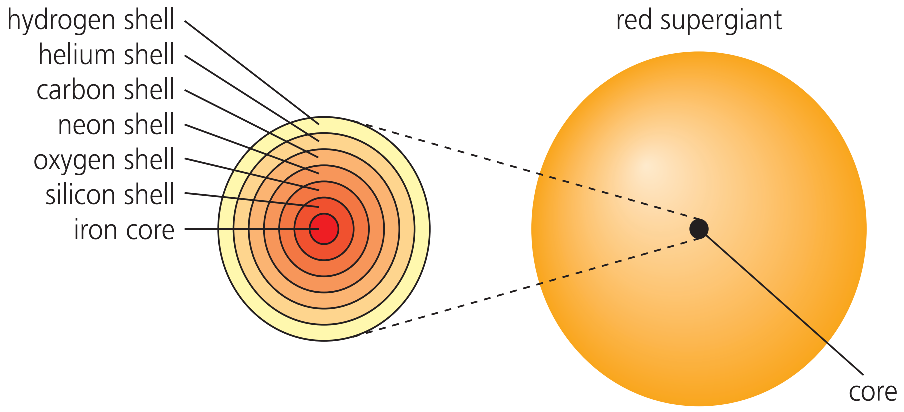


Figure 18.52 The core of a red supergiant

67 Explain why higher temperatures are needed for the fusion of heavier elements and why that only occurs in more massive stars.

68 Explain why iron forms at the core of a collapsing red supergiant star.

69 Where did all the atoms in your body originally come from?

70 Write a nuclear equation for any stellar nucleosynthesis.

Evolutionary paths of stars and stellar processes

Once a star has been 'born' its future can be entirely predicted depending upon its mass. We will now look in more detail at the connection between a star's mass and its luminosity and then discuss in more detail what happens to stars after they finish the major part of their 'lifetimes' as main sequence stars.

Mass–luminosity relation

E.5.4 **Apply** the mass–luminosity equation.

The greater the mass, m, of a main sequence star, the greater the rate of fusion and the greater its luminosity, L. Mass and luminosity are connected by the relationship shown below, which is derived from the basic principles of nuclear physics and confirmed by observations on binary stars for which we know both their luminosity and their mass. This equation cannot be used with stars which are not on the main sequence. Conversely, if the mass and luminosity of a star conform to this relationship, it can be assumed that the star is from the main sequence.

$$L \propto m^n \qquad (\text{where } 3 < n < 4)$$

This equation is given in the IB *Physics data booklet*.

The uncertainty in n arises because of uncertainties in observation and because the same single power law is not appropriate for all stars. The relationship is also represented by the graph in Figure 18.53.

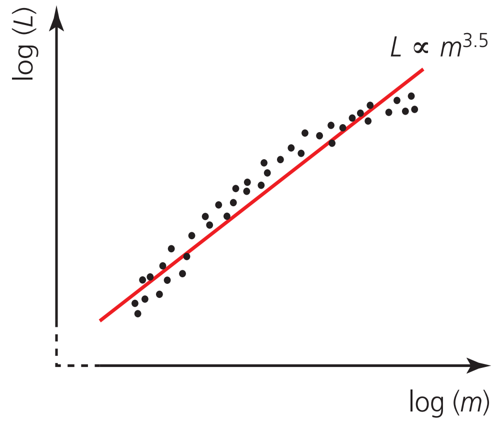


Figure 18.53 How the mass of a star affects its luminosity

Worked example

11 If a star has a mass four times the mass of our Sun, what does this equation predict for the luminosity of the star?

Luminosity is proportional to $m^{3.5}$, so the star will be approximately $4^{3.5} = 128$ times more luminous than the Sun, so its luminosity equals $128 \times (3.8 \times 10^{26}) = 4.9 \times 10^{28}$ W.

This equation is very useful because it can be used to calculate the mass of a main sequence star from its luminosity.

It should also be noted that this equation confirms that more massive stars have shorter lifetimes. For example, if star A had *twice* the mass of star B, calculations show that star A will have a luminosity about *11* times greater than star B. Star A may have twice the mass of hydrogen of star B, but it will use its supply up much quicker because the rate of fusion (as judged by its luminosity) is so much greater.

Looking again at the HR diagram (Figure 18.29), moving along the band of main sequence stars from lower right to upper left, the stars have greater luminosity, higher temperatures and shorter lifetimes simply *because* they are more massive.

The fates of red giants and red supergiants

E.5.6 **Compare** the fate of a red giant and a red supergiant.

When the nuclear reactions in the core of a *red giant* have finished, the star will contract again under the effects of gravity. The energy released will eject the outer layers as a **planetary nebula** (a misleading term because it has nothing to do with 'planets' as the term is usually used). Figure 1.1b on page 1 of the Student Book shows an example. The core which is left behind is much smaller than the original star and is called a *white dwarf*. A star can remain stable as a white dwarf for a long time owing to the behaviour of electrons within it, in a process known as **electron degeneracy pressure** (details are *not* needed for the IB Physics course).

The end of a *red supergiant* is much more complicated and much more violent. A shockwave travels quickly through the star resulting in a vast explosion of the outer layers, called a supernova, leaving behind a very dense core called a **neutron star**. A neutron star can remain stable for a long time because of **neutron degeneracy pressure** (details are *not* needed for the IB Physics course). But, if the core is really massive, this pressure is not great enough to stop the star collapsing further to become an incredibly dense **black hole** – so dense and massive that the escape velocity is greater than the speed of light, so that not even light can be emitted from a black hole.

Chandrasekhar and Oppenheimer–Volkoff limits

E.5.5 **Explain** how the Chandrasekhar and Oppenheimer–Volkoff limits are used to predict the fate of stars of different masses.

Once the mass of a main sequence star has been determined from its luminosity, the **Chandrasekhar** and **Oppenheimer–Volkoff limits** can be used to predict whether the star will become a red giant or a red supergiant and then a neutron star or a black hole.

- The Chandrasekhar limit is the maximum mass of a star that can become a stable white dwarf star.

This mass is approximately 1.4 times the mass of the Sun. A red giant of mass less than this limit will form a white dwarf. A red giant of mass larger than this limit is called a red supergiant and will end its life in a supernova.

- The Oppenheimer–Volkoff limit is the maximum mass of a star that can become a neutron star.

This mass is approximately three times the mass of the Sun. A red supergiant of mass less than this will end its life (after a supernova) as a neutron star. A red supergiant of larger mass will end its life (after a supernova) as a black hole. Figure 18.54 shows a simplified summary of how the masses of stars affect their fates.

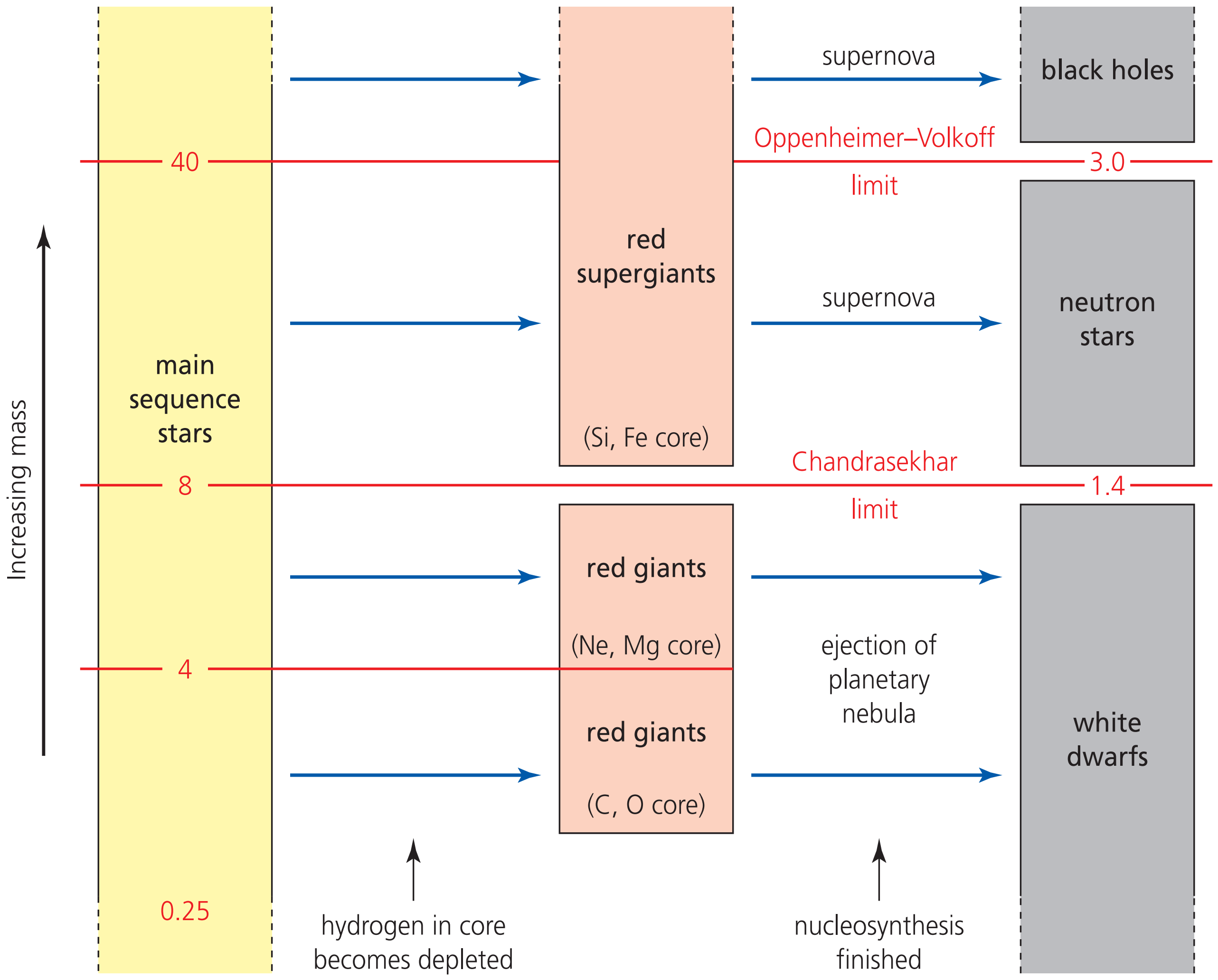


Figure 18.54 Evolution of stars of different masses (the numbers shown represent the approximate mass limits of the stars as multiples of the current mass of the Sun)

Evolutionary paths of stars on HR diagrams

E.5.7 Draw evolutionary paths of stars on an HR diagram.

When a main sequence star expands to a red giant, or a red supergiant, its luminosity and surface temperature change, and this, and subsequent changes over time, can be can be tracked on an HR diagram. It is known as a star's **evolutionary path**. Typical evolutionary paths of low mass and high mass stars are shown in Figure 18.55.

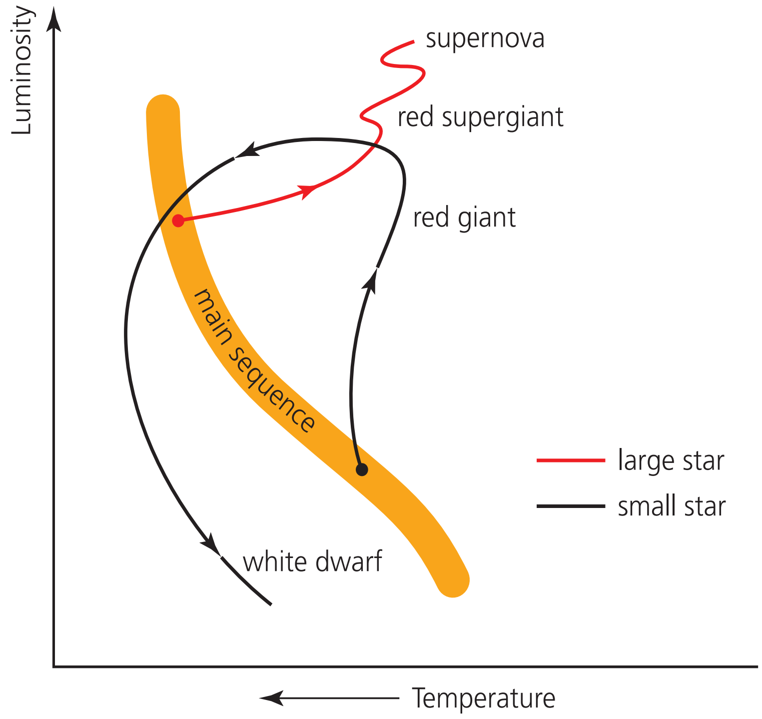


Figure 18.55 Evolutionary paths of stars after they leave the main sequence

Pulsars

E.5.8 Outline the characteristics of pulsars.

In 1967, a new kind of star was observed, one from which the radiation received varied at very regular intervals, and those intervals were unusually small. An event occurring with a period of, say, one second, is exceptionally quick on the astronomical scale. These pulses of radiation led to the stars being called **pulsars** (a shortened version of pulsating star). When they were first discovered by Jocelyn Bell-Burnell (Figure 18.56), the cause of the pulses was unknown and some astronomers even wondered briefly if such regular and precise bursts of radiation were a sign of intelligent life elsewhere in the universe.

It is now believed that pulsars are neutron stars emitting radiation, not as pulses, but continuously as a beam which rotates around, like the beam from a lighthouse (Figure 18.57).

Figure 18.56 Jocelyn Bell-Burnell was the first to discover pulsars in 1967

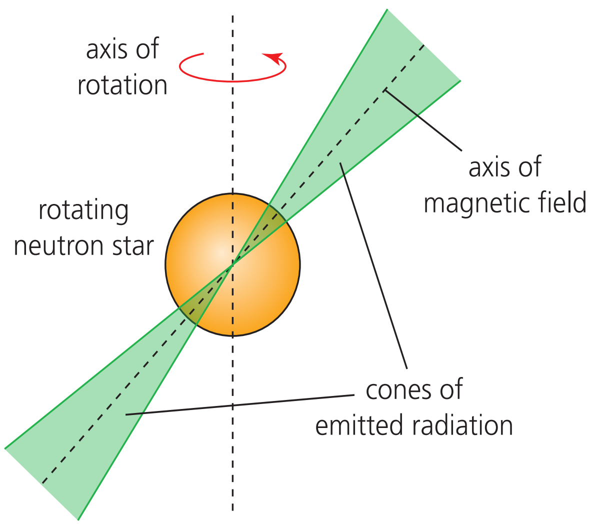


Figure 18.57 Radiation emitted from a rotating pulsar

Stars rotate because they were formed from rotating gas and dust clouds. When a star contracts, its speed of rotation will increase (because its angular momentum is conserved – this theory is not needed for the IB Physics course). Unlike other stars, the mass of a neutron star is so highly compressed that it will rotate at a surprisingly quick rate, with a period of typically less than a second.

The enormously compressed nature of a neutron star explains why they rotate fast enough for their periods to be so short and it also helps to explain the origin of the radiation – charged particles are accelerated in their intense magnetic fields. The beam of radiation rotates because the axis of rotation of the pulsar is not the same as the axis of the magnetic field. (In a similar way the Earth's magnetic poles are not aligned with its geographic poles – which are on the axis of rotation.)

71 A main sequence star has a luminosity which is half of the value of our Sun.
- **a** Is it more massive or less massive than the Sun (2.0×10^{30} kg)?
- **b** Estimate its mass in kilograms.
- **c** Will its lifetime be longer or shorter than that of the Sun? Explain your answer.

72 Sketch the evolutionary path of our Sun on an HR diagram.

73 After a supernova, a black hole or neutron star will be formed. Why can't we show these on an HR diagram?

74 If no radiation is emitted from black holes, how can we know where they are?

75
- **a** Confirm that if star A has twice the mass of star B (both main sequence stars), it will have about 11 times the luminosity.
- **b** Assuming (for simplicity) that the lifetime of a star is proportional to its mass and inversely proportional to its luminosity, what would be the ratio of the main sequence 'lifetimes' of star A and star B?

76 Explain why the Chandrasekhar limit is such an important number in astronomy.

77 A pulsar has a period of 1.36 ms. If it is a neutron star with a diameter of 21 km and mass of 5.1×10^{30} kg, calculate:
- **a** the speed of a point on its equator
- **b** its average density.
- **c** Compare these values with the equivalent values for the Earth. (Use the Internet to research relevant data.)

78 Use the Internet to learn more about electron and neutron degeneracy pressures.

E6 Galaxies and the expanding universe

In this section we will look at calculations to find the recession speeds of galaxies and, consequently, the rate of expansion of the universe. This enables us to make a prediction of the age of the universe.

Galactic motion

A galaxy is a group of billions of stars drawn together by gravity. They are commonly described by their shape as being *spiral* (Figure 18.58), *elliptical* or *irregular*. Our galaxy, the Milky Way, is a spiral galaxy.

Figure 18.58 Spiral Galaxy M81

Clusters of galaxies

E.6.1 Describe the distribution of galaxies in the universe.

Galaxies are distributed throughout space, but not in a completely random way. For example, the Milky Way is one of a group of about 20 galaxies known as the 'Local Group'. Larger groups of galaxies, called **clusters**, are bound together by gravitational forces (see Figure 18.59 for an example). Clusters may contain thousands of galaxies and much intergalactic gas and undetected 'dark matter'. (The term galactic cluster is commonly used for a group of stars within a galaxy.)

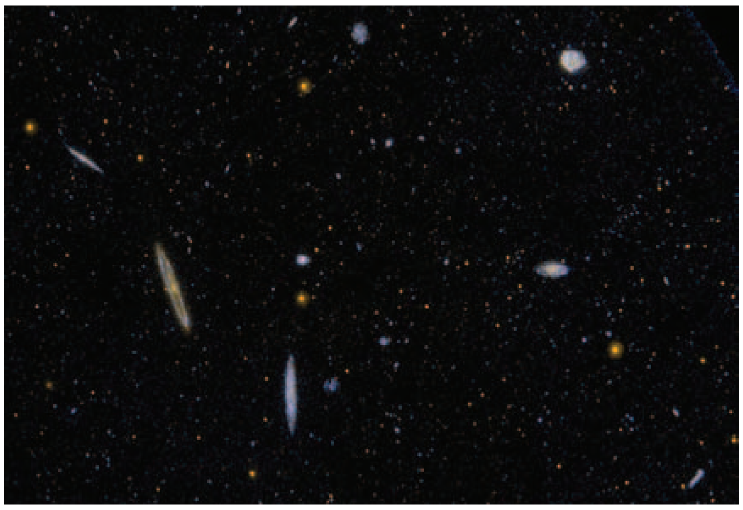


Figure 18.59 Virgo cluster of galaxies

Clusters of galaxies are not distributed evenly throughout space but are themselves grouped together in what are known as **superclusters**. Superclusters of galaxies may be the largest 'structures' in the universe.

Using red-shift to determine recession speed of galaxies

E.6.2 Explain the red-shift of light from distant galaxies.

As explained earlier in this chapter, radiation from receding galaxies is red-shifted and this observation led to the conclusion that the universe is expanding. The expansion of space itself provides the explanation for the red-shift – radiation emitted with a certain wavelength from a galaxy a long time ago will now be received here on Earth with a 'stretched' wavelength.

The following equation can be used to calculate the recession speed of a galaxy from its red-shift.

$$\frac{\Delta\lambda}{\lambda} \approx \frac{v}{c}$$

(This equation is listed in the IB *Physics data booklet*, and it is similar to the Doppler effect equation for frequencies, used in Chapter 11.)

λ is the original wavelength, $\Delta\lambda$ is the red-shift (the increase in wavelength) and v is the recession speed (c = speed of light = $3.00 \times 10^8\,\text{m s}^{-1}$).

Worked example

E.6.3 Solve problems involving red-shift and the recession speed of galaxies.

12 A line in the hydrogen spectrum has wavelength of 4.34×10^{-7} m. When detected on Earth from a distant galaxy, the same line has a wavelength of 4.76×10^{-7} m. What is the recession speed of the galaxy?

$$\Delta\lambda = (4.76 \times 10^{-7}) - (4.34 \times 10^{-7}) = 4.2 \times 10^{-8}\,\text{m}$$

$$\frac{\Delta\lambda}{\lambda} \approx \frac{v}{c}$$

$$\frac{4.2 \times 10^{-8}}{4.34 \times 10^{-7}} = \frac{v}{3.00 \times 10^8}$$

$$v = 2.90 \times 10^7\,\text{m s}^{-1}$$

79 What is the recession speed of a galaxy (km h^{-1}) if radiation of original wavelength 6.5×10^{-7} m undergoes a red-shift of 3.7×10^{-8} m?

80 A star receding at a velocity of $9.2 \times 10^3\,\text{km s}^{-1}$ emits radiation of wavelength 410 nm. What is the value of the red-shift of this radiation when it is received on Earth and what is its received wavelength?

81 Hydrogen emits radiation of frequency 6.17×10^{16} Hz. What frequency will be detected on Earth from a galaxy moving away at $1.47 \times 10^7\,\text{m s}^{-1}$?

82 Only a very tiny percentage of galaxies are moving towards us. Research the blue-shift of the Andromeda galaxy, one of the galaxies in the Local Group.

Hubble's law

E.6.4 State Hubble's law.
E.6.5 Discuss the limitations of Hubble's law.
E.6.6 Explain how the Hubble constant may be determined.

In the mid 1920s, the American astronomer Edwin Hubble compared information about the velocities of recession of relatively nearby galaxies (obtained from the red-shift of the light received) with the distances of the galaxies from Earth that were determined by using Cepheid variables within the galaxies. By 1929 Hubble had gathered enough data to publish a famous graph of his results for Cepheids within distances of a few Mpc from Earth. Figure 18.60 shows more results and for greater distances.

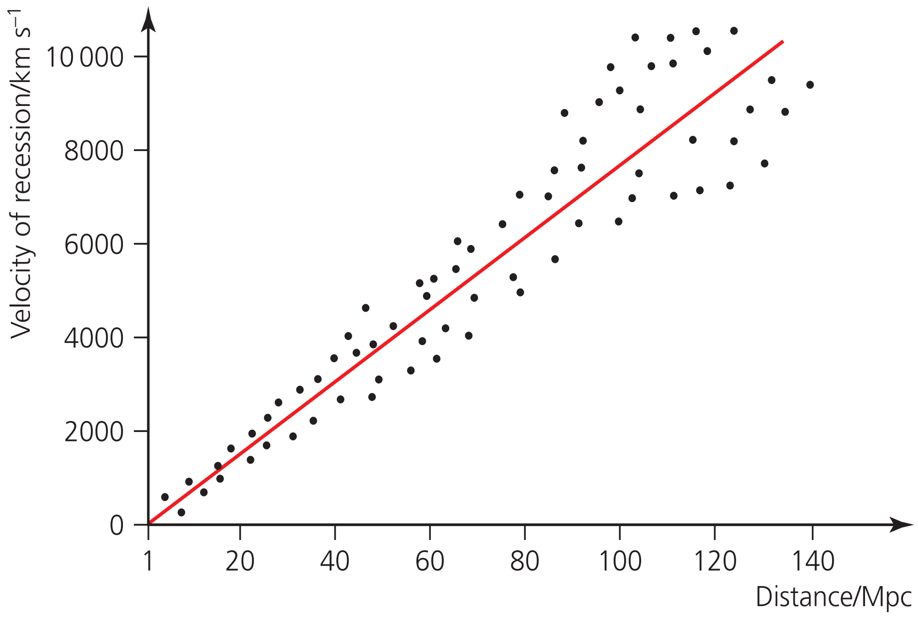


Figure 18.60 Variation of recession velocities of galaxies with their distances from Earth

Even today there are significant uncertainties in the data represented on this graph (although error bars are not shown on Figure 18.60). These uncertainties are mainly because the precise measurement of distances to galaxies is difficult, but also because galaxies move within their clusters. Nevertheless, the general trend is very clear and was first expressed in **Hubble's law**.

The current velocity of recession (the speed at which a galaxy appears to be moving directly away from Earth), v, of a galaxy is proportional to its distance away, d, (from Earth).

This can be written as:

$v = H_0 d$ This equation is listed in the IB *Physics data booklet*.

H_0 is the gradient of the graph and is known as the **Hubble constant**. Because of the uncertainties in the points on the graph, the Hubble constant is also not known accurately. The currently accepted value is about $74\,\text{km s}^{-1}\,\text{Mpc}^{-1}$ ($\text{km s}^{-1}\,\text{Mpc}^{-1}$ is more widely used than the SI unit, s^{-1}). In 2011, data from the Hubble telescope (appropriately) produced a value of $H_0 = 73.8 \pm 2.4\,\text{km s}^{-1}\,\text{Mpc}^{-1}$, but alternative methods have resulted in different values.

Hubble's law can be applied to the radiation received from all galaxies which are moving free of significant 'local' gravitational forces from other galaxies. That is, the law can be used for isolated galaxies or clusters (considered as one object), but less accurately for individual galaxies moving randomly within a cluster because the resultant velocity of an individual galaxy is the combination of its velocity with respect to the cluster and the recession velocity of the cluster as whole. A few galaxies even have a resultant velocity *towards* Earth at this time, and radiation received from such galaxies is blue-shifted (see question **82**).

The use of the Hubble constant with the recessional speeds of distant galaxies provides astronomers with a way to calculate the distance to galaxies which are too far away to use alternative methods.

Age of the universe

E.6.7 Explain how the Hubble constant may be used to estimate the age of the universe.

Figure 18.60 shows that the current velocity of recession of a galaxy is proportional to its distance away. From this we can make the very important conclusion that at some time in the past all galaxies originated at the same point in a 'Big Bang' (see Section E4). Hubble's constant can be used to obtain an estimate of when this happened (the age of the universe), although the uncertainty in the result is large, as shown in Worked example **14**.

Worked example

E.6.8 Solve problems involving Hubble's law.

13 Estimate the gradient of the graph in Figure 18.60 and compare it with the Hubble constant.

$$\text{gradient, } H_0 = \frac{9000}{120} = 75\,\text{km s}^{-1}\,\text{Mpc}^{-1}$$

This value is close to the accepted value of the Hubble constant (within 2% and within the range quoted).

14 Estimate the age of the universe in years using the Hubble constant of value $74\,\text{km s}^{-1}\,\text{Mpc}^{-1}$.

$$\text{time, } t, = \frac{\text{distance}}{\text{velocity}} = \frac{1}{H_0} \quad \text{(the reciprocal of the gradient of the graph)}$$

The value of H_0 in SI units is $\dfrac{74 \times 10^3}{3.26 \times 10^6 \times 9.46 \times 10^{15}} = 2.4 \times 10^{-18}\,\text{s}^{-1}$

So, $t = \dfrac{1}{H_0} = \dfrac{1}{(2.4 \times 10^{-18})} = 4.2 \times 10^{17}\,\text{s}$

Converting this to years gives $t = \dfrac{4.2 \times 10^{17}}{365 \times 24 \times 3600} = 1.3 \times 10^{10}\,\text{y}$

It is very important to realize that uncertainties in the data mean that the age of the universe calculated using this method is just an estimate. The calculation also wrongly assumes that the galaxies have always had the same velocities, ignoring both the assumed deceleration caused by gravitational attraction and the recently discovered accelerating rate of expansion.

Nevertheless, the latest estimate of the age of the universe, 13.7 ± 0.1 billion years (found from measurements of the cosmic microwave background (CMB) radiation), is close to the value calculated using Hubble's law (Worked example **14**).

History of the universe

E.6.9 Explain how the expansion of the universe made possible the formation of light nuclei and atoms.

Understanding the development of the early universe is very complex and the details are certainly not needed for this course, but most astronomers now agree on theories which explain events as early as 10^{-43} s after the Big Bang, when the temperature was about 10^{32} K. Figure 18.61 outlines some later history of the universe as it expanded and cooled, showing the approximate times and temperatures at which particle energies had reduced sufficiently for them to combine.

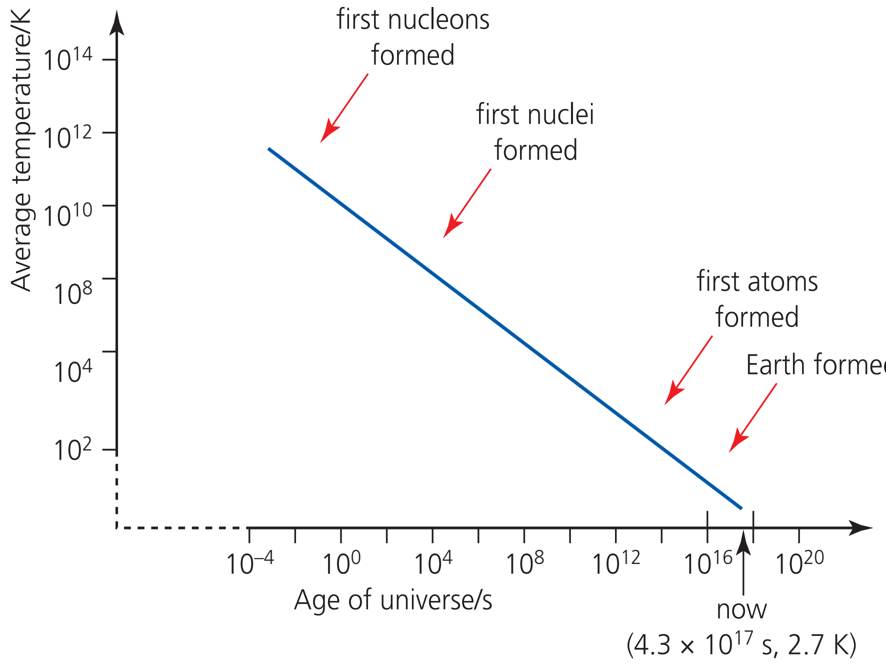


Figure 18.61 How the universe has cooled down (note that both scales are logarithmic)

83 What is the recession speed of a galaxy which is 25 Mpc from Earth?

84 How far away is a galaxy which is travelling at 1% of the speed of light?

85 Galaxy A is a distance of 76 Mpc from Earth and receding at a velocity of $5500\,km\,s^{-1}$. Another galaxy, B, is receding at $7300\,km\,s^{-1}$. Without using a value for H_0, estimate the distance to galaxy B.

86 The value of H_0 is linked to the universal gravitation constant and the critical density of the universe.
a Convert $74\,km\,s^{-1}\,Mpc^{-1}$ to SI units and substitute it into the equation $\rho_0 = 3H_0^2/8\pi G$ to determine a value for the critical density of the universe, ρ_0.
b Approximately how many nucleons per cubic metre is that density equivalent to?

87 A spectral line of normal wavelength $3.9 \times 10^{-7}\,m$ is shifted to $4.4 \times 10^{-7}\,m$ when it is received from a certain distant galaxy.
a How fast is the galaxy receding?
b How far away is it?

88 Explain why there were no atoms in the universe until about 10^5 years after the Big Bang.

89 The very early universe had almost equal amounts of matter and antimatter. Research into why there is no significant amount of antimatter in the universe now.

90 Find out what is meant by the 'cosmological principle'.

SUMMARY OF KNOWLEDGE

E1 Introduction to the universe

- Our Sun is a star around which planets, comets and asteroids orbit in elliptical paths because of the effects of gravity. Moons orbit some planets.
- The inner planets are solid and the outer planets are mainly gaseous. Their approximate relative sizes and distances from the Sun (relative to Earth) should be known. The further a planet is from the Sun, the longer its period.
- Comets are relatively small lumps of rock and ice which have very eccentric orbits and long periods. Asteroids are also rocky but much smaller than planets. They are mostly located in a 'belt' between the orbits of Mars and Jupiter.
- Stars appear as points of whitish light in the night sky and appear to stay in fixed (imagined) patterns which we call constellations. The stars in constellations should not be assumed to be 'close' to each other in space. Our view of the star patterns changes during the night as the Earth rotates. Different places on Earth will see different parts of the star patterns, and our view also changes during the year as the Earth moves around the Sun.
- Some groups of stars of common origin are actually relatively close together in space because of gravitational attraction. These are called stellar clusters.
- Groups of billion of stars orbiting a common centre of mass are called galaxies.
- The brightness of any star as seen from Earth depends on the power of the radiation that it emits (its luminosity) and its distance away. We cannot simply assume that brighter stars are more luminous, or closer.
- Astronomers use several different units for measuring distance. The light year, ly, is defined as the distance travelled by light in vacuum in one year. The astronomical unit (AU) is equal to the mean distance between the Earth and the Sun. The parsec (pc) = 3.26 ly.
- The order of magnitude of the diameter of the observable universe is 10^{11} ly. A typical galaxy has a diameter of about 10^4 ly and the distance between galaxies is typically 10^6 ly.

E2 Stellar radiation and stellar types

- Over a long period of time gravity pulls together the dust and gases in interstellar space. Eventually hydrogen atoms gain enough kinetic energy (that is, the temperature becomes extremely high – millions of kelvin) to overcome the electric repulsion between positive charges. Fusion can take place, energy is emitted and the star is 'born'.
- Nuclear fusion is the dominant energy transformation in all stars in the universe.
- For main sequence stars the principal fusion process can be summarized as: $4\,^1_1\text{H} = \,^4_2\text{He} + 2\,^0_1\text{e}$ + other particles/radiations (neutrinos and gamma ray photons).
- The thermal gas pressure and the emitted radiation pressure outwards oppose the gravitational pressure inwards. These pressures will remain equal and opposite for a very long time, during which the star will remain about the same size and stable. During this period the star is known as a main sequence star.
- The luminosity, L, of a star is defined as the total power (W) that it radiates.
- The apparent brightness, b, of a star is defined as the intensity (W m^{-2}) received on Earth, $b = L/4\pi d^2$.
- Stars are almost 'perfect' emitters of radiation (black bodies). Intensity–wavelength graphs are used to represent the continuous spectra emitted at different temperatures.
- Wien's (displacement) law provides a direct link between the surface temperature of a star and the wavelength at which the maximum intensity of radiation is received: $\lambda_{max}T$ = constant.
- The Stefan–Boltzmann law is used to relate the luminosity of a star to its surface area and surface temperature (K): $L = \sigma AT^4$. σ is the Stefan–Boltzmann constant.
- Atoms in the outer layers of a star absorb some of the radiation emitted from the core. The absorption (line) spectrum detected on Earth can be used to identify the elements present in the star.
- The frequencies of the absorption lines in the spectrum from a galaxy (or star) moving at very high speed away (receding) from Earth are lower than frequencies from the same source on Earth. This is an example of the Doppler effect, and the change of frequency is called a red-shift. The size of the 'shift' can be used to calculate the recession speed of the galaxy.

- A few galaxies are moving towards Earth and radiation from them exhibits a blue-shift.
- Stars are classified by astronomers into different spectral classes, depending on their colour, which itself depends on surface temperature. Starting with the hottest, the classes are: OBAFGKM.
- White dwarf stars are relatively hot and therefore blue-white in colour, but their luminosity is relatively low because they are small in size.
- Red giant stars and red supergiants are relatively cool and therefore yellow-red in colour. However, they have a higher luminosity than many stars because they are large.
- Cepheid variables are an unusual type of star because their luminosities vary.
- It is common for two stars to orbit each other (or, more precisely, orbit their common centre of mass). This is called a binary star system. Observations of this kind of system are very important in determining the mass of stars using calculations which involve the stars' separation and their orbital period.
- Most binary stars appear to be a single star because they are relatively close together, but it is still possible to know that there are two stars by close examination of the light received. Spectroscopic binaries are detected because of the different shifts (red- and blue-shifts) from the two stars. Eclipsing binaries are detected because the intensity of the received radiation changes.
- The Hertzsprung–Russell (HR) diagram is a common way of representing many different stars on the same chart. The (logarithmic) axes of the diagram are luminosity and temperature (reversed), although the diagram may alternatively be drawn with magnitude and spectral class for the axes.
- The majority of stars are in the main part of their 'lifetimes' and are located somewhere along a diagonal line from top left to bottom right of the HR diagram. This is called the main sequence. The only major difference between these stars is their mass – which results in different luminosities and temperatures because of the different rates of fusion.
- The other types of stars can be located in other parts of the HR diagram.

E3 Stellar distances

- The distance to nearby stars can be calculated from a measurement of the parallax angle between their apparent positions (against a background of more distant, fixed stars) at two times separated by six months.
- One parsec is defined as the distance to a star which has a parallax angle of one arc-second.
- For stars further away than a few hundred parsecs the stellar parallax method is not possible because the parallax angle is too small to measure.
- The method of spectroscopic parallax can be used for distances up to a few thousand parsecs. Wien's law is used to find the surface temperature from measurements of the star's spectrum, then the HR diagram can be used to determine its luminosity. Finally, $b = L/4\pi d^2$ can be used to determine the distance.
- For distances up to a few million parsecs, astronomers use Cepheid variables. The outer layers of these stars expand and contract at a regular rate. There is a close relationship (shown graphically) between the period of the resulting luminosity variations and the star's luminosity. Again, $b = L/4\pi d^2$ can be used to determine the distance once luminosity has been determined.
- If the distance to a Cepheid variable is determined, it can then be assumed that the galaxy it is in is also at the same distance. In this way, Cepheid variables are known as 'standard candles'.
- To determine distance to galaxies even further away, red-shift measurements or observations on supernovae are used.
- Astronomers use the apparent magnitude (m) scale to describe the apparent brightness of a star as observed on Earth. This is a reversed, logarithmic scale, originally based on giving the brightest stars a magnitude of 1, and the dimmest visible stars a magnitude of 6. A difference in apparent magnitude of 1 corresponds to a difference in apparent brightness of a factor of $\times 2.512$.

- The absolute magnitude, M, of a star is defined as the apparent magnitude it would have if it was observed from a standard distance of 10 pc. It is the equivalent of luminosity.
- Apparent and absolute magnitudes are connected by the following equation: $m - M = 5 \lg (d/10)$, where d is the distance to the star in parsecs.

E4 Cosmology

- In the Newtonian model, the universe was infinitely large and unchanging (static). In this model, the universe had always been the way it is, and always would be the same. In Newton's model, on the large scale, the universe was uniform and more or less the same everywhere.
- But if Newton's model was true, there would be an infinite amount of light, and certainly no night time. This is known as Olbers' paradox, and it can be verified using the inverse square law of luminosity.
- The Big Bang model of the universe resolved Olbers' paradox. In the Big Bang model the universe began at a point about 14 billion years ago. The Big Bang created space and time. What we are able to see of the universe (the 'observable universe') is limited by how far light can travel in 14 billion years.
- The evidence for the Big Bang comes from: (i) red-shift measurements, (ii) cosmic microwave background radiation.
- Radiation from (nearly) all galaxies (and stars) is red-shifted. This implies that the galaxies are all moving apart from each other and that at an earlier time they were closer together, beginning at a point. The universe is expanding, but not into a void like an explosion. Space itself is expanding.
- As the universe expands its average temperature falls. The Big Bang model predicts that the average temperature of the universe after 14 billion years should be 2.7 K. Black-body radiation corresponding this temperature was found to be coming almost equally from all directions by Penzias and Wilson. This was the conclusive evidence that confirmed the Big Bang model.
- The universe is still expanding but the force of gravity should slow down the rate of expansion. The future fate of the universe depends on how much mass it contains – which determines the size of gravitational forces. Astronomers refer to the average density of the universe.
- If the average density of the universe is less than a certain value, called the critical density, at some time in the future the universe will begin to collapse back inwards. This would be called a closed universe.
- If the average density is exactly equal to the critical density the universe will converge to a limit and stay at that value. This would be called a flat universe.
- If the average density is less that the critical density, the universe will continue to expand for ever. This would be called an open universe.
- The three possibilities for the fate of the universe are commonly represented graphically on a plot of the size of the universe against time.
- Current evidence suggests that the universe is open, but that its density is close to the critical value. The latest research has revealed that the rate of expansion is actually increasing and the concept of dark energy has been introduced to explain this.
- The average mass and density of the universe can be obtained by measurements on the rotation of galaxies. However the results of these calculations produce results far greater than the mass of galaxies determined by adding up the masses of the stars that can be seen. Therefore, most of the mass of galaxies is not accounted for, and is known as dark matter. Much on-going research is directed at determining the nature of the 'missing' mass, including possibilities such as MACHOs and WIMPs.
- Astrophysics research is very expensive and has uncertain practical benefits. There is a wide range of ethical and economical arguments for and against such projects.
- The JUICE Project and the James Webb Space Telescope are two major future projects currently being planned and which involve international co-operation, with the sharing of costs and experience.

E5 Stellar processes and stellar evolution (Higher Level only)

- The formation of heavier elements by the process of nuclear fusion is called nucleosynthesis. In main sequence stars hydrogen nuclei undergo fusion to form helium, with the release of energy.
- Heavier nuclei can only be formed at higher temperatures because the fusing nuclei need greater kinetic energies to overcome the increased electric repulsive forces that exist between nuclei with more protons. Greater temperatures occur in the latter stages of massive stars.
- When the hydrogen in the core of a main sequence star begins to be used up, the outwards thermal gas pressure and radiation pressure are reduced and the star begins to collapse because of gravity. The core heats up and the fusion of some heavier elements may now occur. The star then expands and becomes a red giant (or a red supergiant). What happens next depends on the mass of the star.
- If the initial mass of the star is less than about four times the mass of the Sun, only carbon and oxygen can be formed in the core. When nucleosynthesis is finished, the red giant will eject a planetary nebula and collapse to form a white dwarf star. It can remain in this form for a long time because of electron degeneracy pressure.
- If the initial mass of the star is between about four times and eight times the mass of the Sun, the greater temperatures in the core will enable the fusion of neon and magnesium. This red giant will also emit a planetary nebula and collapse to a white dwarf.
- The Chandrasekhar limit is the maximum mass of a white dwarf star (about 1.4 times the mass of Sun) – formed from a star of initial mass about eight times the mass of the Sun.
- If the initial mass of a star is greater than approximately eight times the mass of the Sun, it will become a red supergiant and the core temperatures will be high enough for the fusion of silicon and, finally, iron.
- When nucleosynthesis is finished in a red supergiant, it will undergo a supernova. If the mass at that time is greater than about three times the mass of Sun, the result will be a black hole. This is called the Oppenheimer–Volkoff limit.
- A red supergiant of smaller mass than this limit will end as a neutron star. A neutron star can remain stable for a long time because of neutron degeneracy pressure.
- These changes in the nature of stars can be drawn on HR diagrams. They are known as the evolutionary paths of the stars.
- To know the fate of a main sequence star, we need to know its mass. Stellar mass is related to luminosity by the following equation: $L \propto m^n$ (where $3 < n < 4$).
- A pulsar is a neutron star which rotates quickly and has a very strong magnetic field. Emitted radiation is confined to narrow cones and these rotate with the star. In this way the radiation reaching Earth arrives as pulses.

E6 Galaxies and the expanding universe (Higher Level only)

- Galaxies are attracted to other galaxies because of gravitational forces. This results in groups of galaxies called clusters.
- Clusters of galaxies are not distributed randomly throughout space, but are found in groups called superclusters.
- The recession velocities of galaxies can be calculated from the Doppler effect equation: $\Delta\lambda/\lambda \approx v/c$.
- Hubble's law states that the recession velocity of a galaxy is proportional to its distance from Earth. However the significant uncertainties in the measurement of galactic speeds and distances place limitations on this relationship.
- The gradient of a recession velocity–distance graph is equal to the Hubble constant, although its value is also uncertain.
- Hubble's constant can be used to estimate the age of the universe. However such a calculation makes the unreasonable assumption that the universe always expanded at the same rate.
- The early universe was incredibly hot, but as it expanded and cooled it became possible for particles to combine to form nucleons, then nuclei, then atoms and then stars and galaxies.

Examination questions – a selection

Paper 3 IB questions and IB style questions

Sections E1 to E4

Q1 a i What is the main energy source of a star? [1]

ii Explain how it is possible for a main sequence star to remain stable for billions of years. [2]

b i Define the luminosity of a star. [1]

ii Explain why main sequence stars can have very different luminosities. [2]

c i Define the apparent brightness of a star. [2]

ii Give two reasons why stars may have different apparent brightnesses. [2]

d Antares is a red supergiant 170 pc from Earth. Its luminosity is 2.5×10^{31} W and its surface temperature is 3400 K.

i Calculate the apparent brightness of Antares as seen from Earth. [2]

ii Explain what is meant by the term *red supergiant*. [2]

iii At what wavelength is the maximum intensity of the spectrum from Antares? [2]

iv What is the spectral class of Antares? [1]

Q2 This question is about the characteristics of the stars Procyon A and Procyon B.

a The stars Procyon A and Procyon B are both located in the same stellar cluster in the constellation Canis Minor. Distinguish between a constellation and a stellar cluster. [2]

b The table shows some data for Procyon A and Procyon B.

	Apparent magnitude	Absolute magnitude	Apparent brightness/W m^{-2}
Procyon A (P_A)	+0.400	+2.68	2.06×10^{-8}
Procyon B (P_B)	+10.7	+13.0	1.46×10^{-12}

Explain, using the data from the table, why:

i as viewed from Earth, P_A is much brighter than P_B [2]

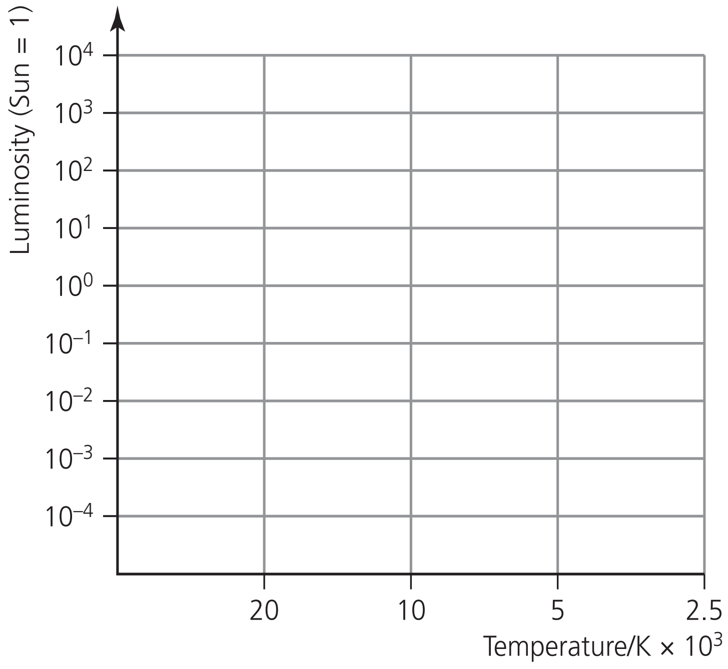
ii the luminosity of P_A is much greater than that of P_B. [3]

c Deduce, using data from the table in **b**, that P_A and P_B are approximately the same distance from Earth. [2]

d State, using your answers to **a** and **c**, why P_A and P_B might be binary stars. [1]

e Calculate, using data from the table in **b**, the ratio $\frac{L_A}{L_B}$ where L_A is the luminosity of P_A and L_B is the luminosity of P_B. [2]

f The surface temperature of both P_A and P_B is of the order of 10^4 K. The luminosity of P_A is of the order $10L_S$, where L_S is the luminosity of the Sun. The diagram shows the grid of a Hertzsprung–Russell diagram.



Copy the grid and label the approximate position of:

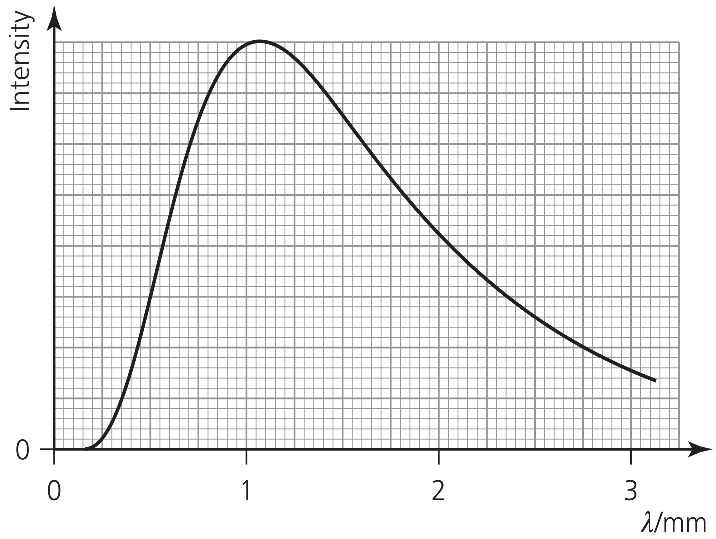
i star P_A with the letter A [1]

ii star P_B with the letter B. [1]

g Identify the nature of star P_B. [1]

Standard and **Higher** Level Paper 3, Nov 10

Q3 This question is about cosmic microwave background radiation. The graph shows the spectrum of the cosmic microwave background radiation.



The shape of the graph suggests a black-body spectrum, i.e. a spectrum to which the Wien displacement law applies.

- **a** Use the graph to estimate the black-body temperature. [2]
- **b** Explain how your answer to **a** is evidence in support of the Big Bang model. [2]
- **c** State and explain another piece of experimental evidence in support of the Big Bang model. [2]

Standard and **Higher** Level Paper 3, May 09 TZ1

Q4 This question is about models of the universe. Observations of the night sky indicate that there are many regions of the universe that do not contain many stars.

- **a** Explain why this observation contradicts Newton's model of the universe. [3]
- **b** Outline how the Big Bang model of the universe is consistent with this observation. [3]

Standard and **Higher** Level Paper 3, May 09

Sections E5 and E6 (Higher Level only)

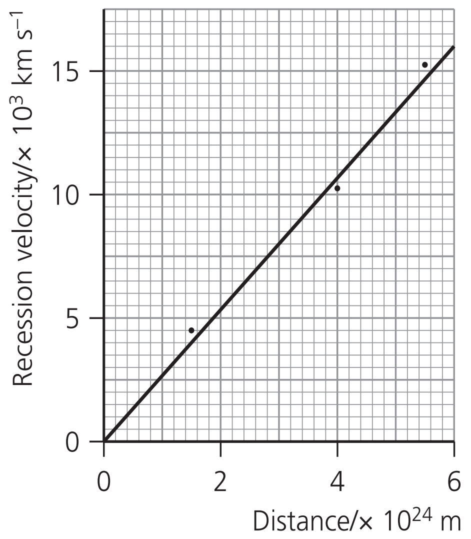
Q5 This question is about the mass–luminosity relation and also the evolution of stars.
The mass–luminosity relation for main sequence stars is assumed to be $L \propto M^{3.5}$, where L is the luminosity and M is the mass. Star X is 8×10^4 times more luminous than the Sun and 25 times more massive than the Sun.

- **a** Deduce that star X is a main sequence star. [2]
- **b** Outline with reference to the Oppenheimer–Volkoff limit, the evolutionary steps and the fate of star X after it leaves the main sequence. [3]

Higher Level Paper 3, May 10 TZ2, QE4

Q6 This question is about Hubble's law and the expansion of the universe.

- **a** The spectrum of the cluster of galaxies Pegasus I shows a shift of 5.04 nm in the wavelength of the K-line. The wavelength of this line from a laboratory source is measured as 396.8 nm. Calculate the velocity of recession of the cluster. [2]
- **b** The graph shows the recession velocities of a number of clusters of galaxies as a function of their approximate distances.



  - **i** State **one** method by which the distances shown on the graph could have been determined. [1]
  - **ii** Use the graph to show that the age of the universe is about 10^{17} s. [2]

Higher Level Paper 3, May 10 TZ2, QE5

Option G

Chapter 19 Electromagnetic waves

STARTING POINTS

- Electric fields exist around charges. Magnetic fields exist around currents (moving charges).
- Wave speed equals frequency multiplied by wavelength: $v = f\lambda$.
- The oscillations in all transverse waves are perpendicular to the direction in which the energy is propagated.
- Electromagnetic radiation may be considered as waves or as photons. The energy carried by an individual photon can be calculated from $E = hf = hc/\lambda$.
- Electromagnetic waves may refract when they enter a medium in which their speed changes. Refractive index = $\dfrac{\text{speed in a vacuum (air)}}{\text{Speed in the medium}}$
- The permittivity and permeability of free space describe the electrical and magnetic properties of a vacuum.
- Waves are diffracted when they pass through apertures or around obstacles. Diffraction is most significant when the wavelength is equal to the size of the gap or obstacle.
- Intensity = $\dfrac{\text{incident power}}{\text{area}}$
- Photons are emitted when an electron moves from a higher to a lower energy level within an atom. The photon's energy is equal to the difference in energy levels of the electron.
- Analysis of emission and absorption spectra leads to a determination of energy levels within atoms.
- When waves from different sources arrive at the same point, the principle of superposition can be used to find the resultant displacement at any moment.
- If waves arrive in phase, constructive interference occurs. If waves arrive completely out of phase, destructive interference occurs.
- In order to observe an interference pattern, the waves which interfere must be coherent. The resultant at any one point can be determined by considering the path difference between the waves.
- Electrons accelerated across a vacuum gain kinetic energy $\frac{1}{2}mv^2 = eV$. The kinetic energy gained by an electron accelerated by a p.d., V, of one volt is 1.6×10^{-19} J, which is also known as one electronvolt (1 eV).
- Most solids have some regularity in the way that their particles are arranged.

G1 The nature of EM waves and light sources

Nature and properties of EM waves

G.1.1 Outline the nature of electromagnetic (EM) waves.

In order to develop a deeper understanding of the nature of electromagnetic waves, we will first review the connection between electric and magnetic fields. The magnetic fields created whenever charges move (along wires) were discussed in Chapter 6. Consider Figure 19.1, in which a potential difference applied along a length of wire sets up an electric field. Charges (electrons) flow along the electric field

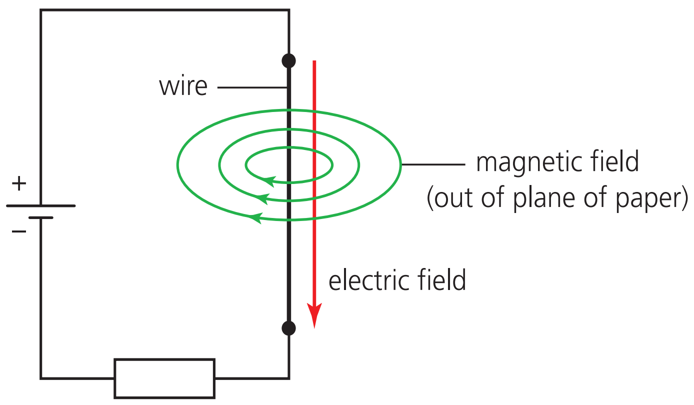


Figure 19.1 Electric and magnetic fields associated with a direct current in part of a circuit

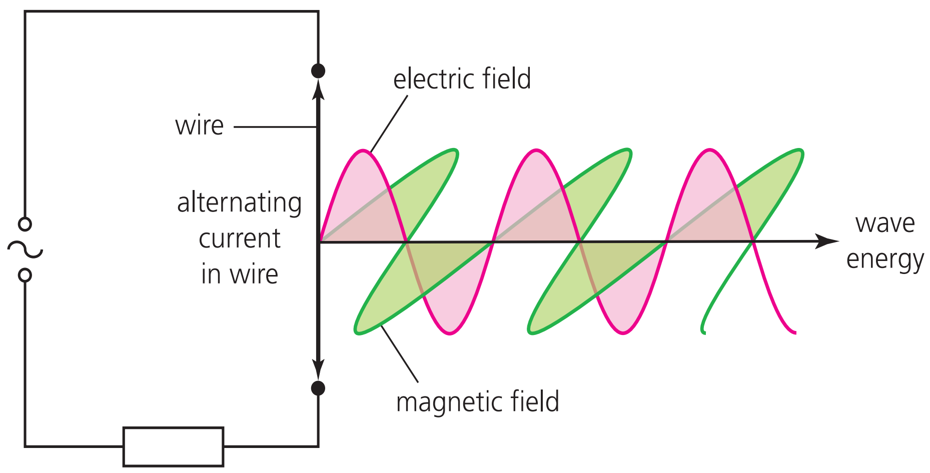


Figure 19.2 Perpendicular electric and magnetic fields spreading away from an alternating current

lines, setting up a magnetic field in a plane perpendicular to the electric field.

If the cell is replaced by an ac power supply, as shown in Figure 19.2, *changing* electric and magnetic fields (perpendicular to each other) are created around the wire. These disturbances spread away into the surroundings in a similar way to the waves created by other types of oscillating sources, as discussed in Chapter 4. This is an example of an electromagnetic (EM) wave. Unlike other waves, however, these disturbances of electric and magnetic fields do not need a medium through which to travel. They can travel (**propagate**) through a vacuum.

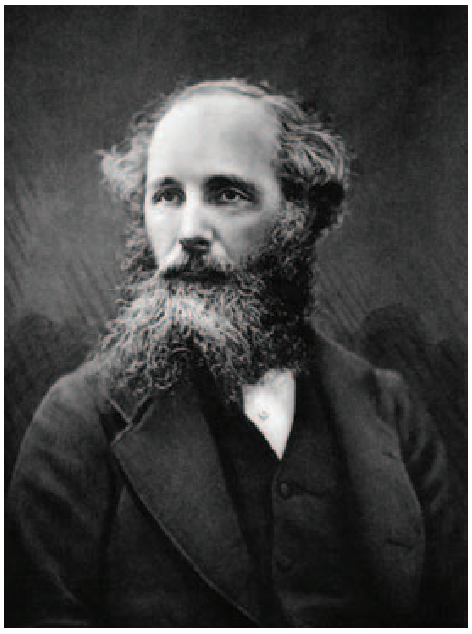


Figure 19.3 James Clerk Maxwell

James Clerk Maxwell (1831–79, Figure 19.3), is considered to be one of the greatest of physicists. This is largely because of his work on the theory of EM waves, which are undoubtedly one of the most important features of the physical universe. Maxwell's theoretical work was done about 20 years before EM waves were first artificially produced and detected by Heinrich Hertz in 1886. Maxwell was able to predict the speed of EM waves, c, to be $3.0 \times 10^8\,\mathrm{m\,s^{-1}}$ from a knowledge of the permeability and permittivity of free space. (He showed that $c = 1/\sqrt{(\mu_0\varepsilon_0)}$.) This speed was equal to the known speed of light. Up until that time, the nature of light had always been a major issue in science but, after Maxwell's work, it became clear that light was an EM wave. Maxwell's work on EM waves led to a very important and *powerful synthesis* of important concepts in physics.

An EM wave produced artificially as shown in Figure 19.2 is usually called a *radio wave*, but it will only spread away efficiently from a conductor (which would then be described as an **aerial** or an **antenna**) if the wavelength is comparable to the length of the wire (remember diffraction theory from Chapter 4).

TOK Link: Synthesis of ideas

Some of the most significant advances in human knowledge occur when very different areas of human experience are seen to have common connections. Ideas can be combined (synthesized) and widespread patterns can then be identified.

In the history of physics, Maxwell's work on *electromagnetic waves* is an excellent example, combining ideas on electricity, magnetism, light and other parts of the electromagnetic spectrum. Joule's work on *heat*, as discussed in Chapter 3, was also a major turning point in the development of physics.

Questions

1 Explain in what way the work of James Prescott Joule united different areas of physics knowledge.

2 Choose one physicist from the following list and discuss in what way(s) any of their work can be described as a *synthesis* of ideas: Galileo, Newton, Einstein, Rutherford.

Worked example

1 Suggest a frequency of an EM wave which would spread efficiently away from an aerial of length 10 cm (the size of a typical mobile phone).

$v = f\lambda$

Using the aerial size as the approximate wavelength:

$3.0 \times 10^8 = f \times 0.10$

$f = 3.0 \times 10^9\,\mathrm{Hz}$

With modern electronics, producing high-frequency currents to transmit radio (micro-) waves like this is not a problem, but such facilities were not available to the early radio pioneers. Hertz used sparks to transmit the first artificial EM waves, and spark gap transmitters were used extensively (with Morse code) in the early years of radio communication. Sparks were used because they involve high-frequency oscillating currents.

As the oscillating electromagnetic field spreads out and passes through a distant receiving aerial it induces a tiny oscillating current, similar to the one which originally produced it. The current can then be detected and amplified. A simple, artificially produced wave like the one described above will be polarized (see Chapter 11), which suggests that the transmitting and receiving aerials should point in similar directions.

Origin and speed of all EM waves

All EM waves are produced as a result of the changing electric and magnetic fields that occur when charged particles oscillate or otherwise change their motion (accelerate). For example, radio waves are produced when the current in the aerial changes, light waves are emitted from an atom when there are changes to an electron's energy level, and gamma rays are emitted when there are changes inside a nucleus.

All EM waves propagate away from their source as linked electric and magnetic fields, perpendicular to each other and perpendicular to the direction in which the wave energy is being transferred, as shown in Figure 19.4. It should be clear from this description that *all* EM waves are transverse in nature.

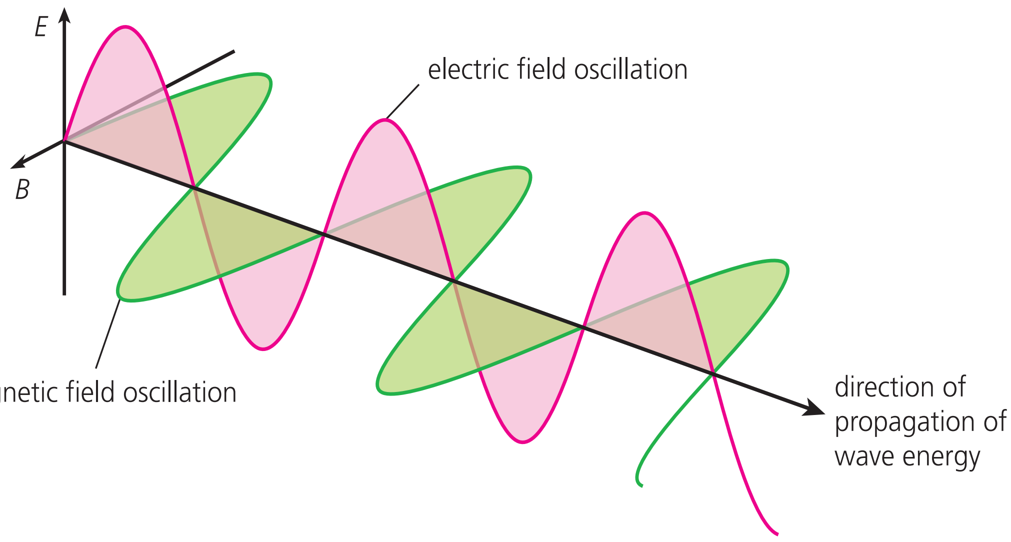


Figure 19.4 Transverse electromagnetic waves

The speed of EM waves depends on the electric and magnetic properties of a vacuum (free space) or the medium through which they are travelling. All EM waves travel at the same speed in a vacuum ($3.00 \times 10^8\,\mathrm{m\,s^{-1}}$), but with differing and slower speeds in other media.

For the majority of this chapter we will concentrate on the wave nature of light and other parts of the EM spectrum (and their representation by rays), but it should always be remembered that many properties of EM radiation require a 'particle' (photon) explanation, as discussed in Chapters 7 and 13.

■ Additional Perspectives

Measuring the speed of light

Early attempts to measure the speed of light (by Galileo and others) tried to measure the time it took to travel a known distance, which was made as large as possible. Such experiments were unsuccessful because the speed of light is too fast for the delay (between the production of the light at its source and its observation at a different location) to be detected with the equipment available at the time. It was widely, but wrongly, believed that light might travel at an infinite speed, so that it was received at the same instant that it was emitted.

In order to measure the speed of light with any degree of accuracy, or even obtain a reasonable estimate, it is necessary to use a very large distance and/or have the means of accurately measuring *very* small intervals of time. The Danish astronomer Ole Røemer made the first measurement, in 1676, by using observations of one of Jupiter's moons.

The first accurate estimate of the speed of light which involved measurements made entirely on Earth was made by Armand Fizeau in 1849 using a toothed wheel rotating at high speed. A light beam was made to pass through a gap between two adjacent teeth on the wheel. The light beam then travelled to a mirror several kilometres away and was reflected back to the wheel. The speed of the wheel was adjusted so that the returning light beam passed through the next gap on the wheel. For example, if a wheel has 100 teeth and rotates 500 times a second, it can effectively measure a time interval of $1/(100 \times 500)$ s, or 20 μs.

Figure 19.5 shows the principle of a technique involving adjusting the speed of a rotating mirror, used by Albert Michelson and others. In this example, an octagonal mirror reflects the light to fixed mirrors a large distance away and the returning light beam needs to strike the next side of the mirror at the correct angle in order to be seen by an observer using the telescope.

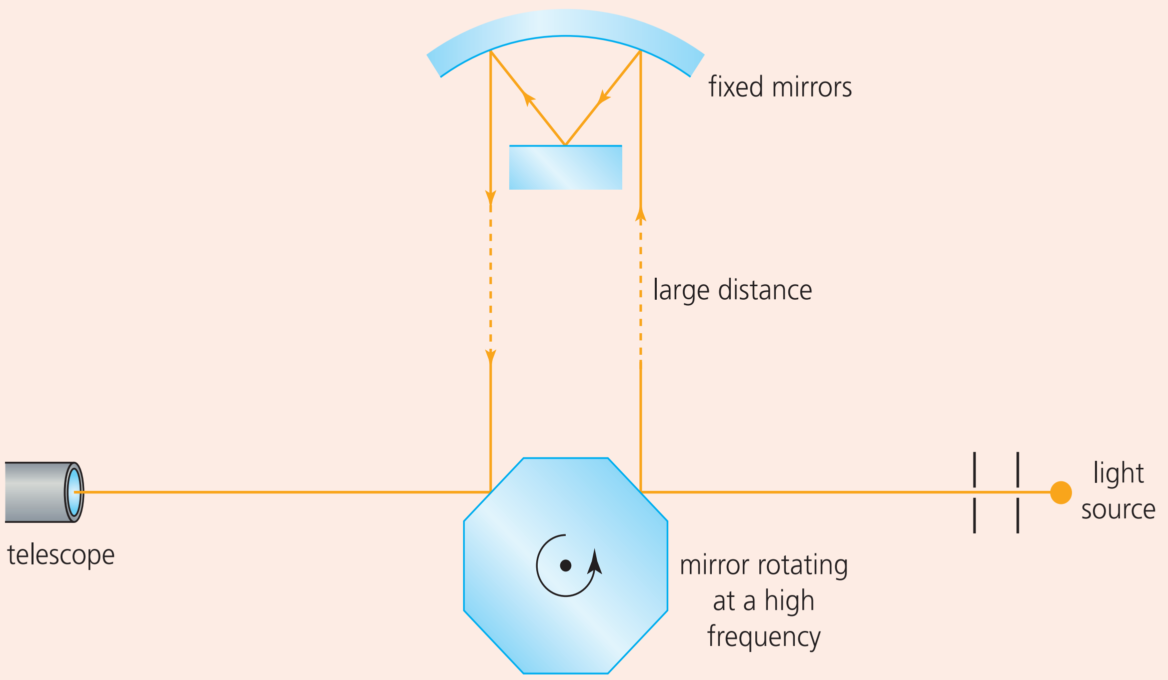


Figure 19.5 Using a rotating mirror to measure the speed of light

With modern electronics and oscilloscopes with fast time bases, it is now relatively easy to measure the speed of a pulse of light travelling down an optical fibre in a school classroom.

Questions

1 Find out how Røemer was able to estimate the speed of light using a moon of Jupiter.
2 Estimate the minimum frequency of rotation of the mirror needed (Figure 19.5) in order to measure the speed of light if the distance to the fixed mirrors is 500 m.

Electromagnetic spectrum

G.1.2 Describe the different regions of the electromagnetic spectrum.

Table 19.1 (overleaf) gives some brief details of the different regions of the electromagnetic spectrum. This table is a repeat of Table 4.1 from Chapter 4 with the addition of orders of magnitude for typical frequencies and for typical photon energies. Most people prefer to think of EM waves in terms of their wavelengths, rather than their frequencies. However, frequency is the basic property of a wave, whereas the wavelength of a radiation will change if it enters a medium in which its speed is different ($c = f\lambda$).

Table 19.1 Regions of the electromagnetic spectrum

Name	Typical order of magnitude			Origins	Some uses
	Wavelength /m	**Frequency /Hz**	**Photon energy /J**		
Radio waves	10^{2}	10^{6}	10^{-26}	Electronic circuits /aerials	Communications, radio, TV
Microwaves	10^{-2}	10^{10}	10^{-23}	Electronic circuits /aerials	Communications, mobile phones, ovens, radar
Infrared (IR)	10^{-5}	10^{13}	10^{-20}	Everything emits IR. Hotter objects (like 'heaters') emit *much* more IR than cooler objects	Lasers, heating, cooking, medical treatments, remote controls
Visible light	5×10^{-7}	6×10^{14}	4×10^{-19}	Very hot objects, light bulbs, the Sun	Vision
Ultraviolet (UV)	10^{-8}	10^{16}	10^{-17}	Sun, UV lamps	Fluorescence
X-rays	10^{-10}	10^{18}	10^{-15}	X-ray tubes	Medical diagnosis and treatment, investigating the structure of matter
Gamma rays	10^{-13}	10^{21}	10^{-12}	Radioactive materials	Medical diagnosis and treatment

When the 'sizes' of photon energies are compared (the energies of X-rays and microwaves, for example, are in the ratio of 100 000 000 : 1), it is not surprising that different regions of the spectrum have very different properties. Radiation from the high-frequency end of the spectrum is a potential health hazard to humans because of the large amount of energy carried by individual photons, but low-frequency radiation, for example from transmission lines and household electrical appliances, is considered by most scientists to be harmless.

Dispersion of EM waves

G.1.3 Describe what is meant by the dispersion of EM waves.
G.1.4 Describe the dispersion of EM waves in terms of the dependence of the refractive index on wavelength.

All EM waves travel at exactly the same speed in a vacuum ($c = 2.998 \times 10^{8}\,\mathrm{m\,s^{-1}}$ to four significant figures). The speed of light in air is $2.997 \times 10^{8}\,\mathrm{m\,s^{-1}}$. The difference between these figures is usually considered insignificant. When EM waves travel through other mediums (media) they travel more slowly than in air, but it is important to realize that waves of different frequencies do not have identical speeds in the same medium. Because of this, a beam of radiation containing EM waves of different frequencies will be *dispersed* (spread out and separated) into its constituent frequencies when it passes obliquely from one medium to another.

Consider Figure 19.6, in which a ray is used to show the direction of travel of EM waves of two different frequencies approaching a boundary between two media. The EM waves have the same speed in medium 1, but travel more slowly in medium 2. This causes a change of direction (*deviation*), called *refraction*, as discussed in Chapter 4. In medium 2, waves with frequency A travel faster than waves with frequency B, so they are refracted less and the two sets of EM waves travel in different directions and become dispersed.

All EM waves may be dispersed in this way when they pass into suitable media, but the most well-known example is the dispersion of white light into the colours of the visible spectrum (see Figure 19.7). Figure 19.7 shows that violet light is refracted the most; this is because its speed changes the most when it enters glass (violet light travels more slowly in glass than the other colours).

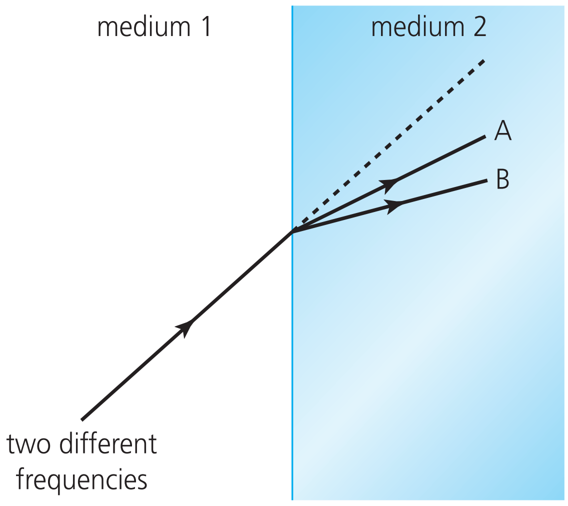


Figure 19.6 A change of speed causes deviation and dispersion

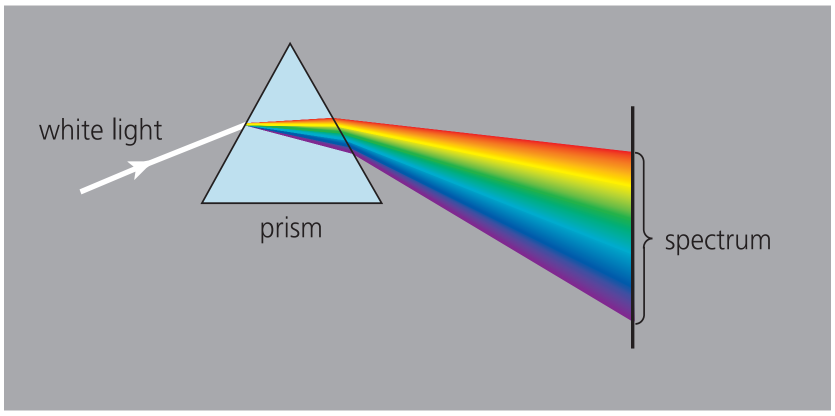


Figure 19.7 Dispersion of white light by a prism

In Chapter 4 we saw that the refractive index of a material is the ratio of the speed of the wave in air (or a vacuum) to the speed of the waves in the material:

$$n_{glass} = {}_{air}n_{glass} = \frac{v_{air}}{v_{glass}}$$

The refractive index of a material (glass, for example) is not constant, but depends on the wavelength/frequency of the light. The dispersion of colours in Figure 19.7 shows that the refractive index for glass is greater for violet light than for red light.

Transmission, absorption and scattering

G.1.5 Distinguish between transmission, absorption and scattering of radiation.
G.1.6 Discuss examples of the transmission, absorption and scattering of EM radiation.

Electromagnetic waves can travel unaffected through a vacuum, but when they travel through another medium they may be **transmitted** (unaffected), **absorbed** or **scattered**. A material is described as being **transparent** if radiation is mostly transmitted through it without being affected in any way, or **opaque** if none of the radiation is transmitted. In general the gradual decrease in intensity as radiation passes through a medium is called **attenuation**.

If electromagnetic waves are absorbed, energy is transferred to the atoms and molecules of the medium through which it is passing and this will result in an increase in temperature. Examples of absorption of EM waves include light photons exciting electrons to higher energy levels, infrared photons transferring their energy to increased oscillations of molecules, and X-ray photons being partially absorbed in human tissues and bones. For a photon to be absorbed it must have an energy which is suitable for exciting some process in the medium. If the photon's energy is too little or too large, the radiation will be mostly transmitted. For example, radiation originating from alternating currents in electrical circuits, radio waves, X-rays and gamma rays are all very penetrating (not easily absorbed).

Absorbed energy may be re-emitted, but it would then be radiated in all directions (not just the original direction), so that absorption will always result in a reduction in the intensity directly transmitted.

Scattering of radiation is also a complicated process. It can be considered as random reflections of radiation in all directions from small particles or molecules in the medium, but without the radiation being absorbed or the nature of the radiation being affected.

The effect of the Earth's atmosphere on EM radiation

- **Colour of the sky.** When we look at the sky our eyes are receiving light that has been scattered by the Earth's atmosphere. In a place where there is no atmosphere, for example on the Moon, the sky would appear black and the Sun would be completely white. Scattering in the Earth's atmosphere is mostly caused by water droplets, particles of dust or gas molecules, and the radiation may be scattered once or many times. On a cloudless day with a clear atmosphere, scattering will be mainly from gas molecules.

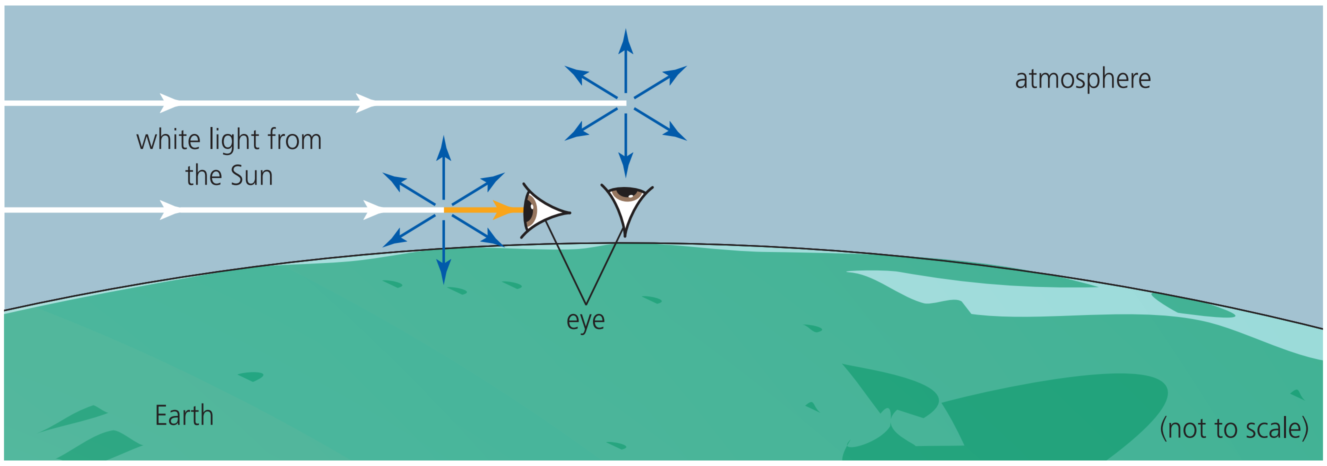


Figure 19.8 Scattering of blue light explains the colour of the sky, and the colours seen at sunrise and sunset

The amount of scattering of EM radiation by molecules in the atmosphere depends on wavelength: the smaller the wavelength, the greater the scattering. This means that the blue (violet) end of the visible spectrum is scattered much more than red light, so that the sky appears blue (see Figure 19.8).

For the same reason, the Sun usually appears pale yellow, rather than white. This is because some blue light has been removed by scattering. When light has to pass through a greater length of the atmosphere to reach our eyes (at sunrise and sunset), the Sun and the sky around the Sun often appear red and orange, as in Figure 19.9.

Figure 19.9 The colours of sunset are caused by the scattering of light by the atmosphere

- **Use of microwaves for satellite communication.** Microwaves can be transmitted through the Earth's atmosphere without any significant absorption or scattering.
- **Greenhouse gases.** Some gases which are present in the atmosphere (for example carbon dioxide) absorb certain infrared frequencies because of molecular resonance. These gases are called greenhouse gases and their effect on the Earth is discussed in Chapter 8.
- **Ozone layer.** Ozone (O_3) is formed in the upper atmosphere when ultraviolet radiation interacts with oxygen (O_2). This process has formed the ozone layer in the atmosphere, high above the Earth's surface. The ozone then interacts with further ultraviolet radiation. These photochemical reactions absorb nearly 99% of the ultraviolet radiation from the Sun which would otherwise be very dangerous for life on Earth. Human activity and the release of certain gases have resulted in a depletion of the ozone layer, creating 'holes' near the poles, especially the South Pole.

1 In what regions of the EM spectrum are the following wavelengths and frequencies?
a 2 nm **b** 20 MHz **c** 200 mm **d** 2000 GHz

2 **a** Approximately how many light photons are emitted every second by a household light bulb rated at 12 W, if it has an efficiency of 24% in producing visible light?
b Draw a very small circle, the area of which is to represent the energy carried by a light photon. On the same page draw another circle representing the energy carried by an X-ray photon.
c Explain why X-rays are more dangerous to the human body than light waves.

3 In what ways are ultraviolet radiations dangerous to humans?

4 The transmitter on the top of the tower shown in Figure 19.10 emits radio waves, which are represented by circular wavefronts.
a Estimate the frequency of the waves.
b Why is the aerial placed at the top of the tower?

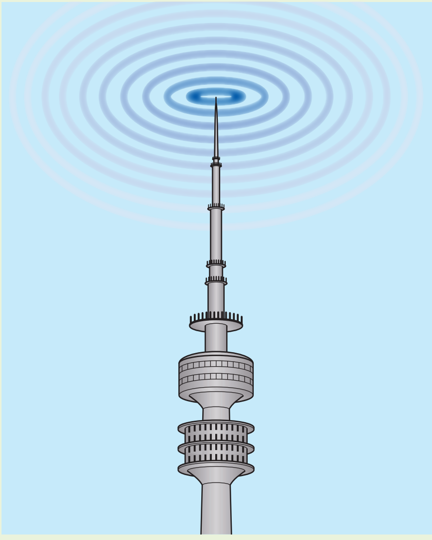


Figure 19.10

5 Carbon dioxide gas absorbs infrared radiation of wavelength 4.3 μm.
a What is the frequency of this radiation?
b In what region of the electromagnetic spectrum is this radiation?
c What important effect does this absorption produce?

6 The speed of red light in a certain kind of glass is $2.21 \times 10^8\,\text{m s}^{-1}$, while violet light travels at $2.14 \times 10^8\,\text{m s}^{-1}$ in the same glass.
a Calculate the refractive indices for this glass for these two colours.
b What is the angle of dispersion in glass between these two colours if white light enters the glass at an angle of incidence of 53.5°?

7 Research into, and explain, one non-medical use of gamma rays.

8 Some recent reports suggest that the rate of depletion of the ozone layer is not as rapid as formerly believed, and that it may 'repair itself'. Use the Internet to investigate into the latest research on this important issue.

Lasers

G.1.9 Outline the mechanism for the production of laser light.

All waves present in the EM spectrum occur naturally, but scientists have developed ways of producing, controlling and using EM waves of many different frequencies. Examples include radio waves and X-rays. One of the most useful types of radiation is that from **lasers**. Unlike other radiation sources, laser beams diverge very little, which means that they can still have high intensities at considerable distances from their sources. This property gives lasers a wide range of applications.

Before we discuss how a laser works, we should first review (from Chapter 7) the process by which light waves/photons are normally emitted from atoms. Atoms are usually in their *ground state*, which means that their electrons are in the lowest possible energy levels. When atoms are *excited* (given energy, possibly by heating or by passing an electric current through a substance), their electrons are raised to one of a number of higher energy levels (see the example shown in Figure 19.11). The electrons usually only stay in this excited state for a very short length of time (perhaps 10^{-8} s) before returning to a lower energy level. When there is a transition to a lower energy level the atom emits a photon of energy equal to the difference in energies between the two levels.

In the example shown in Figure 19.11, a photon of energy $(E_3 - E_1)$ would be emitted $(= hf)$, but other transitions are also possible, producing photons of different energies. Such transitions are spontaneous and uncontrolled. In any normal source of light, a very large number of atoms are emitting photons with a range of different frequencies, at different times and in different directions. The photons will *not* be in phase with each other.

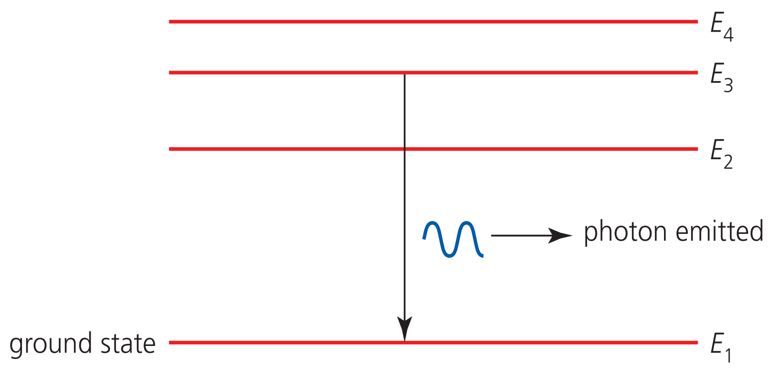


Figure 19.11 Transitions between electron energy levels produce photons of radiation

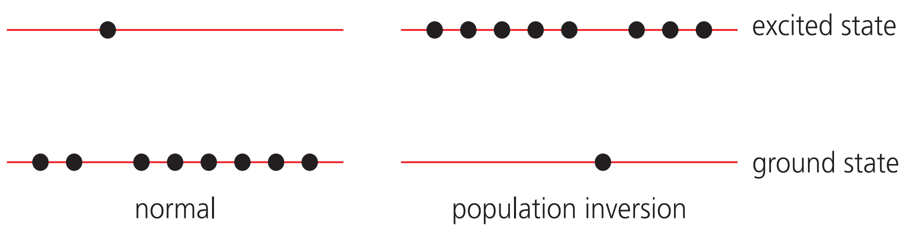


Figure 19.12 In a population inversion there are more electrons in an excited (metastable) state

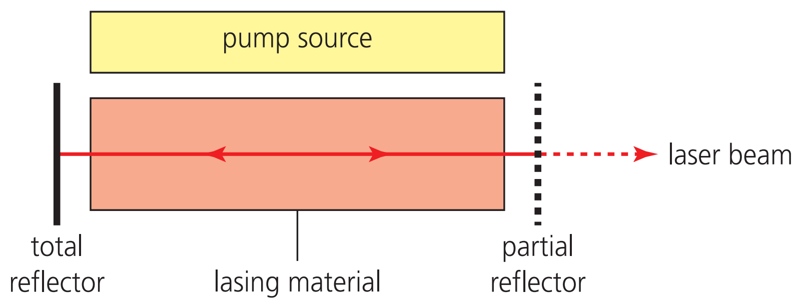


Figure 19.13 Basic components of a simple laser

In order to produce laser light we need to encourage many atoms to emit identical photons (from identical, single transitions of electrons), *at the same time* and in the same direction. To do this, we need more atoms in a particular excited energy level than in the ground state, and a means of getting those atoms to all emit photons together. As stated, atoms are usually in their ground state, and getting more of them to move into a higher energy level creates what is known as a **population inversion**, as represented by Figure 19.12. This is only possible in substances which allow electrons to stay in higher energy levels (called **metastable states**) for much longer times than usual (perhaps 10^{-3} s).

An atom with an electron in a metastable state can be encouraged (*stimulated*) to emit a photon of energy E when it interacts with another photon of the same energy, E. The two photons will have the same frequency and phase. If photons are reflected backwards and forwards between mirrors, more and more photons will be produced and the light becomes *amplified*. Figure 19.13 shows the basic components of a simple laser. The **pump source** provides the energy (for example, by using a flash-lamp) to excite the atoms in the **lasing material**. The photons which are then emitted move backwards and forwards along the 'optical cavity' and some of them emerge from the partial reflector, which transmits only a small percentage of the light incident upon it.

The word 'laser' stands for **Light Amplification by Stimulated Emission of Radiation**. Although the name specifically refers to light, similar processes can be used to produce, for example, coherent beams of microwaves, infrared and ultraviolet radiation.

Laser light is monochromatic and coherent

G.1.7 Explain the terms monochromatic and coherent.
G.1.8 Identify laser light as a source of coherent light.

Lasers produce intense beams of coherent light (or other EM waves) of a single wavelength (monochromatic).

These terms were used in Chapters 4 and 11, but it is important to stress their meanings here. The term **monochromatic** comes from ancient Greek, meaning having one colour. In scientific terms, monochromatic means EM radiation that has only one frequency (or contains only a very narrow range of frequencies). This is illustrated in Figure 19.14, which compares monochromatic radiation with the kind of *broad spectrum* radiation that would be emitted from most light sources. Laser light is monochromatic because all the photons are emitted from exactly the same energy level transition.

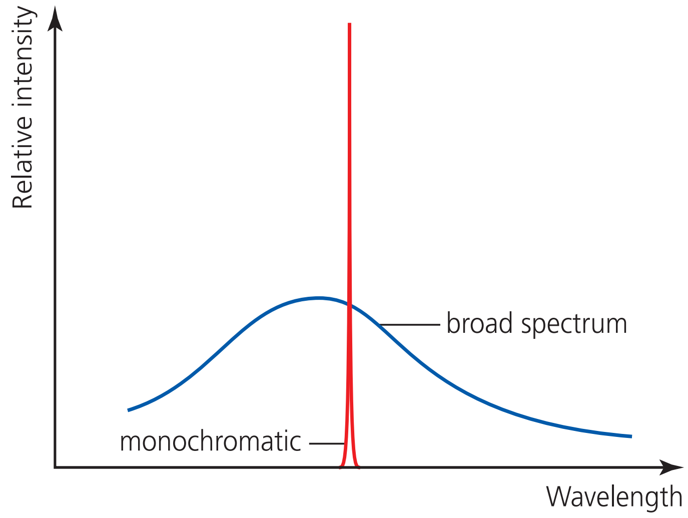


Figure 19.14 Monochromatic and broad spectrum radiation

As explained, the emissions of photons from the atoms in a laser are linked; they are not just a large number of random and spontaneous events. This results in photons emerging from the laser which are moving in the same direction and are in phase with each other. This property of lasers makes them very useful for producing interference effects. Interference was introduced in Chapter 4. Figure 19.15 indicates the difference between the light emitted from a laser and light emitted from a household light bulb.

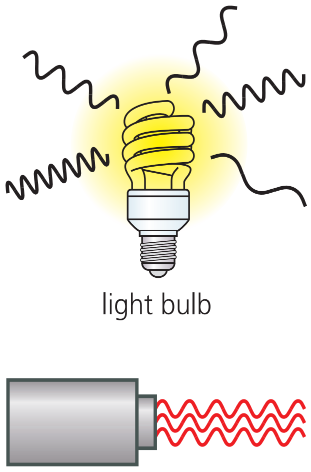


Figure 19.15 Comparing light from a household bulb with light from a laser

A series of waves which are in phase with each other are described as being **coherent**. A series of waves which are exactly out of phase are also coherent. So, to generalize, waves are coherent if they have a *constant phase difference*.

Some uses of lasers

G.1.10 Outline an application of the use of laser light.

The unique properties of lasers make them ideal for many applications in medicine, industry, communications and the storage of digital information.

- **Medical applications**. The fact that light from a laser does not diverge significantly means that a high intensity and easily controllable beam can be directed accurately onto a very small area of the human body. This makes lasers ideal for cutting and sealing tissue without damaging nearby areas, for example in the correction of eye defects (see Figure 19.16), or for surgery in which incisions (cuts) using lasers are less liable to cause excessive bleeding than those made using a scalpel (knife). Lasers are also used to unblock arteries and help destroy tumours or unwanted tissue, including marks on the skin.
- **Bar-code scanners**. Lasers are used to 'read' bar-codes for stock control in shops and factories, on supermarket check-outs and even on physics examination papers. The light and dark bars on the bar-code are simply a way of representing a number which identifies the item to which the bar-code is attached. A scanning laser sends a narrow red, or infrared, beam across the bar-code and the reflections are then detected by a sensor. The dark bars reflect much less back to the detector than the light bars.

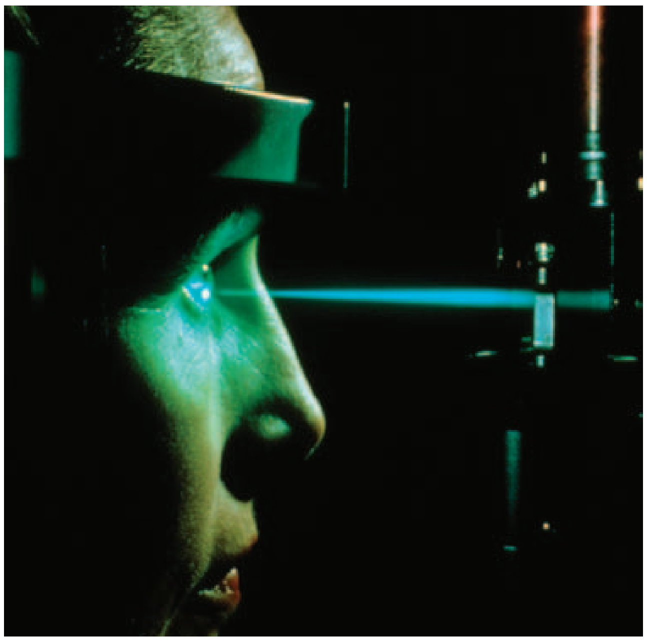


Figure 19.16 Laser eye surgery

- **Industrial uses**. High-power lasers can transfer enough energy to produce very high temperatures, which makes them very useful for cutting, welding and drilling holes in metals and some other materials (see Figure 19.17).
- **Surveying**. The fact that laser beams can travel large distances in straight lines (without diverging and losing their intensity) makes them suitable for surveying, measuring (large) distances and in weapon control.
- **Communications**. The intensity of a directed laser beam makes it ideal for transmitting digital information in the form of light pulses down optical fibres.
- **Digital information**. The coherence of very narrow laser beams makes them ideal for reading (and writing) the digital data stored on optical discs (CDs, DVDs, Blu-ray). Details of this application are discussed in Chapters 11 and 14.

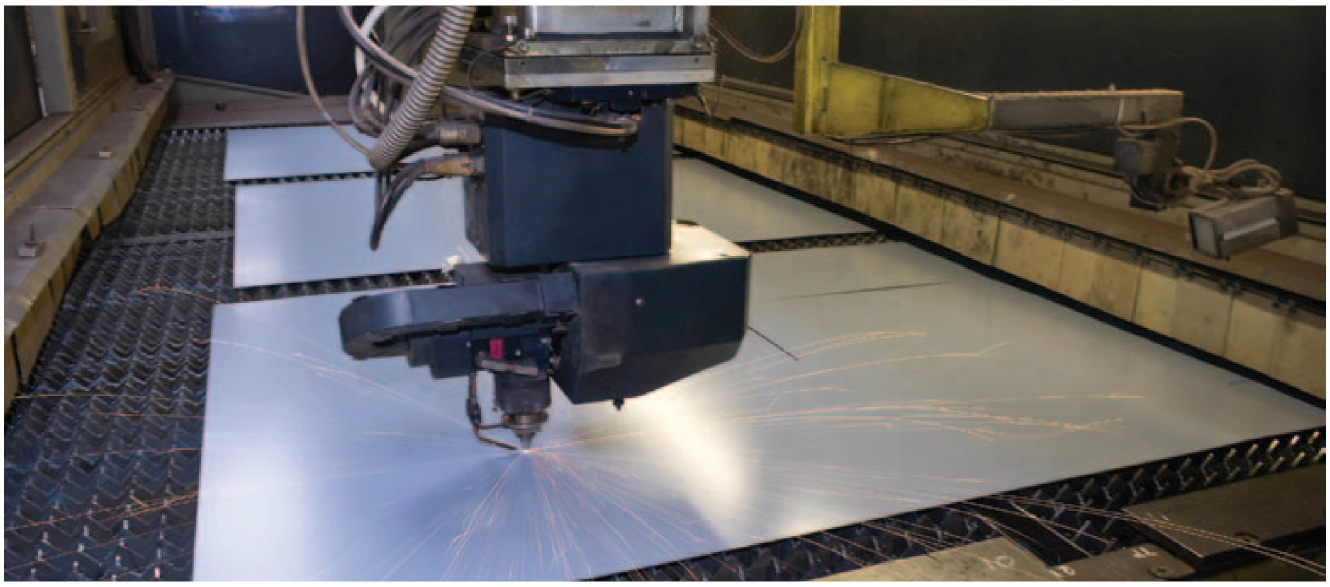


Figure 19.17 Laser cutting

9 Find out about the 'Universal Product Code', UPC.

10 a What is the intensity of radiation from a 4.10 mW, 633 nm laser emitted through a circular aperture of diameter, $b = 2.28$ mm?

b The beam is not exactly parallel, but spreads out with an angle of approximately λ/b (wavelength/diameter of aperture) to the normal. Estimate the diameter of the beam at a distance of 2.0 m from the aperture.

c What is the intensity at this distance?

d Compare your answer with the intensity at the same distance from the source of light from a domestic light bulb which emits 15 W of visible light equally in all directions.

11 Use the Internet to find out what a hologram is.

12 Suggest why using infrared radiation is preferable to visible light when sending digital data along optical fibres.

13 In a population inversion in a certain kind of laser, electrons are raised to an energy level above their ground state. These electrons then emit energy as they move to an energy level which is 1.97 eV lower. What is the wavelength (nm) emitted by the laser?

G2 Optical instruments

Seeing images

We see an object when light from it enters our eyes. Some objects emit light, but we are able to see most things because the light waves striking them are scattered in all directions and some of those waves (spreading away from the same point on the object) are brought back together by our eyes. Figure 19.18 (on page 746) shows this using rays to represent the directions in which the waves are travelling. The representation of an object that our eyes and brain 'see' is called an **image**. The term **object** is generally used to describe the thing that we are looking at.

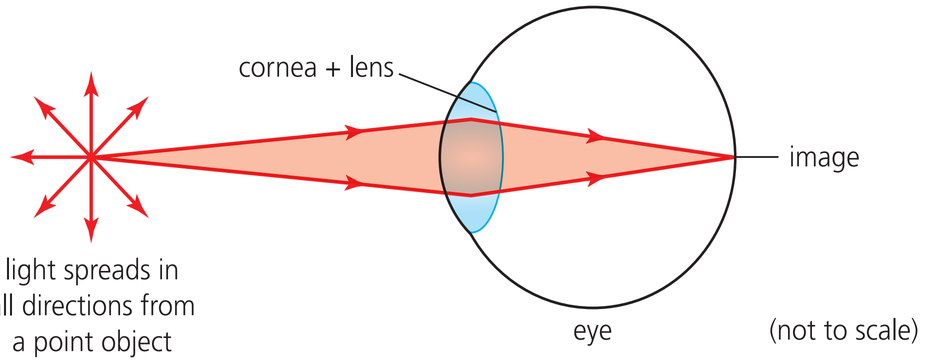


Figure 19.18 The eye focusing light to form an image

The eye uses refraction (Chapter 4) to bring the light rays diverging from a point on an object back to a point on the image. This process is called **focusing** the light to form an image. Light waves can also be focused by reflection, but this is not included in the course.

Lenses

The eye contains a lens which helps to focus the light. In general, **lenses** are made of transparent materials with smooth curved surfaces and they are used to refract EM waves into converging (or diverging) beams. Figure 19.19 shows the effects of the two basic types of lens on plane wavefronts (parallel rays). The wavefronts inside the lenses have not been included in these diagrams. Light rays will refract and change direction (deviate) at *both* surfaces of the lens, unless they are incident along a normal, but for simplicity all the diagrams in this chapter show the changes of direction only in the centre of the lens.

In Figure 19.19a the wavefronts converge to a focus; for this reason this type of lens is often called a **converging lens**. Because of the shape of its surface, this type of lens is also called a **convex lens**. Despite their name, converging lenses do not *always* converge light (magnifying glasses are discussed on page 754). Figure 19.19b shows the action of a **diverging (concave) lens**. Detailed knowledge of diverging lenses is *not* needed for this course.

Lenses are made in a wide variety of shapes and sizes, but all lenses can be described as either converging/convex or diverging/concave.

Glass lenses have many applications in which they are used to refract light waves to form images. For ease of explanation, diagrams and theory in this chapter are presented in terms of rays rather than waves.

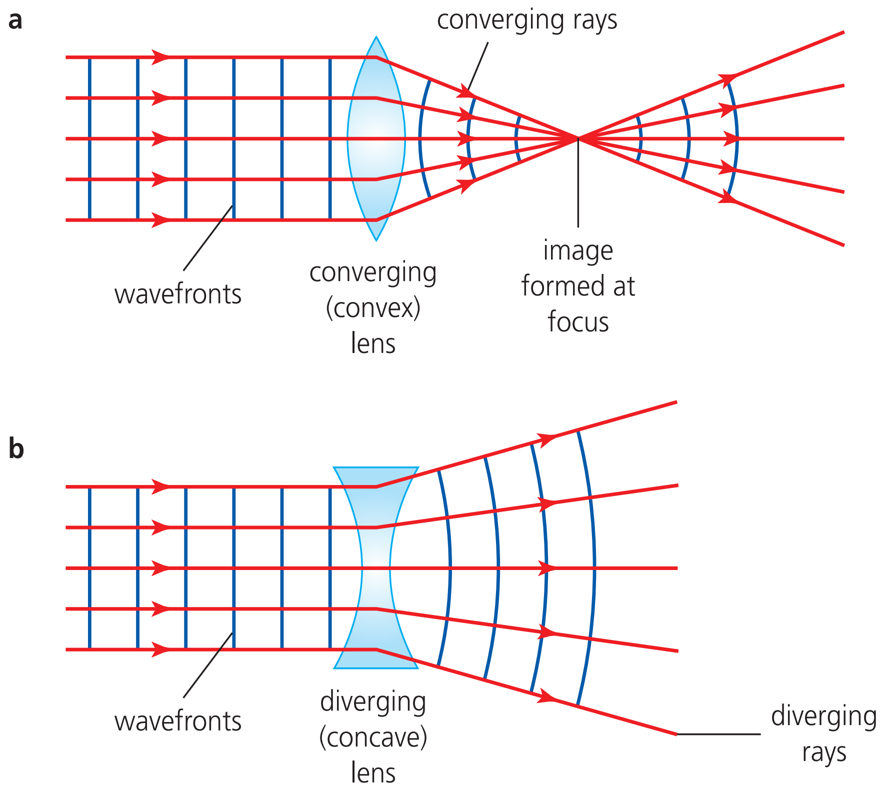


Figure 19.19 Two basic types of lens and how they affect light waves (and rays)

Lens terminology

G.2.1 (part) Define the terms *principal axis*, *focal point*, *focal length* as applied to a converging (convex) lens.
G.2.2 Define the *power* of a *convex lens* and the *dioptre*.

Figure 19.20 illustrates the basic terms used to describe lenses.

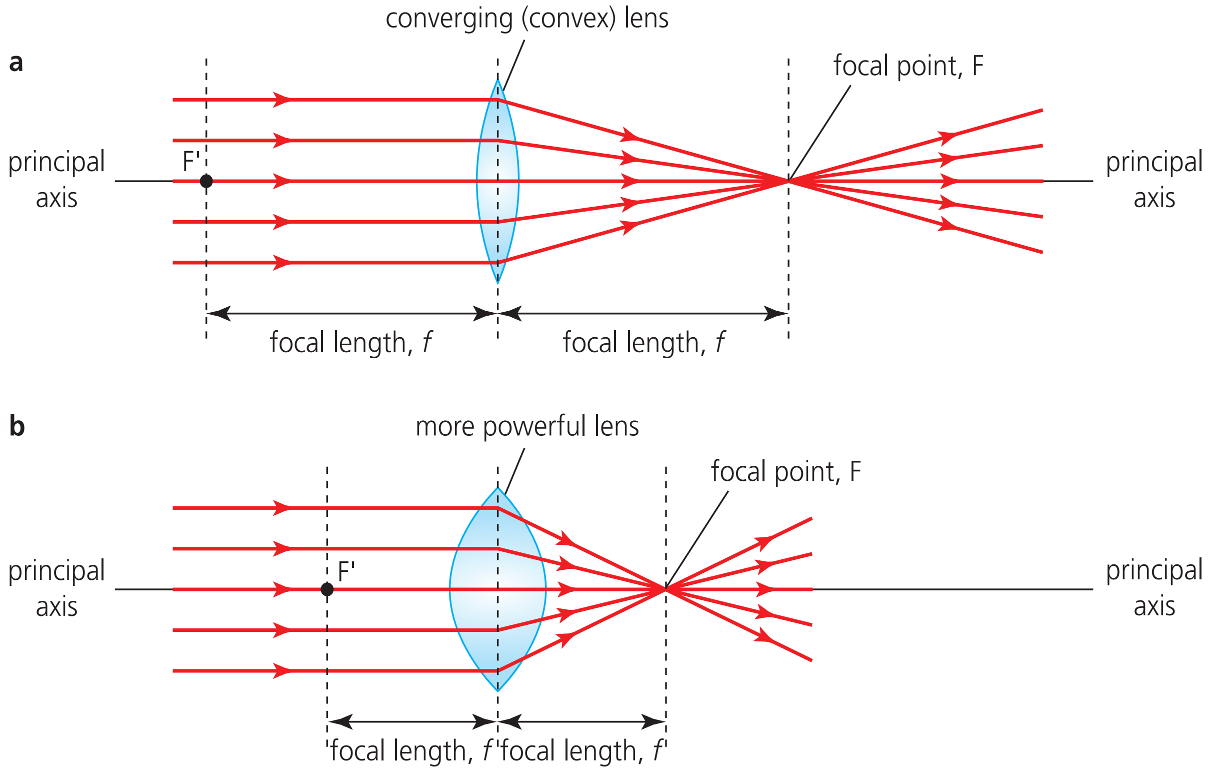


Figure 19.20 Defining the basic terms used to describe lenses

The **principal axis** of a lens is defined as the (imaginary) straight line passing through the centre of the lens which is perpendicular to the surfaces.

Light rays may be focused in different places depending on how close the object is to the lens, but a lens is defined in terms of where it focuses **parallel** rays of light that are incident upon it.

The **focal point** of a convex lens is defined as the point through which all rays parallel to the principal axis converge, after passing through the lens.

The focal point is sometimes called the *principal focus*. A lens has two focal points, the same distance from the centre of the lens on either side. These are shown as F and F' in Figure 19.20.

Lenses are not perfect and not all rays will pass exactly through the focal point. The best focus will usually be formed by using a thin lens of small diameter. Lens *aberrations* (differences from perfect lenses) are discussed in more detail on page 759. In this chapter, unless stated otherwise, we shall assume that we are dealing with light rays striking near to the middle of a thin lens.

The **focal length**, f, of a lens is defined as the distance between the *centre* of the lens and the focal point.

Focal length is typically measured in centimetres, although the SI unit is metres. The focal length of a lens is the essential piece of information about a lens which tells us how it affects light passing through it. The longer the focal length of a lens, the less effect it has on light. The shorter the focal length of the lens, the greater the refraction of the light and the lens is described as being more *powerful*.

To determine the focal length of a lens experimentally it is necessary to use parallel rays of light. These are conveniently obtained from any distant object. Spherical wavefronts from a point source become effectively parallel if they are a large distance from their origin.

The focal length of a lens depends on the curvature of the surfaces and the refractive index of the material(s) from which it is made. Simple lenses have surfaces which are *spherical* – the same shape as part of a sphere. A lens which has a smaller radius of curvature, or a higher refractive index, will have a smaller focal length and be more powerful (see Figure 19.20b).

People who work with lenses, such as optometrists and opticians, usually classify different lenses according to their (optical) power, a term which is not connected in any way to the more general meaning of power as the rate of transfer of energy. Optical power is defined as follows:

$$\text{power} = \frac{1}{\text{focal length}}$$

$$P = \frac{1}{f}$$

This equation is given in the IB *Physics data booklet.*

The unit for (optical) power is the **dioptre**, **D**, which is defined as the power of a lens which has a focal length of 1 m.

That is:

$$P(\text{D}) = \frac{1}{f(m)}$$

Worked example

2 What are the powers of lenses which have focal lengths of:
a 2.1 m
b 15 cm?

a $P = \frac{1}{f} = \frac{1}{2.1} = 0.48\,\text{D}$

b $P = \frac{1}{f} = \frac{1}{0.15} = 6.7\,\text{D}$

14 **a** What is the focal length of a convex lens with a power of 2.5 D?
b Make a sketch of a lens (of power 2.5 D) and then draw next to it a lens of the same diameter which has a much smaller focal length.
c What assumption did you make?

15 **a** Calculate the power of a lens with diameter 4.0 cm and focal length 80 mm.
b How is it possible that another lens of exactly the same shape could have a focal length of 85 mm?

16 When two lenses are placed close together, their combined power is equal to the sum of their individual powers. A lens of what focal length can be combined with a lens of power 5 D to make a combined power of 25 D?

Forming images with convex lenses

G.2.4 Construct ray diagrams to locate the image formed by a convex lens.

We can locate the position and nature of an image by constructing (to scale) a **ray diagram**, or mathematically by using the **thin lens formula** (see page 751). We will consider ray diagrams first.

Figure 19.21a shows rays that are coming from the top of an extended object (that is, it is not a point object) being focused to form an image. All rays incident on the lens are focused to the same point. If part of the lens was covered, an image would still be formed at the same point by the remaining rays.

The predictable paths of three rays coming from the top of the object are highlighted. These same three rays can be used to locate the image in *any* situation.

- A ray parallel to the principal axis passes through the focal point.
- A ray striking the centre of the lens is undeviated.
- A ray passing through the focal point emerges from the lens parallel to the principal axis.

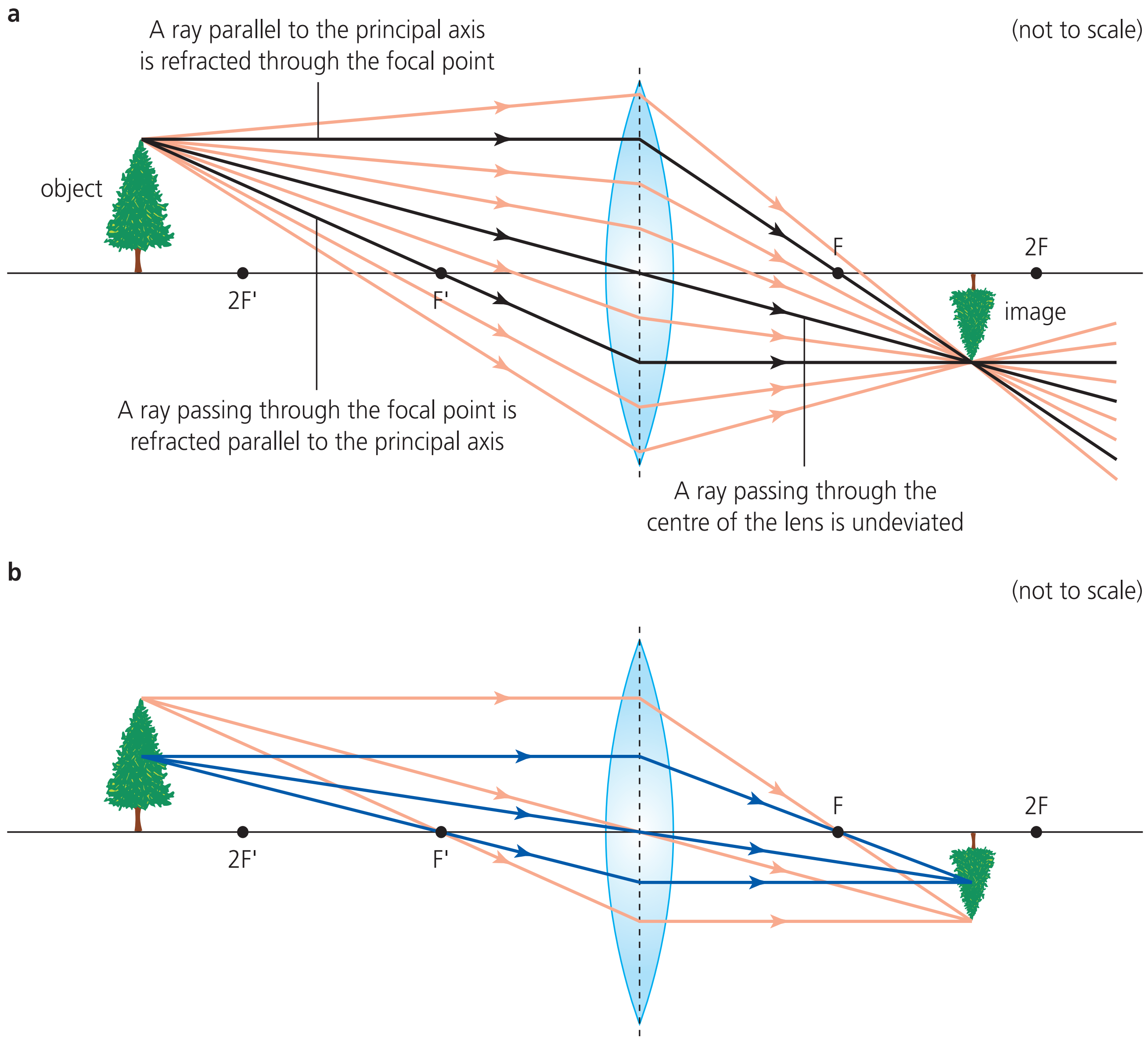


Figure 19.21 Predicting the paths of rays between an object and its image using three standard rays

Note that the vertical scale of the diagram may be misleading – a light ray striking the centre of a small thin lens from an object some distance away will be incident almost normally, which is not apparent in this diagram. Of course, all the light from an object does not come from one point at the top. Figure 19.21b also shows the paths of three rays going from the middle of the object to the middle of the image.

Nature of an image

G.2.1 (part) Define the term *linear magnification* as applied to a converging (convex) lens.
G.2.3 Define *linear magnification*.
G.2.5 Distinguish between a real image and a virtual image.

After drawing scale ray diagrams, we can predict the following about the nature of an image:

- Its *position*.
- Its *size* (and whether it is *magnified* or *diminished*).

The **linear magnification, *m***, of an image is defined as the ratio of the height of the image to the height of the object:

$$m = \frac{h_i}{h_o}$$

This equation is given in the IB *Physics data booklet*.

Because m is a ratio it has no unit. If m is greater than one the image is magnified; if m is less than one the image is diminished (smaller).

- Whether the image is *upright* or *inverted* (the same way up as the object or upside down).
- Whether the image is *real* or *virtual*. **Real images** are formed where rays of light actually converge. **Virtual images** are formed when diverging rays enter the eye and the image is formed where the rays appear to have come from. (For example, the images seen when looking at ourselves in a plane mirror or using a magnifying glass are virtual.)

In the example shown in Figure 19.21, we can see from the ray diagram that the image between the positions F and 2F (the point 2F is a distance 2*f* from the centre of the lens) is diminished, inverted and real. If the lens shown was replaced by a less powerful lens, the image would be further away, bigger and dimmer (but it would remain inverted and real).

If a lens and object are brought closer together, the image stays real and inverted but becomes larger and further away from the lens (as well as dimmer). But if the object is placed at the focal point, the rays will emerge parallel and not form a useful real image (it is at infinity). Figure 19.22 represents the possibilities in a series of diagrams for easy comparison.

If an object is placed closer to the lens than the focal point, the emerging rays diverge and cannot form a real image. Used in this way, a lens is acting as simple magnifying glass, and a virtual, magnified image is formed (see page 754).

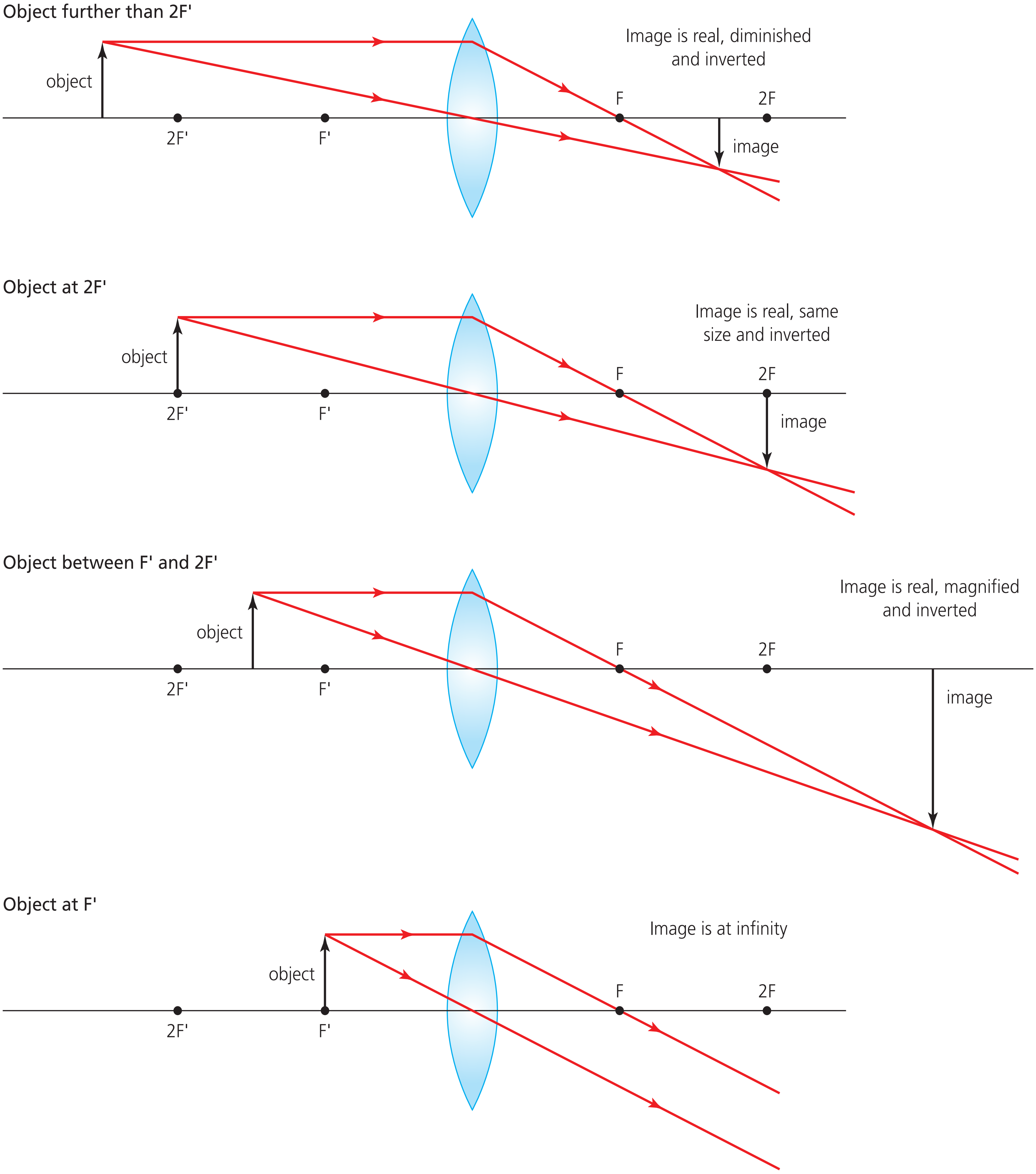


Figure 19.22 How the image changes when the object and lens become closer together

17 **a** Draw a ray diagram to determine the position and size of the image formed when an object 10 mm tall is placed 8.0 cm from a convex lens of focal length 5.0 cm.
b What is the linear magnification of the image?

18 **a** Draw a ray diagram to determine the position and size of the image formed when an object 20 cm tall is placed 1.20 m from a convex lens of power 2.0 D.
b What is the linear magnification of the image?

19 Construct a ray diagram to determine where an object must be placed in order to project an image of linear magnification 10 onto a screen which is 2.0 m from the lens.

20 An image of an object 2.0 cm in height is projected onto a screen which is 80 cm away from the object. Construct a ray diagram to determine the focal length of the lens if the linear magnification is 4.0.

21 **a** Describe the images that are normally formed by cameras.
b Draw a sketch to show a camera forming an image of a distant object.
c How can a camera focus objects that are different distances away?

Thin lens formula

G.2.6 **Apply** the convention 'real is positive, virtual is negative' to the thin lens formula.

The thin lens formula provides a mathematical alternative (to scale drawings) for determining the position and nature of an image. In this formula the symbol u is used for the distance between the object and the centre of the lens (called the *object distance*) and the symbol v is used for the distance between the image and the centre of the lens (the *image distance*), as shown in Figure 19.23.

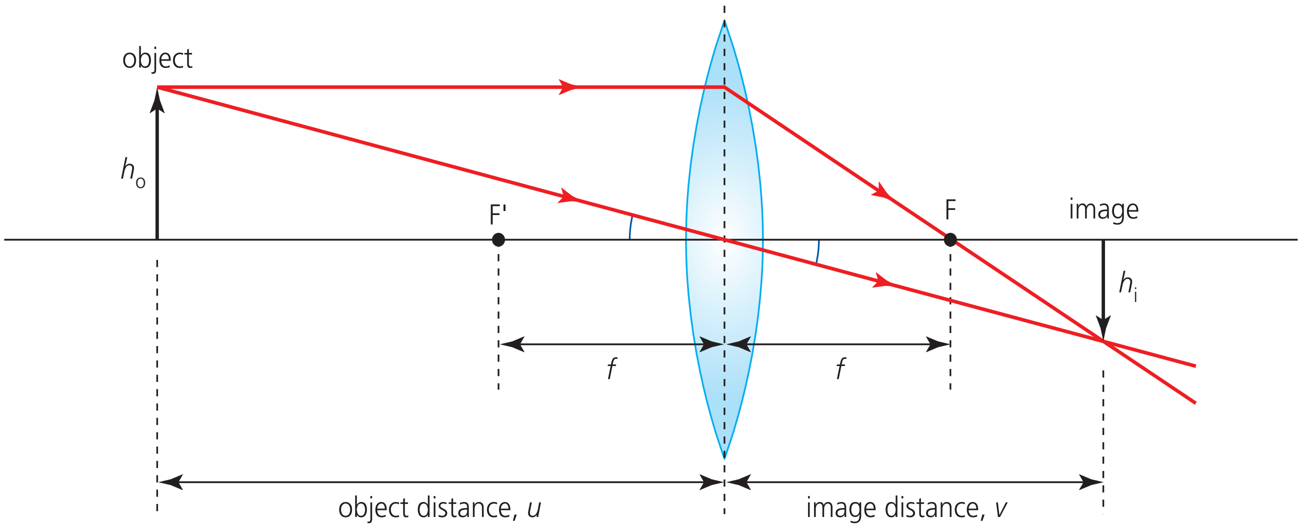


Figure 19.23 Object and image distances

The thin lens formula is:

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$$

This equation is given in the IB *Physics data booklet.*

Later in this chapter we will use this equation with a simple magnifying glass and virtual images. In those examples it will be important to use the *'real is positive, virtual is negative' convention*. This convention means that the distance to a virtual image always has a negative value. But, for the moment, we will concentrate on the use of convex lenses to form real images.
Looking at the two similar triangles with marked angles in Figure 19.23, it should be clear that:

$$\frac{h_o}{u} = \frac{h_i}{v}$$

Therefore, the magnitude of the linear magnification can also be calculated from object and image distances:

$$m = \frac{h_i}{h_o} = \frac{v}{u}$$

If a negative sign is introduced into this equation, then all real, inverted images will have *negative* magnifications, and all virtual, upright images will have *positive* magnifications (because v is negative). In other words, the sign can be used to indicate whether the image is upright (positive) or inverted (negative).

$$m = \frac{h_i}{h_o} = -\frac{v}{u}$$

This equation is given in the IB *Physics data booklet*.

Additional Perspectives

Deriving the thin lens formula

Consider Figure 19.23 again. The ray passing through the focal point on the right-hand side of the lens forms the hypotenuse of two similar right-angled triangles. Comparing these two triangles, we can write:

$$\frac{h_o}{f} = \frac{h_i}{(v - f)}$$

$$\frac{h_i}{h_o} = \frac{(v - f)}{f}$$

But we have already seen that $h_i/h_o = v/u$.
Comparing the last two equations, it is clear that:

$$\frac{v}{u} = \frac{(v - f)}{f}$$

$$vf = uv - uf \quad \text{or} \quad vf + uf = uv$$

Dividing by uvf, we get:

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$$

The important simplifying assumptions made in this derivation are that:

1 the ray parallel to the principal axis changes direction (once) in the middle of the lens and
2 the ray passing through the middle of the lens does not deviate because it is incident normally.

These assumptions are only valid for rays close to the principal axis striking a thin lens.

Question

1 Draw a ray diagram showing the formation of a real image by the refraction of rays at *both* surfaces of a converging lens.

Worked example

3 a Use the thin lens formula to calculate the position of the image formed by a convex lens of focal length 15 cm when the object is placed 20 cm from the lens.
b What is the linear magnification?
c Is the image upright or inverted?

a $$\frac{1}{f} = \frac{1}{u} + \frac{1}{v}$$

$$\frac{1}{15} = \frac{1}{20} + \frac{1}{v}$$

$$v = 60\,\text{cm}$$

b $m = -\frac{v}{u}$

$m = -\frac{60}{20} = -3.0$

c The negative sign confirms that the image is inverted.

G.2.7 Solve problems for a single convex lens using the thin lens formula.

Use the thin lens formula to answer the following questions about forming real images with convex lenses.

22 **a** Determine the position of the image when an object is placed 45 cm from a convex lens of focal length 15 cm.
b Calculate the linear magnification.

23 **a** Where must an object be placed to project an image on to a screen 2.0 m away from a lens of focal length 20 cm?
b What is the linear magnification?

24 An object is placed 10 cm away from a convex lens and forms an image with linear magnification of −3.5. What is the focal length of the lens?

25 What power lens is needed to produce an image on a screen which is 12 cm away, so that the length of the image is 10% of the length of the object?

Using optical instruments to see more detail

Normal vision

G.2.8 Defne the terms *far point* and *near point* for the unaided eye.

The human eyeball is between 2 cm and 3 cm in diameter and the focal length of its lens system must be about the same length so that parallel light from distant objects is focused on the back of the eye (the **retina**).

Muscles in the eye alter the shape of the lens in order to change its focal length (power) so that objects at different distances can be focused on the retina. These muscles are more relaxed when viewing distant objects and most strained when viewing close objects. However, the normal human eye is not powerful enough to focus light from an object which is closer than about 25 cm.

A ray diagram, or the use of the thin lens formula, will confirm that the images formed on the retina are always real, inverted and diminished.

The nearest point to the human eye at which an object can be clearly focused (without straining) is called the **near point**.

By convention, the distance from the eye to the near point for a person with 'normal' eyesight (without any aid) is agreed to be 25 cm. This distance is given the symbol D.

The furthest point from the human eye that an object can be clearly focused (without straining) is called the **far point**.

A 'normal' eye is capable of focusing things that are a long way away (although objects cannot be seen in detail). The far point is assumed to be at infinity for normal vision.

Lenses ('glasses', spectacles or contact lenses) can be used to correct the vision of people with eye defects and they are also widely used in various optical instruments for making things appear bigger (than they would without the instrument). In other words, lenses can produce an **angular magnification**. Angular magnification is a much more useful concept than linear magnification when discussing optical instruments. Examples of optical instruments include magnifying glasses and microscopes for making close objects seem larger than they are, and telescopes and binoculars for seeing objects that are a long way away more clearly.

Angular magnification

G.2.9 Define *angular magnification*.

Figure 19.24 shows an eye looking at an object and the same eye looking at an image of that object formed by some type of optical instrument (not shown).

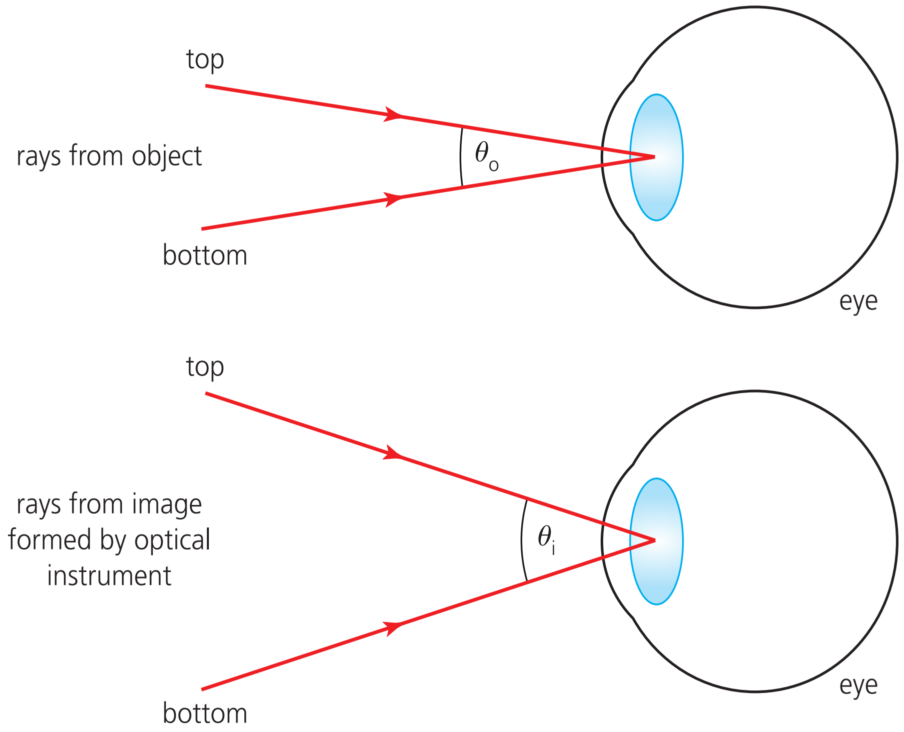


Figure 19.24 The concept of angular magnification

Angular magnification, M, is defined as:

$$M = \frac{\text{angle subtended at the eye by the image}}{\text{angle subtended at the eye by the object}}$$

$$M = \frac{\theta_i}{\theta_o}$$

This equation is given in the IB *Physics data booklet.*

Angular magnification is a ratio and therefore has no units.

Simple magnifying glass

A single convex lens can be used to make a small object, which is placed close to the lens, seem bigger (it produces both an angular and a linear magnification). Used in this way it is called a **simple magnifying glass**. The object is placed closer to the lens than the focal point so that the rays diverge into the eye, which then sees a virtual image. Figure 19.25 shows an image being formed at the near point when the object is at a distance u from the lens. (The image distance v is equal to D, assuming that the lens is very close to the eye.)

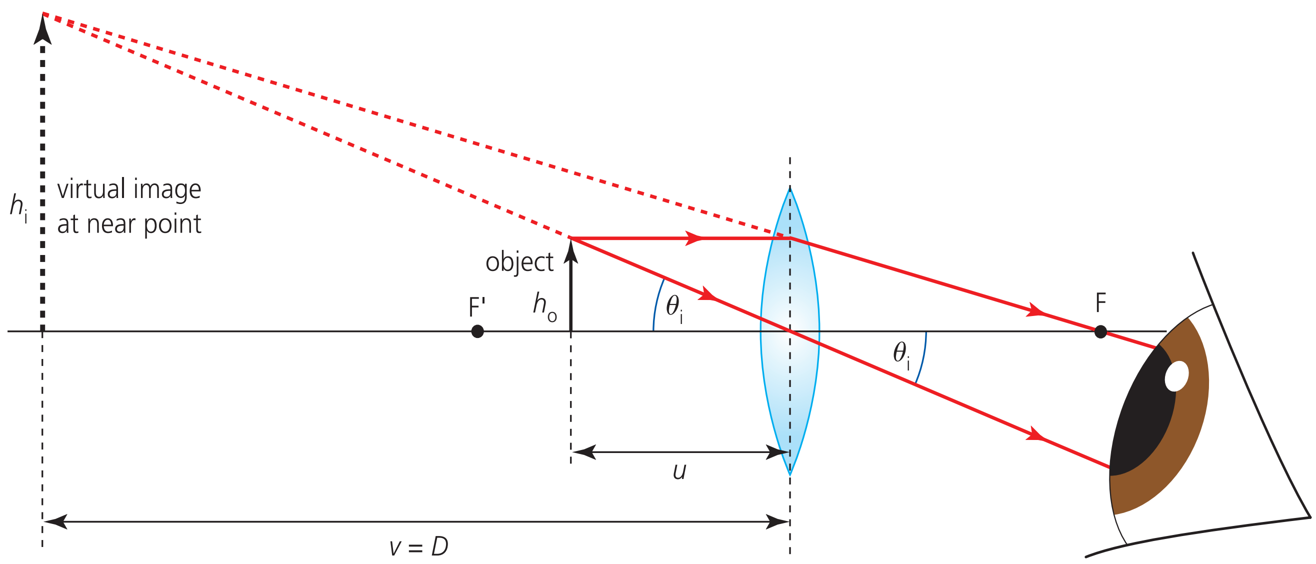


Figure 19.25 Using a convex lens as a magnifying glass to form a virtual image at the near point of the eye

Worked example

4 A convex lens of focal length 8.0 cm is used to look at an object 2.0 mm in height.
   a Where must the object be placed to form an image at the near point ($v = 25$ cm)?
   b What is the height of the image?
   c Is the image upright or inverted?

This question could be solved by drawing a ray diagram, but we will use the thin lens formula.

a $\frac{1}{f} = \frac{1}{u} + \frac{1}{v}$

$\frac{1}{8.0} = \frac{1}{u} + \frac{1}{-25}$ remembering that a virtual image must be given a negative image distance.

$u = 6.1$ cm

b $m = -\frac{v}{u} = -\frac{(-25)}{6.1} = 4.1$

So the image size $= 4.1 \times 2.0 = 8.2$ mm

c The magnification is positive, which means that the image is upright.

Angular magnification of a simple magnifying glass for an image formed at the near point

G.2.10 Derive an expression for the angular magnification of a simple magnifying glass for an image formed at the near point and at infinity.

When the image is formed at the near point, the lens produces the greatest possible magnification. The equation for the magnification can be derived in the following way.

Remember that for small angles, θ in radians $\approx \tan\theta$, so that the angle subtended at the eye by an object at the near point (without the lens) is $\theta_o = h_o/D$. From Figure 19.25 we see that when the lens is used such that the image is formed at the near point, $\theta_i = h_i/D$, so that:

$$\mathrm{M} = \frac{\theta_i}{\theta_o} = \frac{h_i / D}{h_o / D} = \frac{h_i}{h_o}$$

Looking at the similar triangles containing the angle θ_i, we see that:

$$\mathrm{M} = \frac{h_i}{h_o} = \frac{D}{u}$$

But we want an equation which gives us the angular magnification in terms of f, not u. Multiplying the lens equation ($1/f = 1/u + 1/v$) throughout by v gives us:

$$\frac{v}{f} = \frac{v}{u} + \frac{v}{v}$$

Remember that in this situation $v = -D$ (the negative sign is included because the image is virtual). Therefore:

$$-\frac{D}{f} = -\mathrm{M} + 1$$

Or:

$$\mathrm{M} = \frac{D}{f} + 1$$ for an image at the near point

This equation for the *greatest* angular magnification of a simple magnifying glass is given in the IB *Physics data booklet*.

Angular magnification of a simple magnifying glass for an image formed at infinity

The equation for the angular magnification when the image is at infinity (when the eye is most relaxed and the magnification is *least*) can be derived as follows.

Consider Figure 19.26. The image is at infinity because the object is placed exactly at the focal point. As before, $\theta_o = h_o/D$, for an eye looking without a lens at an object as close as possible.

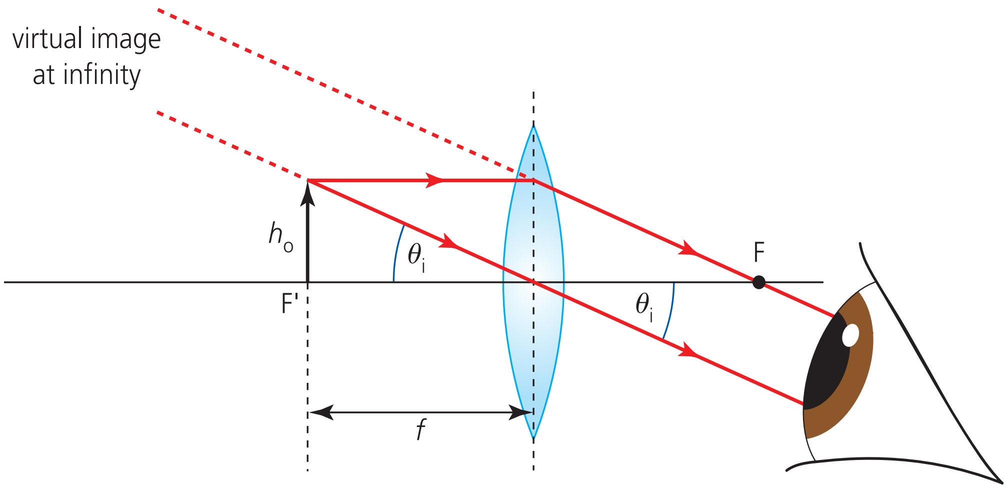


Figure 19.26 A simple magnifying glass with the image at infinity

From Figure 19.26 we see that $\theta_i = \dfrac{h_o}{f}$, so that:

$$M = \frac{\theta_i}{\theta_o} = \frac{h_o / f}{h_o / D}$$

$$M = \frac{D}{f}$$ for an image at infinity

This equation for the *least* angular magnification of a simple magnifying glass is also given in the IB *Physics data booklet*.

By adjusting the distance between object and lens, the angular magnification can be adjusted from D/f to $D/f + 1$, but the lens aberrations (see page 759) of high-power lenses limit the magnification possible with a single lens. A typical focal length for a magnifying glass is about 10 cm, which will produce an angular magnification between 2.5 and 3.5.

26 **a** Draw an accurate ray diagram to show the formation of the image when an object is placed 5.0 cm away from a convex lens of focal length 8.0 cm.
b Use the diagram to determine the linear magnification.

27 Use the thin lens formula to predict the nature, position and linear magnification of the image formed by a convex lens of power 20 D when it is used to look at an object 4.0 cm from the lens.

28 What is the focal length of a convex lens which produces a virtual image of length 5.8 cm when viewing a spider of length 1.8 cm placed at a distance of 6.9 cm from the lens?

29 **a** Calculate the angular magnification produced by a convex lens of focal length 12 cm when observing an image at the near point.
b In what direction would the lens need to be moved in order for the image to be moved to infinity and for the eye to be more relaxed?
c When the lens is adjusted in this way, what happens to the angular magnification?

30 What power lens will produce an angular magnification of 3.0 of an image at infinity?

31 Two small objects which are 0.10 mm apart can just be distinguished as separate when they are placed at the near point. What is the closest they can be together and still be distinguished, when a normal human eye views them using a simple magnifying glass which has a focal length of 8.0 cm?

32 **a** Where must an object be placed in order for a virtual image to be seen at the near point when using a lens of focal length 7.5 cm?
b Calculate the angular magnification in this position.

The compound microscope and astronomical telescope

Compound microscope

G.2.11 Construct a ray diagram for a compound microscope with final image formed close to the near point of the eye (normal adjustment).

If we want to observe an image of a nearby object with a greater magnification than can be provided with a single lens, a second convex lens can be used to magnify the first image (see Figure 19.27). The lens which is closer to the object is called the **objective** and the second lens, closer to the eye, is called the **eyepiece**. Two lenses used in this way are called a **compound microscope**.

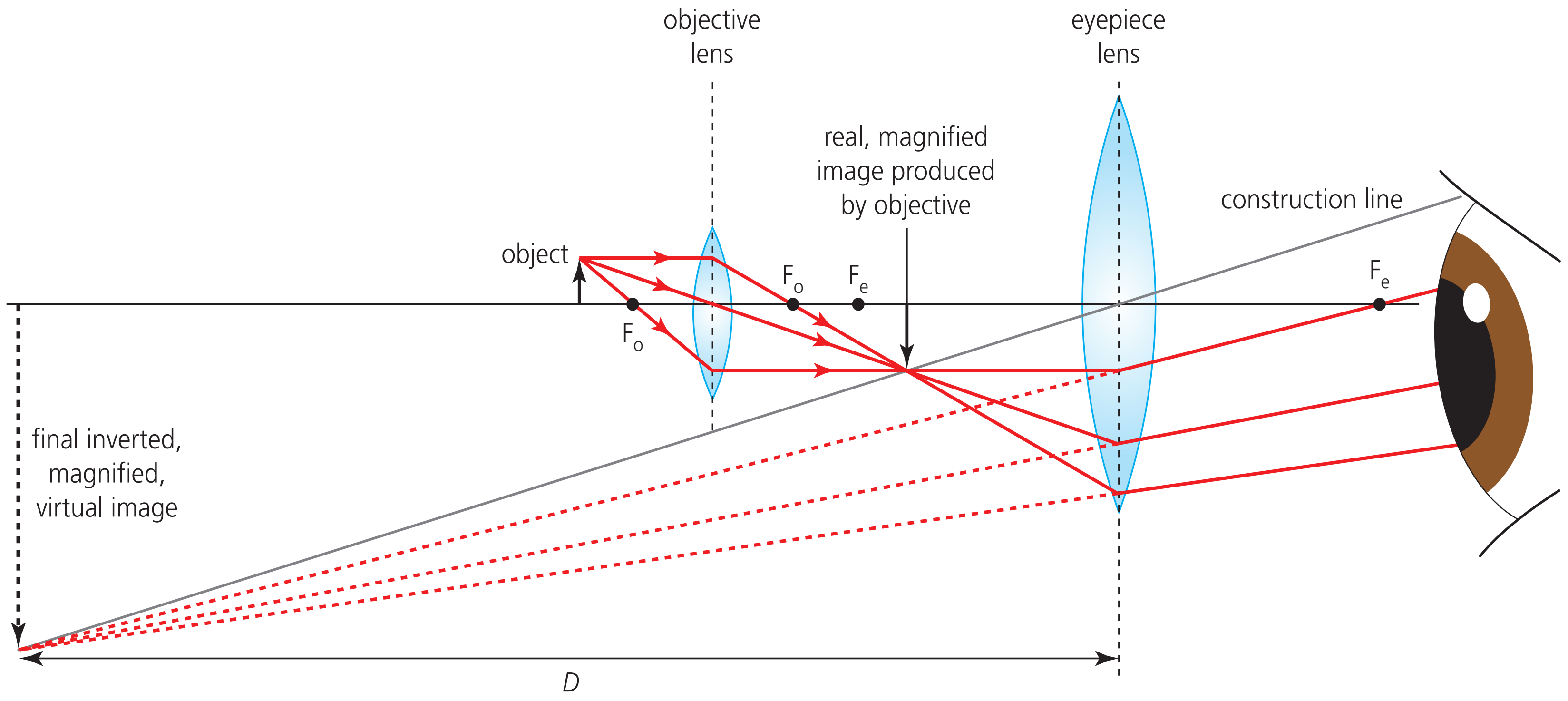


Figure 19.27 Compound microscope in normal adjustment with the final image at the near point

The object to be viewed under the microscope is placed just beyond the focal point of the objective lens, so that a real image is formed between the two lenses with a large linear magnification. The eyepiece is then used as a magnifying glass, and its position is adjusted to give as large an image as possible with the final virtual image usually at, or close to, the near point of the eye. This is called **normal adjustment**. To locate the image by drawing it is necessary to find the point where the *construction line* through the centre of the eyepiece from the top of the first image meets the extension of the ray from the first image which passes through the focal point.

The angular magnification produced by a compound microscope is equal to the product of the angular magnifications of the individual lenses, but in practice aberrations of the individual lenses (see page 759) prevent very large magnifications.

Astronomical telescope

G.2.12 Construct a diagram for an astronomical telescope with the final image at infinity (normal adjustment).
G.2.13 State the equation relating angular magnification to the focal lengths of the lenses in an astronomical telescope in normal adjustment.

A telescope is an optical instrument designed to produce an angular magnification of a distant object. Images may be formed by the processes of refraction or reflection, but in this course we will only discuss refracting telescopes. The linear magnification of a telescope is always very much less than one – the image is always a lot smaller than the object. In a simple refracting **astronomical telescope** there are two lenses which together produce an *inverted*, virtual image. Such a telescope would be of little use for looking at objects on Earth; that is why it is described as 'astronomical' – for use in astronomy.

Figure 19.28 shows the basic construction of an astronomical telescope. The light rays arriving at the objective lens can be considered to be parallel because the source of light is such a large distance away. Consequently, a small, inverted, real image will be formed at the focal point of the objective lens. The eyepiece is then used as a magnifying glass to magnify this image.

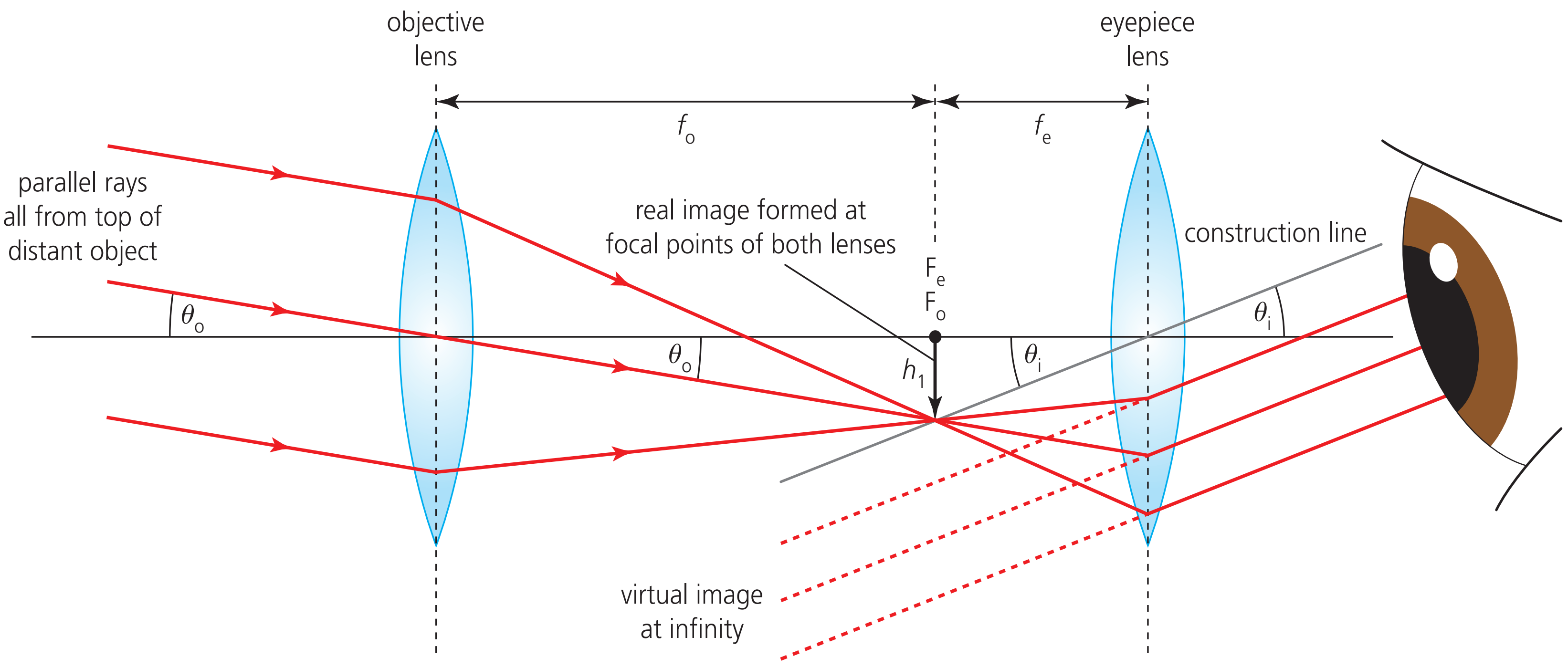


Figure 19.28 Astronomical telescope in normal adjustment with final image at infinity

The telescope is usually adjusted so that the final image is at infinity and the eye can be relaxed when observing it. This is called using the telescope in 'normal adjustment' and the image formed by the objective must be formed at the focal point of the eyepiece. When adjusted in this way, the distance between the lenses is the sum of their focal lengths. The direction to the top of the final image is located by drawing a construction line through the centre of the eyepiece from the top of the first image.

The angular magnification (in this adjustment) can be determined by examining the two triangles involving h_1:

$$M = \frac{\theta_i}{\theta_o} = \frac{h_1 / f_e}{h_1 / f_o} \quad \text{(for small angles)}$$

$$M = \frac{f_o}{f_e}$$

This equation is given in the IB *Physics data booklet*.

Clearly, a greater magnification is obtained by using an objective with a longer focal length (lower power) and an eyepiece with a smaller focal length (higher power). But lens aberrations (see page 759) of a high power eyepiece limit the angular magnification possible with good quality images.

The diameter of the objective lens is also of importance when considering the quality of the image. A larger lens has two advantages: (i) most importantly, it collects more light to produce a brighter image and (ii) there will be less diffraction, which otherwise reduces the resolution of images (as discussed in Chapter 11). However, larger lenses also have abberation problems.

G.2.14 Solve problems involving the compound microscope and the astronomical telescope.

33 A compound microscope has an eyepiece of focal length of 15 cm which is located 16 cm from the objective of focal length 1.2 cm. An object of height 0.5 cm is placed 1.5 cm from the objective. Draw an accurate ray diagram to represent this arrangement and determine the position and size of the final image.

34 The eyepiece of a microscope has a focal length of 12 cm and the objective has a focal length of 2.0 cm. When used in normal adjustment the separation of the two lenses is 15 cm. Calculate:
- **a** the distance from the objective of the image formed by the eyepiece
- **b** the angular magnification of the eyepiece
- **c** the distance of the object from the objective
- **d** the angular magnification caused by the objective
- **e** the overall angular magnification of the microscope.

35 **a** The diameter of the Moon is 3474 km. What angle in radians does it subtend at our eyes when the separation of the Earth and the Moon is 378 000 km?
b What angle in degrees is subtended by the image when using a telescope with an objective of focal length 60 cm and eyepiece of focal length 4.0 cm in normal adjustment?

36 A telescope has two lenses, one of focal length 75 cm and the other of focal length 15 cm.
a What is the distance between the lenses when the telescope is in normal adjustment?
b What is the angular magnification?
c It is suggested that the telescope could be improved by increasing the diameter and power of the lenses. Comment on these suggestions.
d Explain why this may not be a good idea.

Aberrations

G.2.15 **Explain** the meaning of spherical aberration and of chromatic aberration as produced by a single lens.
G.2.16 **Describe** how spherical aberration in a lens may be reduced.
G.2.17 **Describe** how chromatic aberration in a lens may be reduced.

Aberration is the term we use to describe the fact that, with real lenses, all the light coming from the same place on an object does *not* focus in exactly the same place on the image (as simple optics theory suggests). There are two principal kinds of aberration: **spherical** and **chromatic**.

Figure 19.29 represents spherical aberration. Spherical aberration is the inability of a lens, which has surfaces that are spherically shaped, to focus rays which strike the lens at different distances from the principal axis to the same point.

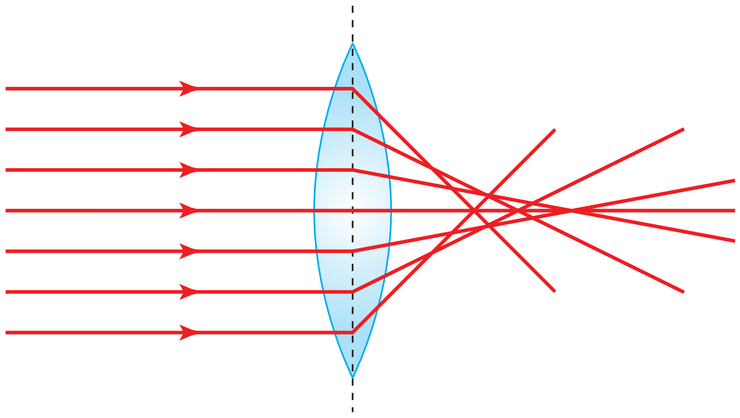


Figure 19.29 Spherical aberration of monochromatic light (exaggerated)

Spherical aberration results in unwanted blurring and distortion of images (see Figure 19.30), but in good quality lenses the effect is reduced by adjusting the shape of the lens. However, this cannot completely remove aberration for all circumstances. The effects can also be reduced by only letting light rays strike close to the centre of the lens. In photography the size of the aperture (opening) through which light passes before it strikes the lens can be decreased to reduce the effects of spherical aberration. This is commonly known as 'stopping down' the lens, but it has the disadvantage of reducing the amount of light passing into the camera and may also produce unwanted diffraction effects.

Figure 19.31 represents chromatic aberration. Chromatic aberration is the inability of a lens to refract light of different colours (wavelengths) to the same focal point. As we have discussed, a medium has slightly different refractive indices for light of different frequencies, so that white light may be dispersed into different colours when it is refracted. Typically, chromatic aberration leads to the blurring of images and gives images red or blue/violet edges.

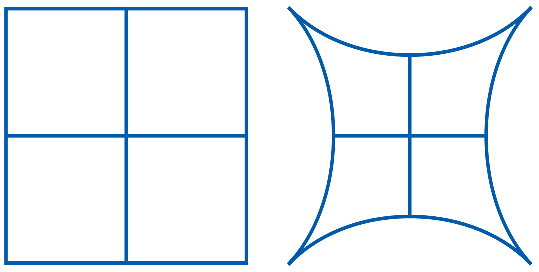


Figure 19.30 Typical distortion produced by spherical aberration (exaggerated)

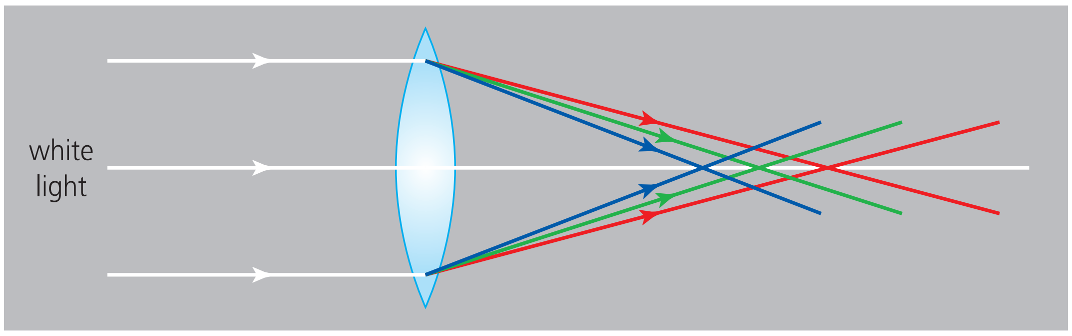


Figure 19.31 Chromatic aberration

Chromatic aberration can be reduced by combining two or more lenses together. For example, a converging lens can be combined with a diverging lens (of different refractive index), so that the second lens eliminates the chromatic aberration caused by the first (see Figure 19.32).

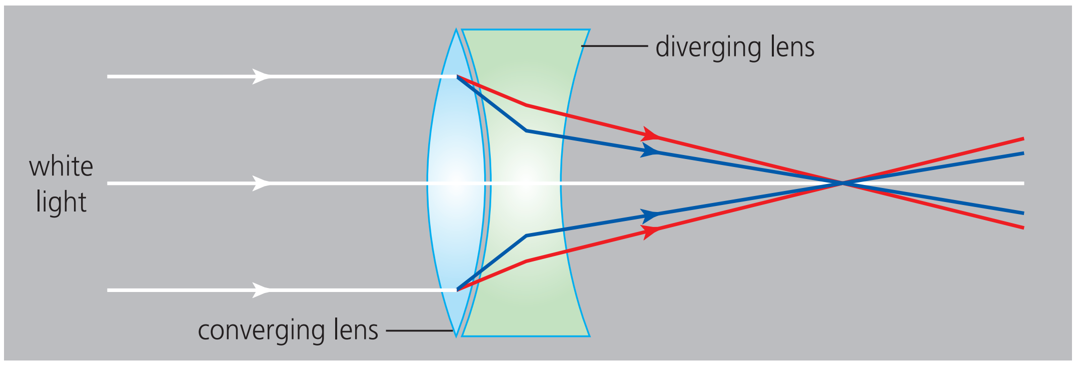


Figure 19.32 Combining lenses of different refractive indices to correct for chromatic aberration

Figure 19.33 This lens achieves top quality images by having a large number of lens elements

In the modern world we are surrounded by optical equipment capable of taking video and still pictures. The quality of the lenses has improved enormously in recent years, as well as the ways in which the images are detected. The quality of the images produced by the best modern camera lenses is highly impressive (Figure 19.33).

37 Make a copy of Figure 19.31 and show on it where a screen would have to be placed to obtain an image which had blue edges.

38 Draw a diagram(s) to illustrate the improved focusing achieved by stopping down a lens.

39 Suggest why lens aberrations tend to be worse for higher power lenses.

G3 Two-source interference of waves

G.3.1 **State** the conditions necessary to observe interference between two sources.
G.3.2 **Explain**, by means of the principle of superposition, the interference pattern produced by waves from two coherent sources.

The interference of waves was introduced in Chapter 4, where it was explained that the necessary condition for interference is that the sources are *coherent*.

Now would be a good point to review the relevant section of Chapter 4, but the important principles will be quickly repeated here. Students are recommended to make use of some of the many computer simulations available on interference and diffraction.

Consider Figure 19.34, which shows two coherent sources, S_1 and S_2, and the waves that they have emitted.

The solid blue lines represent the crests of waves and the troughs of the waves will be midway between them. At a point like P, two crests are coming together, but at the same point a short time later, two troughs will come together. The waves arriving at points like P will always arrive *in phase* and, using the *principle of superposition*, we can determine that *constructive interference* occurs. At points like Q, the two sets of waves

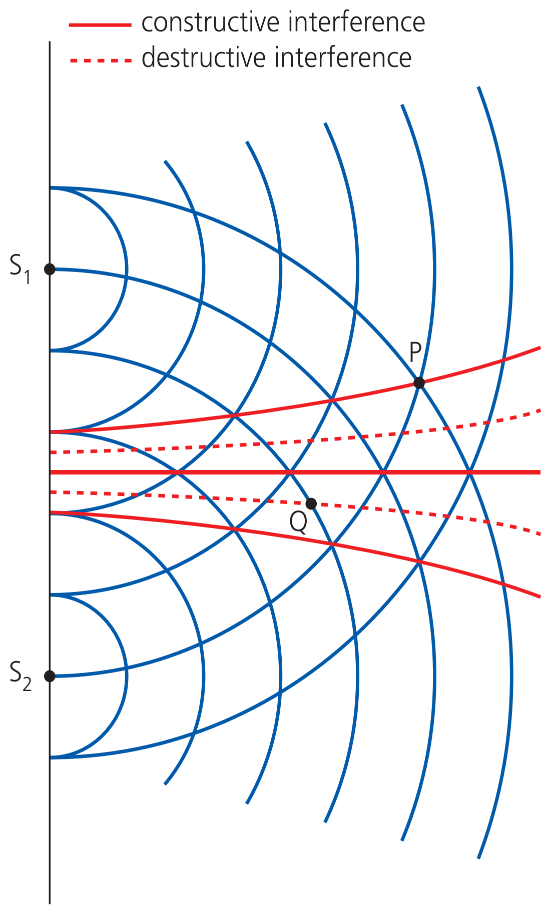


Figure 19.34 Two sets of coherent waves crossing each other to produce an interference pattern

will always be exactly *out of phase* and *destructive interference* occurs at all times. At other places, the interference will not be perfectly constructive or perfectly destructive. In this way progressive waves of any kind from coherent sources can set up stationary interference patterns. Figure 19.35 shows a photograph of an interference pattern produced by water waves in a ripple tank. (A sketch of a similar pattern is shown in Figure 4.64 in Chapter 4.) The water waves are coherent because both dippers producing the waves are being vibrated by the same motor.

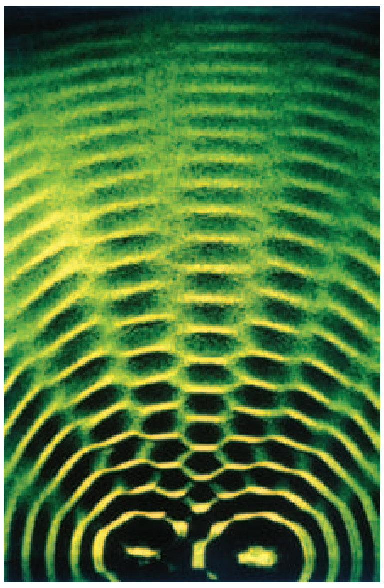


Figure 19.35 Interference pattern with water waves

Interference of light

G.3.3 Outline a double-slit experiment for light and draw the intensity distribution of the observed fringe pattern.

Light waves are *not* usually coherent, so the interference of light is not a phenomenon of which we are usually aware. Light from a laser (which is both monochromatic and coherent) is the ideal source for producing interference patterns, but interference can also be produced without lasers.

Rather than use two separate sources, light from a single source is split into two parts using two narrow slits, which must be close together. The waves passing through each slit then act like separate coherent sources as the waves spread away from the slits because of diffraction (Figure 19.36). In general, we know that for diffraction to be significant, the size of the gap needs to be comparable to the size of the wavelength, so the gaps must be very narrow (since the wavelength of light is so small).

The original source of light should be as bright as possible and as small as possible (that is why a slit is also placed in front of the source). Ideally the light should be monochromatic but a white light source can also be used, perhaps with a filter across the slit to reduce the range of wavelengths.

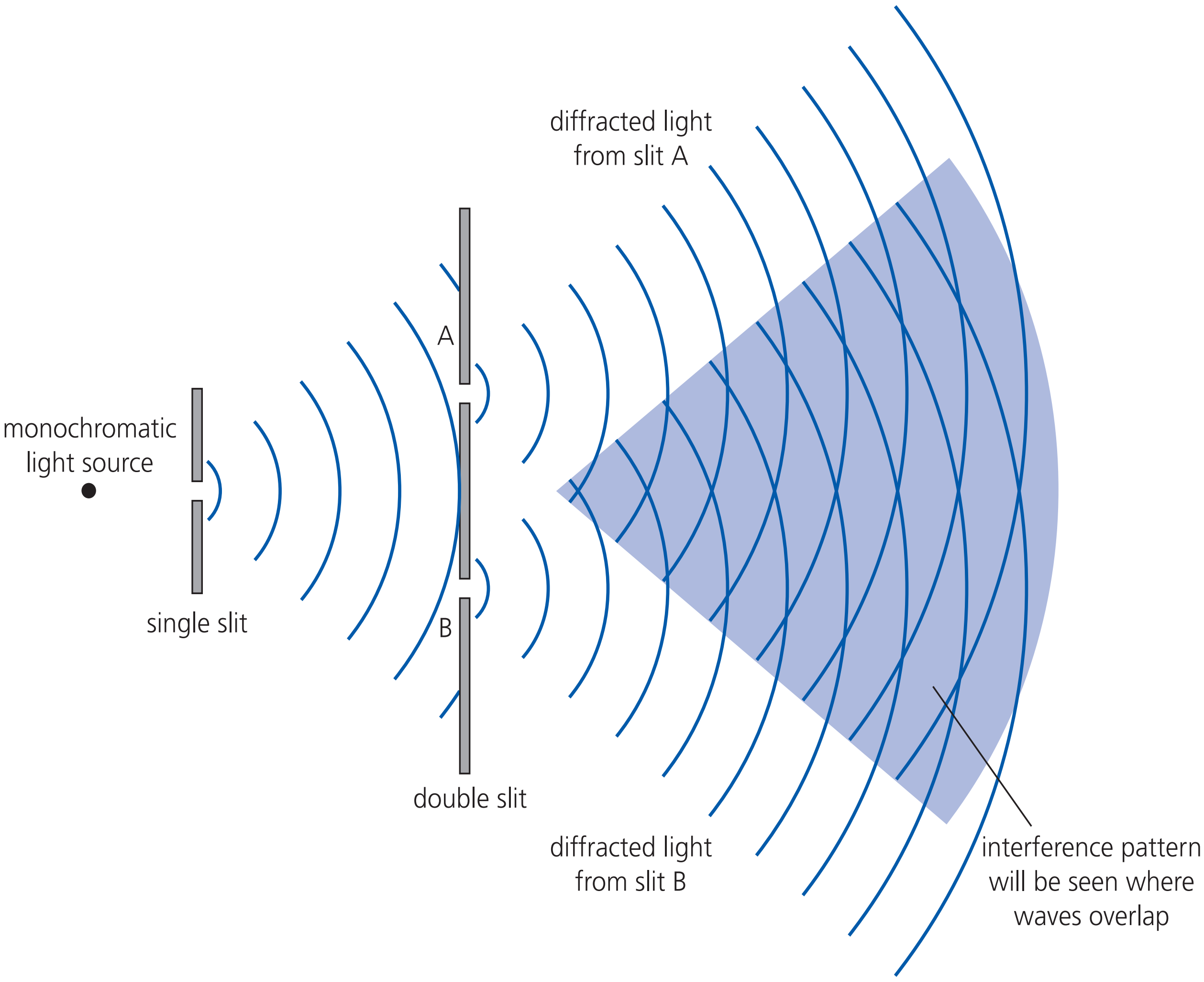


Figure 19.36 Using two slits to produce an interference pattern

If the source is not monochromatic, different colours will produce slightly different patterns and this will blur and confuse what is seen, although it will be brighter. The top image in Figure 19.37 shows a typical pattern of **fringes** produced with monochromatic light, and the bottom image shows what may be seen when white light is used and is slightly dispersed into its colours.

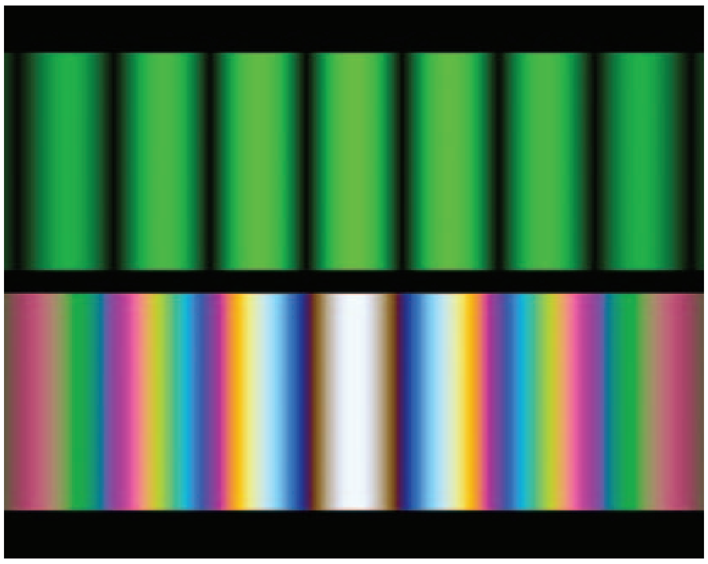


Figure 19.37 Double slit interference patterns of monochromatic light (top line) and white light (bottom line)

The arrangement shown in Figure 19.36 is known as the 'double-slit' experiment, or sometimes as 'Young's experiment', named after Thomas Young, who was the first to carry out this experiment (results published in 1803). This classic physics experiment was very important because it confirmed the wave nature of light (only waves can interfere).

Predicting where constructive interference occurs

To derive an equation which predicts exactly where the interference will be constructive (or destructive) we need to consider *path differences*. In Figure 19.38, rays are used to represent the directions of wave travel and it is clear that the waves arriving at a point P from S_2 have travelled further than the waves arriving from S_1. We say that there is a **path difference** between the waves ($S_2P - S_1P$).

Constructive interference occurs if the path difference between the waves is zero, one wavelength, or two wavelengths, or three wavelengths, etc.

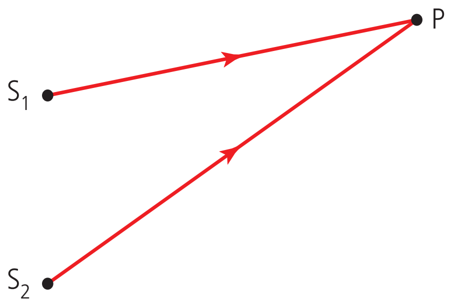


Figure 19.38 Path difference

The condition for constructive interference is that the path difference = $n\lambda$ (where n is an integer: 1, 2, 3, 4, 5, etc.).

The condition for destructive interference is that the path difference equals an odd number of half wavelengths. Path difference = $(n + \frac{1}{2})\lambda$.

These equations for path difference are given in the IB *Physics data booklet* (for Chapter 4).

If we want to predict the exact locations of constructive and destructive interference we must link path differences to the dimensions of the apparatus. Consider Figure 19.39, which shows two identical, very narrow slits separated by a distance d (between the centres of the slits). Note that d is significantly greater than the widths of the two slits. The arrowed lines represent rays showing the direction of waves leaving both slits at an angle θ to the normal. Suppose that these waves meet and interfere *constructively* at a distant point. Of course, if the rays are perfectly parallel they cannot meet, so we must assume that, because the point is a long way away (compared to the slit separation) that the rays are *very nearly* parallel.

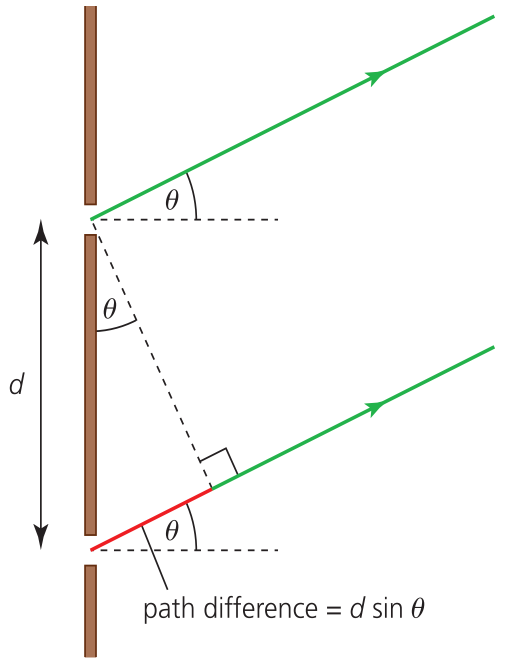


Figure 19.39 Explaining path difference = $d \sin \theta$

From the triangle in Figure 19.39, we can see that the path difference between the rays is $d \sin \theta$ and we know that this must equal $n\lambda$ because the interference is constructive:

$$d \sin \theta = n\lambda$$

Rearranging, the following equation predicts the angles at which constructive interference occurs:

$$\sin \theta = \frac{n\lambda}{d}$$

This equation is given in the IB *Physics data booklet*.

The angle θ is usually small, so that $\sin \theta \approx \theta$ in radians, and the equation can also be written as $\theta \approx n\lambda/d$.

Destructive interference will occur at angles such that $\sin \theta = (n + \frac{1}{2})\frac{\lambda}{d}$.

Figure 19.40 links these equations to a sketch of how the intensity varies across the centre of a fringe pattern. For small angles, the horizontal axis could also represent the distance along the screen (as an alternative to $\sin \theta$ or θ in radians). (This equation and graph can be easily confused with similar work on *single* slit diffraction, covered in Chapter 11.)

In an actual experiment we cannot measure angles directly, so measurements are taken from the screen on which the interference pattern is observed. See Figure 19.41, in which s represents the distance between the centres of adjacent interference maxima (or minima). Here s can be considered to be constant along the screen as long as the angle θ is small enough for $\sin \theta$ to be approximately equal to θ and $\tan \theta$. This also means that the *angular separation* of adjacent fringes can be assumed to be constant.

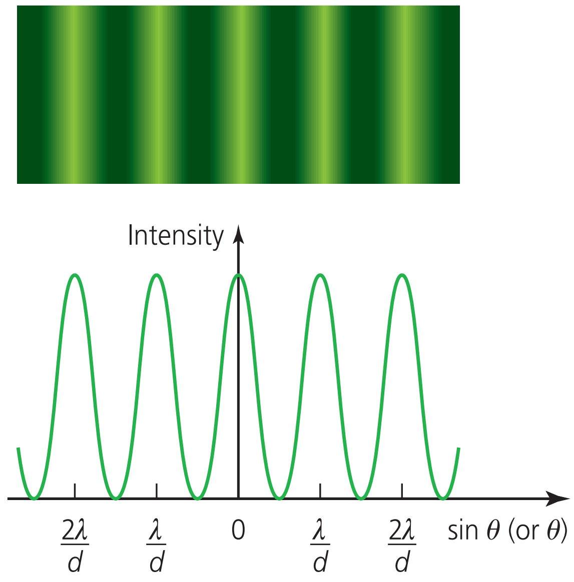


Figure 19.40 Variation of intensity with angle for double-slit interference

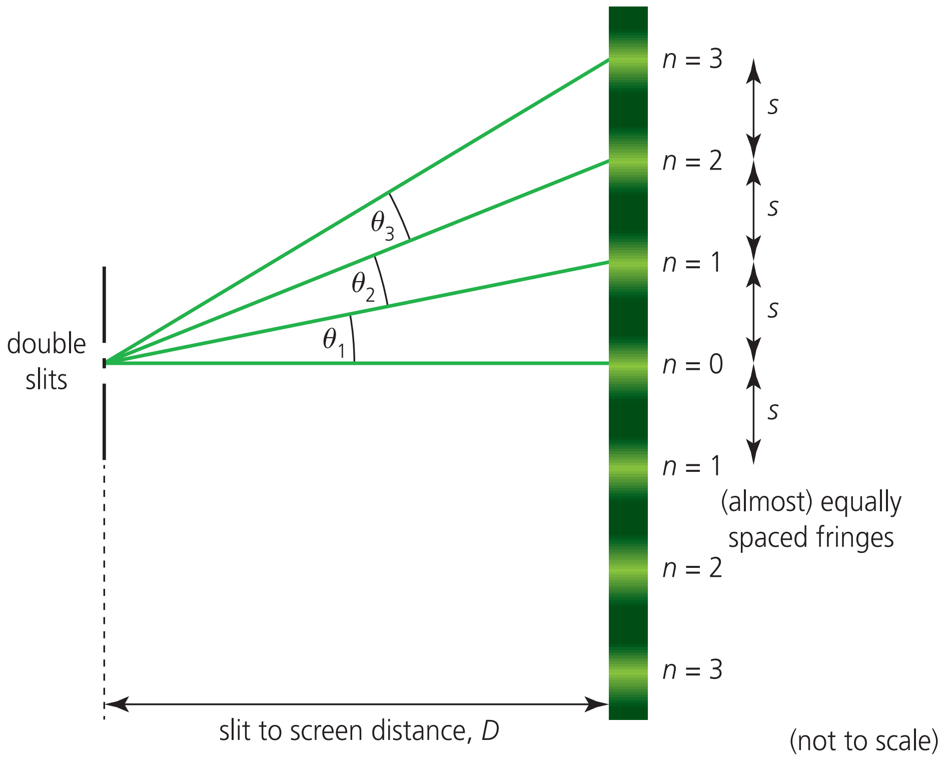


Figure 19.41 Separation and numbering of fringes seen on a screen

The angles at which constructive interference occurs are:

$$\sin\theta_1 = 1 \times \frac{\lambda}{d} \qquad \sin\theta_2 = 2 \times \frac{\lambda}{d} \qquad \sin\theta_3 = 3 \times \frac{\lambda}{d} \qquad \text{etc. } (\sin\theta_n = \frac{n\lambda}{d})$$

In general, if x is the separation between the centres of n bright fringes at small angles:

$\sin\theta_n \approx \frac{x}{D} = \frac{n\lambda}{d}$ for angles at which constructive interference occurs.

If $n = 1$ then x becomes s, so that this can be rewritten as:

$$s = \frac{\lambda D}{d}$$

This equation is given in the IB *Physics data booklet*.

Worked example

5 When monochromatic light of wavelength 5.9×10^{-7} m was directed through two narrow slits at a screen 2.4 m away, a series of light and dark fringes was seen. If the distance between the centres of adjacent bright fringes (and adjacent dark fringes) was 1.1 cm, what was separation of the slits?

$$s = \frac{\lambda D}{d}$$

$1.1 \times 10^{-2} = (5.9 \times 10^{-7}) \times 2.4/d$

$d = 1.3 \times 10^{-4}$ m

G.3.4 Solve problems involving two-source interference.

40 a Light of wavelength 450 nm passes through two slits which are 0.10 mm apart. Calculate the angle to the first maximum.
b Sketch a graph of the intensity variation across the interference pattern.
c How far apart will the fringes be on a screen which is placed 3.0 m from the slits?

41 Give reasons why it is difficult to observe the interference of light from a household light bulb using double slits.

42 The angular separation between the centres of the first and the fourth bright fringes in an interference pattern is 1.23×10^{-3} rad. If the wavelength used was 633 nm, calculate the separation of the two slits which produced the pattern.

43 A Young's double-slit experiment was set up to determine the wavelength of the light used. The slits were 0.14 mm apart and the bright fringes on a screen placed 1.8 m away were measured to be 6.15 mm apart.
a What was the result of the experiment?
b Explain what would happen to the separation of the fringes if the experiment was repeated in water (refractive index = 1.3).

44 In a school experiment to demonstrate interference, two slits each of width 3 cm are placed in front of a source of microwaves. The distance between the slits is 12 cm.
a Suggest why the slit widths were chosen to be 3 cm.
b Some distance away, a microwave detector is moved along a line which is parallel to a line joining the slits. The detector moves 24 cm between two adjacent maxima. Estimate the *approximate* distance between the detector and the slits.

45 The interference pattern produced by double slits with monochromatic light is observed on a screen. What will happen to the pattern if each of the following changes is made (separately)?
a Light of a longer wavelength is used.
b The slit to screen distance is increased.
c The slits are made closer together.
d White light is used.

46 A plane is flying at altitude directly between two radio towers emitting coherent EM waves of frequency 6.0 MHz. If the strength of the signal detected on the plane rises to a maximum every 0.096 s, what is the speed of the plane?

G4 Diffraction grating

Multiple-slit diffraction

G.4.1 Describe the effect on the double-slit intensity distribution of increasing the number of slits.

As we have seen, a pattern of interference fringes can be formed using a beam of light incident on just two slits. If the number of slits (of the same width and separation) is increased, the observed diffraction/interference pattern will have an increasing intensity, but the maxima will still occur at the same angles because the wavelength and slit separation have not changed. However, the most important feature to note is that, the maxima become much *sharper* and more accurately located at those particular angles when more slits are involved. This is shown in Figure 19.42, which compares the patterns seen using monochromatic light for two, five and ten slits.

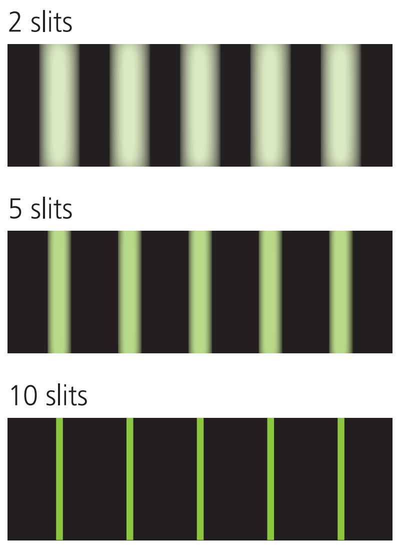


Figure 19.42 How an interference pattern changes as more slits are involved

If the light incident normally on slits is not monochromatic but contains a range of different wavelengths, the maxima of the different wavelengths will occur at slightly different angles, so that the light will be *dispersed*. The different maxima will overlap if a *small* number of slits are used, but the maxima can be separated (*resolved*) by using a *large* number of slits. A very large number of slits close together which is used for dispersing light into different wavelengths is called a **diffraction grating** (see Figure 11.28).

The slits on a diffraction grating are usually called '*lines*'. A typical grating may have 600 lines in every millimetre, which gives a spacing, d, of about 1.7×10^{-6} m. This is approximately three average wavelengths of visible light and it is therefore much smaller than the double-slit separations which we discussed in the previous section. This means that diffraction gratings typically deviate light through much greater angles than a pair of slits.

Diffraction gratings are widely used for analysing light and they offer an alternative to glass prisms, which use the *refraction* (rather than *diffraction*) of light to produce dispersion. A typical application would be the investigation of emission and absorption spectra, as discussed in Chapter 7.

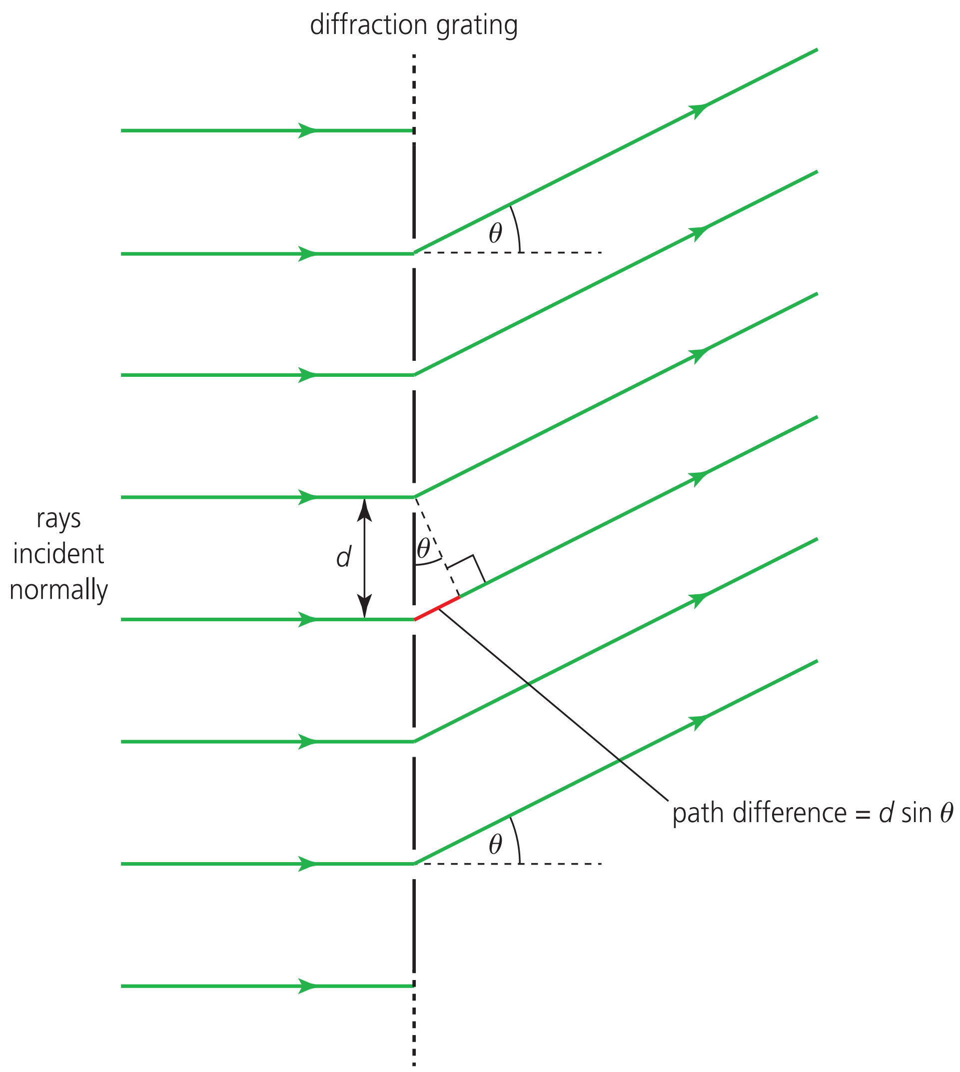


Figure 19.43 Path difference between rays from adjacent slits

Diffraction grating formula

G.4.2 Derive the diffraction grating formula for normal incidence.

The equation $\sin\theta = n\lambda/d$, which predicts the angles at which maxima occur for double-slit interference, can also be used with multiple slits and diffraction gratings. The equation can be derived by considering path differences and using similar principles as were applied to the double-slit situation. Figure 19.43 shows rays emerging at an angle θ from some slits in a small part of a grating. If these rays interfere perfectly constructively at this angle, then the path difference between any, and all, of them must be a whole number of wavelengths ($n\lambda$).

Considering the triangle shown in the diagram, the path difference between rays from adjacent slits must be $d \sin \theta$. If the path difference between rays from adjacent slits is a whole number of wavelengths, then the path difference between rays from *all* slits is a whole number of wavelengths. Therefore, the overall condition for constructive interference is:

$d \sin \theta = n\lambda$ This equation is given in the IB *Physics data booklet*.

Using a diffraction grating to measure wavelength

G.4.3 Outline the use of a diffraction grating to measure wavelengths.

Figure 19.44 shows how a diffraction grating is used to produce an interference pattern (of monochromatic light). In many practical arrangements, a converging lens is used to focus the light on the screen, but that is not shown here. The first maximum out from the middle of the pattern is called the first **diffraction order**, the second is called the second order, and so on.

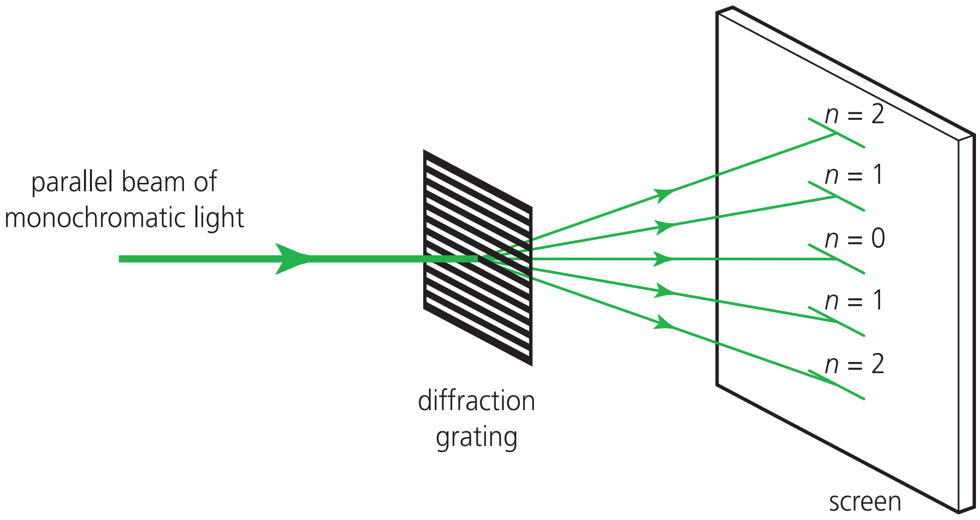


Figure 19.44 Action of a diffraction grating to produce different orders

Figure 19.45 shows sets of rays emerging in different directions. All waves transmitted along the normal will interfere constructively because there is no path difference between them. This is called the zeroth order, $n = 0$.

At angle θ_1, the rays again interfere constructively because there is a path difference of one wavelength between adjacent rays ($n = 1$). This is called the *first order*. The *second order* occurs at an angle θ_2 when there is a path difference of two wavelengths between rays from adjacent rays ($n = 2$). Gratings are usually designed using small slit separations to spread the light out as much as possible and our attention is normally concentrated on the lower orders. Higher orders are not usually needed.

To determine a wavelength using a diffraction grating, measurements are taken from a screen placed at a known distance away. Figure 19.46 shows the arrangement. Because diffraction gratings are usually used to produce large angular separations we *cannot* usually make the approximation $\sin \theta \approx \theta$ without introducing a significant error.

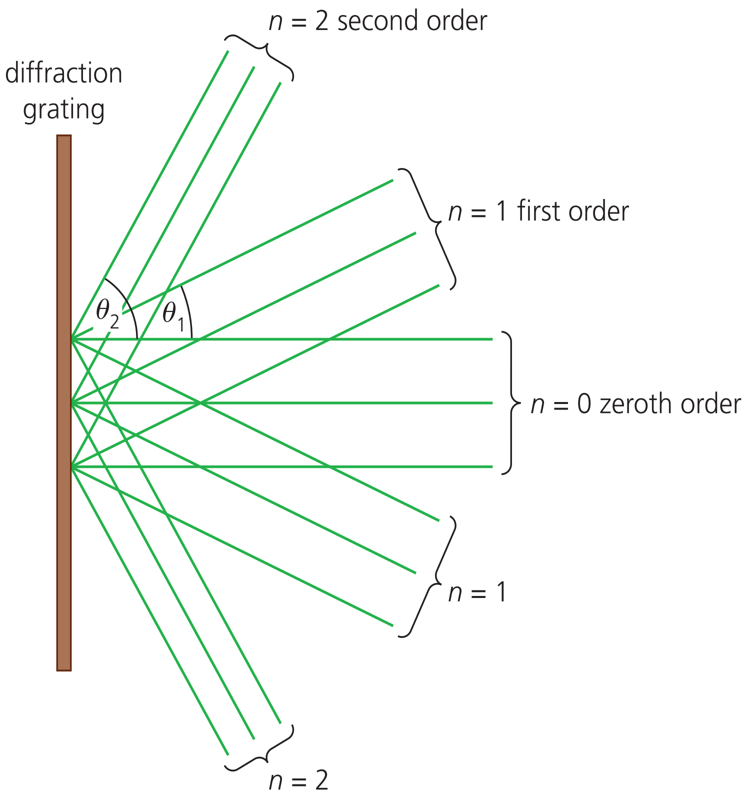


Figure 19.45 Different orders in different directions

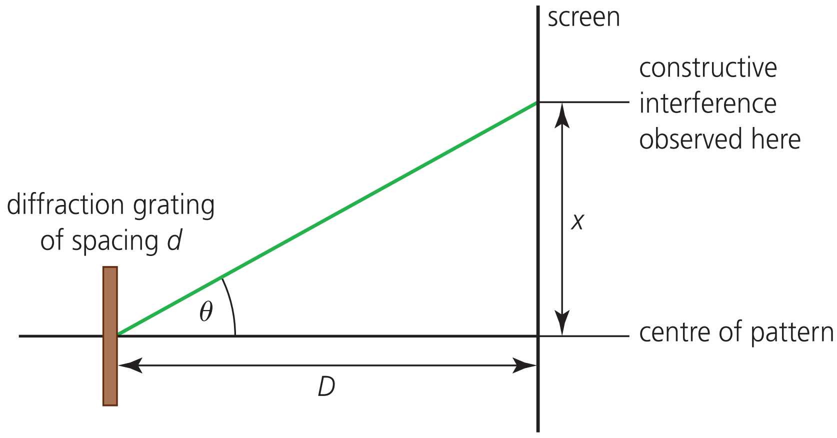


Figure 19.46 Measurements needed to determine wavelength

Worked example

6 Monochromatic light passes through a diffraction grating with 400 lines mm^{-1} and forms an interference pattern on a screen 2.25 m away.

a If the distance between the central maximum and the first order is 57.2 cm, what is the wavelength of the light?

b If a second beam of monochromatic light of wavelength 6.87×10^{-7} m also passes through the grating, how far apart are the first orders on the screen?

a Considering Figure 19.46, $\tan\theta = \frac{0.572}{2.25} = 0.254$, so $\theta = 14.3°$ and $\sin\theta = 0.246$

$d\sin\theta = n\lambda$

$\left(\frac{1.0 \times 10^{-3}}{400}\right) \times 0.246 = 1 \times \lambda$

$\lambda = 6.16 \times 10^{-7}$ m

(Using the $\sin\theta \approx \tan\theta$ approximation produces an answer of 6.36×10^{-7} m.)

b $\sin\theta' = \frac{n\lambda}{d}$

$\sin\theta' = 1 \times \left(\frac{6.87 \times 10^{-7}}{10^{-3} / 400}\right) = 0.275$

So $\theta' = 16.0°$ and $\tan\theta' = 0.286$

$\tan\theta'$ = distance between the central maximum and the first order divided by the slit to screen distance

Distance between the central maximum and the first order = $0.286 \times 2.25 = 0.643$ m. So the separation of the orders on the screen = (64.3 – 57.2) = 7.1 cm.

Using a diffraction grating to analyse spectra containing different wavelengths

The main reason why diffraction gratings are so useful is because they produce very sharp and intense maxima. Figure 19.47 compares the relative intensities produced by double slits (using monochromatic light) with a diffraction grating of the same spacing. Compare this with Figure 19.42, which showed the visual effect of increasing the number of slits.

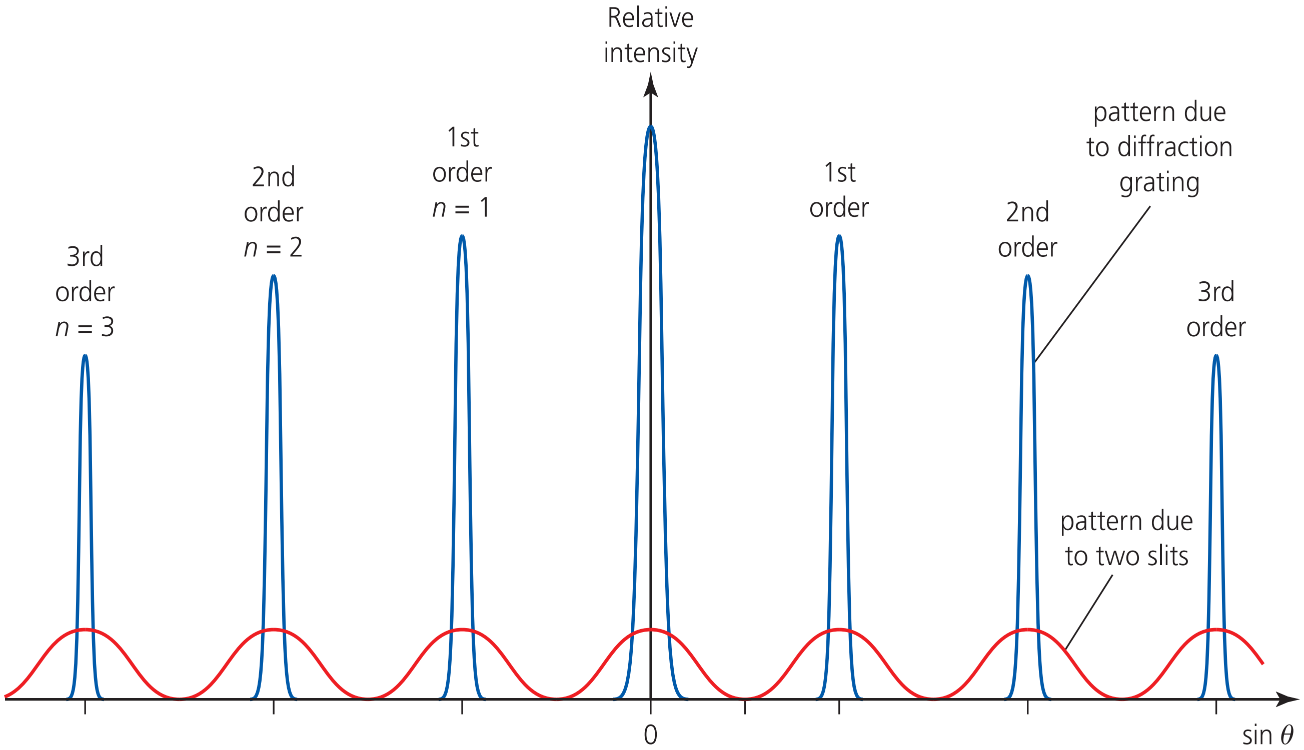


Figure 19.47 Comparing the maxima produced by double slits and a diffraction grating using monochromatic light

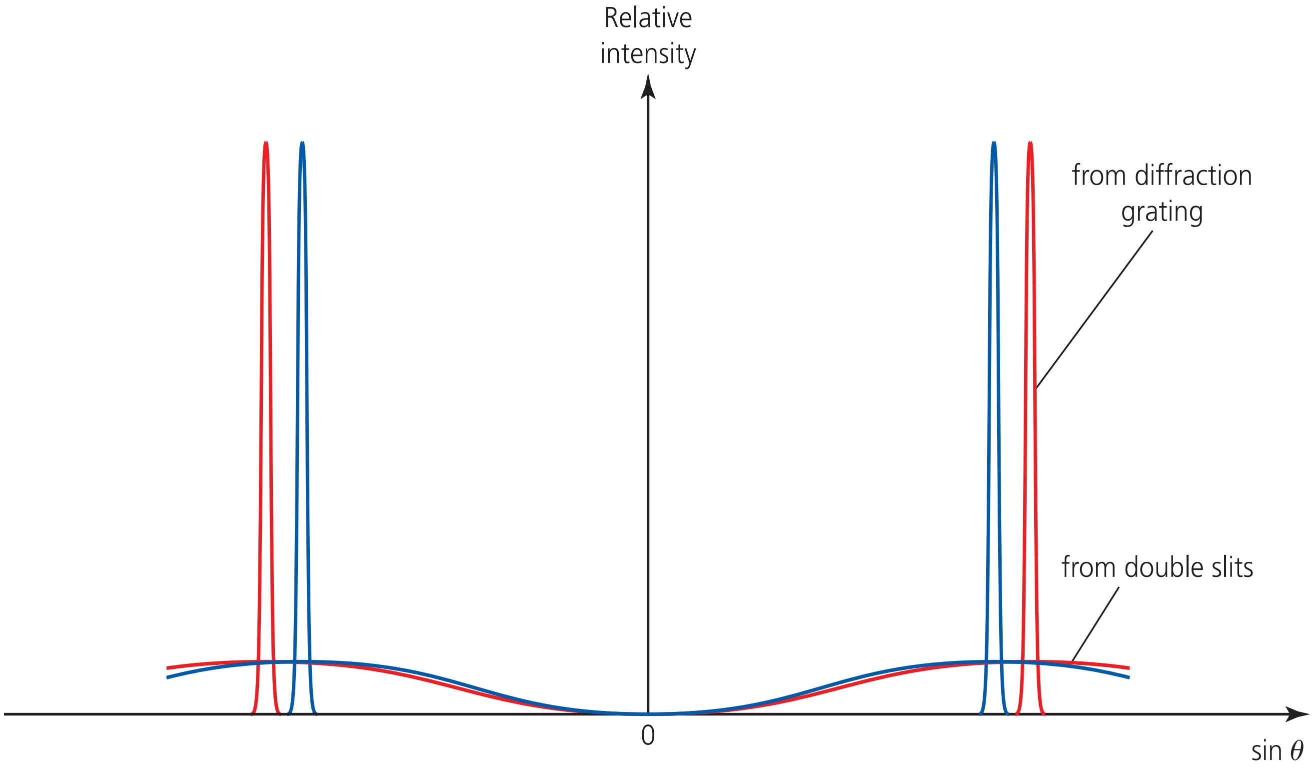


Figure 19.48 High resolution produced by diffraction gratings

If light is incident on a greater number of slits, more waves from different slits will arrive at any particular point on the screen. But, unless the angle is perfect for constructive interference (as represented by the equation $d \sin \theta = n\lambda$), there will be enough waves arriving at that point out of phase to result in overall destructive interference. In practice, light striking a grating will usually be incident upon well over 1000 slits, so that the peaks become very much sharper and more intense than those shown in Figure 19.47 and Figure 19.48.

Diffraction gratings analyse spectra by separating (*resolving*) the maxima of different wavelengths. This is shown in Figure 19.48, which compares the intensities of the first-order spectra from double slits and a diffraction grating (for a spectrum containing only red and blue light).

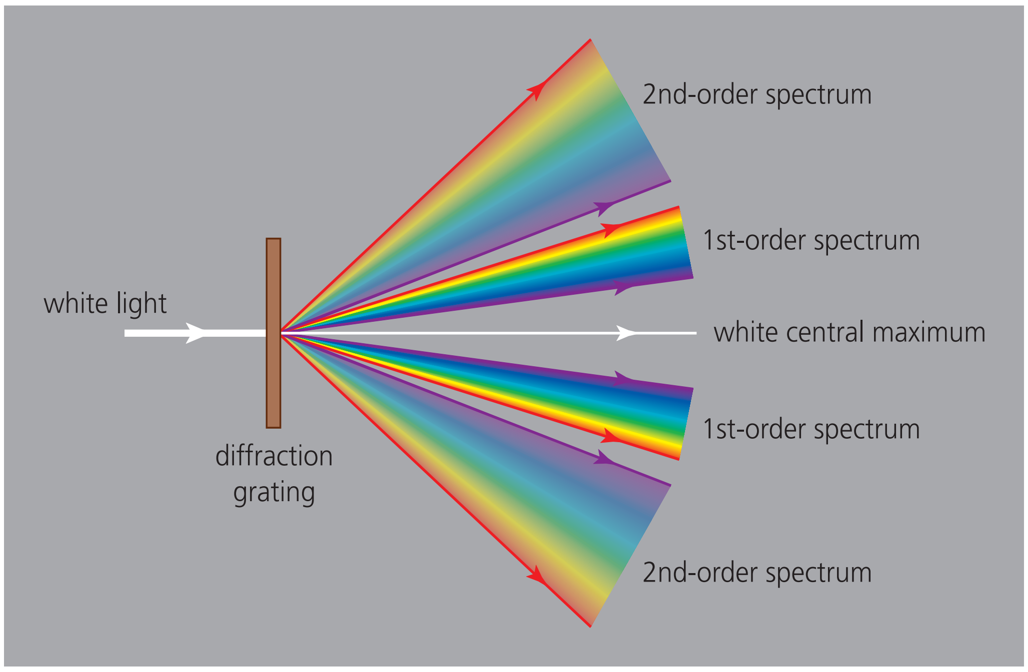


Figure 19.49 White light passing through a diffraction grating

If white light is incident on a grating, a series of continuous spectra is produced, as shown in Figure 19.49. Since the wavelength of red light is less than twice the wavelength of violet light, the first-order and second-order visible spectra cannot overlap, although higher orders do.

If the *emission* or *absorption spectrum* of a substance is being observed, a good quality grating will be able to separate the wavelengths into the different 'lines' of a *line spectrum*, like those shown in Figure 7.10.

The diffraction gratings that we have discussed so far have all been *transmission* gratings, but similar principles can be applied to light *reflected* off a series of very close lines. Reflection gratings may be useful when examining radiation that would otherwise be absorbed in the material of a transmission grating, for example ultraviolet. The manufacture of optical data storage discs (for example, DVDs) produces a regularity of structure that results in them acting as reflection gratings. The colours seen in reflections from their surfaces are a result of this 'diffraction grating effect'. The range of colours seen in the reflection of light off some insects and birds (*iridescence*) can be explained in a similar way.

G.4.4 Solve problems involving a diffraction grating.

47 Light of wavelength 460 nm is incident normally on a diffraction grating with 200 lines per millimetre. Calculate the angle to the normal of the third-order maximum.

48 A diffraction grating is used with monochromatic light of wavelength 6.3×10^{-7} m and a screen a perpendicular distance of 3.75 m away. How many lines per millimetre are there on the diffraction grating if it produces a second-order maximum 38 cm from the centre of the pattern?

49 Monochromatic light of wavelength 530 nm is incident normally on a grating with 750 lines mm^{-1}. The interference pattern is seen on a screen which is 1.8 m from the grating. What is the distance between the first and second orders seen on the screen?

50 **a** A teacher is using a diffraction grating of 300 lines mm^{-1} to show a class a white light spectrum. She wants the first-order spectrum to be 10 cm wide. Estimate the distance that she will need between the grating and the screen.
b A prism can also be used to produce a spectrum. Explain why red light is refracted less than blue light in a prism, but diffracted more by a diffraction grating.

51 Light of wavelength 5.9×10^{-7} m is incident on a grating with 6.0×10^{5} lines per metre. How many orders will be produced?

52 When using white light, explain why red light in the second-order spectrum overlaps with blue light in the third-order spectrum.

53 A second-order maximum of blue light of wavelength 460 nm is sent to a certain point on a screen using a diffraction grating. At the same point the third-order maximum of a different wavelength is detected.
a What is the wavelength of the second radiation?
b In what section of the electromagnetic spectrum is this radiation?
c Suggest how it could be detected.

54 The theory developed in this section has all been for light incident *normally* on a diffraction grating. Suggest what the effect would be on an observed pattern if the grating was twisted so that the light was incident obliquely.

55 Suggest two reasons why an optical diffraction grating would not be of much use with X-rays.

56 Sketch a relative intensity against $\sin\theta$ graph for monochromatic light of wavelength 680 nm incident normally on a diffraction grating with 400 lines per mm.

57 A diffraction grating produced two first-order maxima for different wavelengths at angles of 7.46° and 7.59° to the normal through the grating. This angular separation was not enough to see the two lines separately. What is the angular separation of the same lines in the second order?

58 **a** Make a sketch of the variation of relative intensity with $\sin\theta$ for red light as it passes through a diffraction grating which produces two orders.
b Add to your sketch the variation in intensity of blue light passing through the same grating.
c Draw a separate sketch showing the relative intensities of the same red and blue light after it passes through ten slits (with the same spacing as the grating).
d Use your sketches to help explain why diffraction gratings are so useful for analysing light.

G5 X-rays

X-rays are electromagnetic radiation produced by changes in energy of electrons within atoms. Typical wavelengths for X-rays are between 10^{-9} m and 10^{-11} m.

In recent years technological developments and computerized control have resulted in a significant increase in the use of X-rays in hospitals and for security scanning. Figure 19.50 shows a typical use at an airport. X-rays are useful for these purposes because of their selective absorption: depending on the nature of the medium and the energy of the X-ray photons, X-rays are absorbed by different amounts. X-rays are also widely used by scientists for investigating the structure of crystalline materials.

Figure 19.50 X-ray security at an airport

The production of X-rays

G.5.1 Outline the experimental arrangement for the production of X-rays.

As we have seen, whenever charges accelerate or decelerate, electromagnetic radiation is emitted. The frequency of the emitted radiation depends on the magnitude of the acceleration.

In order to produce X-ray photons, high-speed electrons are very rapidly decelerated. Figure 19.51 shows the basic design of a simple **X-ray tube** used to produce X-rays. This is named a Coolidge tube, after its inventor, William Coolidge (1913).

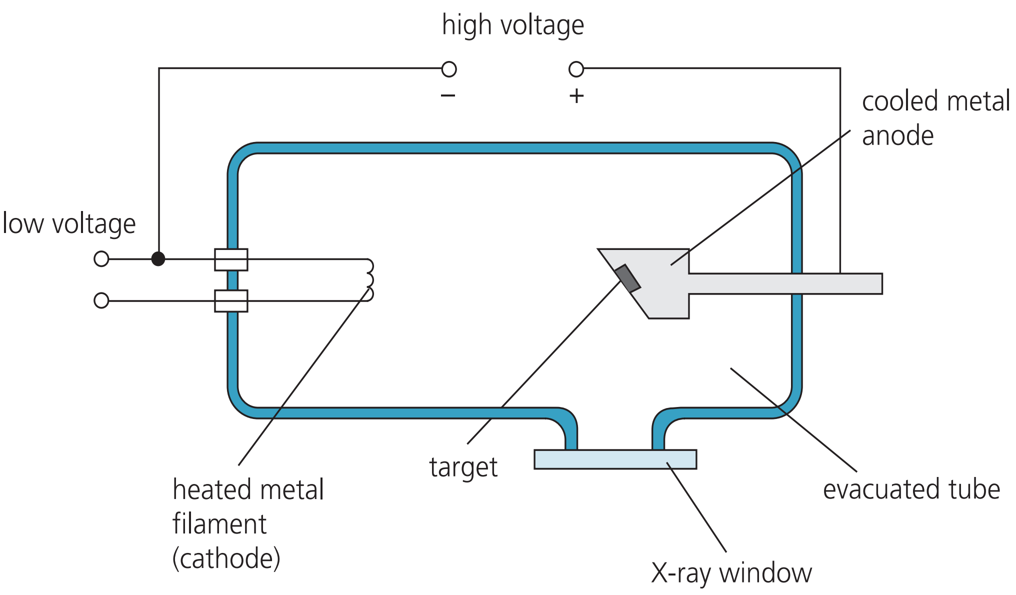


Figure 19.51 X-ray tube

In this basic kind of tube, a relatively low voltage is used to pass a current through the **cathode** to make it hot. High-speed electrons then have enough kinetic energy to escape from the metal's surface, a process known as **thermionic emission**. The electrons are accelerated across the tube by a very high potential difference between the cathode (–) and the 'target', the **anode** (+). When an electron interacts with one or more atoms in the metal anode, it decelerates and emits one or more X-ray photons, which pass out of the transparent 'window'. X-rays cannot pass through any other part of the tube. The tube must be *evacuated* (it contains a vacuum), so that there are no gas molecules for the electrons to collide with as they are accelerated.

An electron of charge e will have electric potential energy in the electric field between the electrodes. As it is accelerated towards the anode by a p.d. V, it will transfer this potential energy to kinetic energy. A typical X-ray tube may use an accelerating voltage of 25 kV. The kinetic energy of an electron arriving at the anode can then be calculated as follows:

energy transferred = charge × p.d.

$$eV = (1.6 \times 10^{-19}) \times (2.5 \times 10^{4}) = 4.0 \times 10^{-15}\,\text{J}$$

More commonly, the electron's energy will be quoted as simply 2.5×10^4 electronvolts (eV) or 25 keV.

Although they produce X-rays, typically most of the kinetic energy of the decelerated electrons is transferred to the atoms of the target in collisions. In this way, the kinetic energy of the metal atoms increases and the target gets hotter. It may become so hot that it needs to be cooled to prevent it overheating or melting.

Intensity and hardness of an X-ray beam

If the current in the cathode is increased, the power will increase, it will get hotter and more electrons will be emitted, so that more X-ray photons will be produced at the anode. In this way the current in the cathode controls the *intensity* (power/area) of the emitted X-rays (for a given accelerating p.d.).

The penetrating power of X-rays is commonly known as their **hardness**. X-rays produced by greater accelerating voltages will be 'harder' because their photons have greater frequencies (smaller wavelengths) and more energy ($E = hf$). Of course, the hardness of the X-rays is chosen to be suitable for their specific purpose, and filters are commonly used to absorb the softer X-rays which are not needed, which might otherwise add to the overall health risk of the X-rays being emitted.

Body scanners used at airports

The X-ray machines that are used to examine the contents of bags and cases, etc. (see Figure 19.50) rely on the selective *absorption* of radiation as it passes through the containers, but human body scanners use the *scattering* of X-rays from the skin. Figure 19.52 shows an example of a human body scanner. The wavelengths of the 'softer' X-rays which are used in back-scattering full-body scanners are selected because they scatter differently off different kinds of materials.

Figure 19.52 X-ray back-scattering scanner

X-ray body scanners are capable of revealing considerable detail of the surface of a human body beneath clothing and this fact has raised many privacy concerns. The computer software which controls the process and the nature of the images produced can be designed to limit how much is revealed, especially of the private parts of the body. However, this obviously also reduces the effectiveness of the security check.

When any part of the body is exposed to X-rays there will be a health risk and the regular and routine scanning of large numbers of people obviously increases that overall risk. However, most studies have suggested that the increased risk to health is almost negligible, although there are some scientists who are more cautious.

Question

1 Suppose that a scientific study suggested that, statistically, the world-wide use of body scanners resulted in an average of an extra 10 premature deaths every year because of cancer. Discuss whether that would be sufficient reason to stop using such scanners.

X-ray spectra

G.5.2 **Draw** and **annotate** a typical X-ray spectrum.

Any particular electron may interact with one or more of the target metal atoms in a variety of ways, so that X-ray photons are emitted with a range of different energies and therefore different wavelengths. Figure 19.53 shows a typical **continuous X-ray spectrum**.

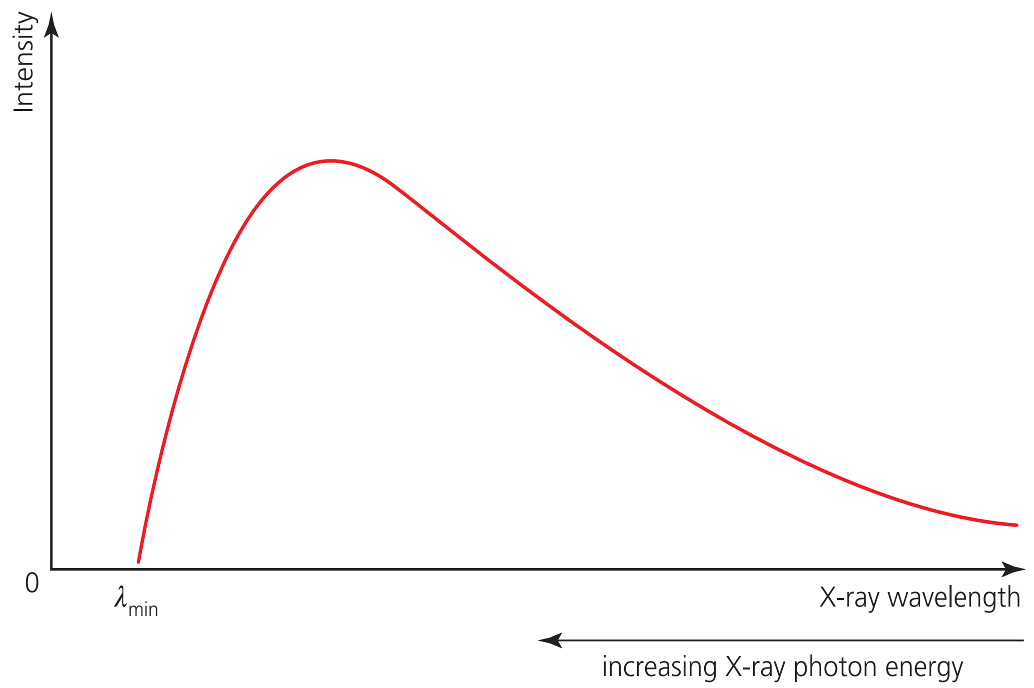


Figure 19.53 Continuous X-ray spectrum (characteristic peaks not shown)

Note that there is a well-defined **characteristic minimum wavelength,** λ_{min} (maximum photon energy), emitted. A photon of *maximum* energy will be released when, after the interaction, *all* of the electron's kinetic energy is transferred to the photon (and none to the atom). In this case:

$$\text{electron's kinetic energy} = \text{photon energy} = hf = \frac{hc}{\lambda}$$

$$eV = \frac{hc}{\lambda_{min}}$$

Or:

$$\lambda_{min} = \frac{hc}{eV}$$

This equation is given in the IB *Physics data booklet*.

Worked example

7 What is the minimum wavelength of X-rays emitted by:

a a 22 kV X-ray tube

b a 44 kV tube (twice the voltage)?

a $\lambda_{min} = \frac{hc}{eV}$

$$\lambda_{min} = \frac{(6.63 \times 10^{-34}) \times (3.00 \times 10^{8})}{(1.60 \times 10^{-19}) \times (2.2 \times 10^{4})}$$

$$\lambda_{min} = 5.7 \times 10^{-11}\,\text{m}$$

b $\lambda_{min} = \frac{5.7 \times 10^{-11}}{2} = 2.8 \times 10^{-11}\,\text{m}$

Characteristic X-rays

G.5.3 **Explain** the origins of the features of a characteristic X-ray spectrum.

Some of the kinetic energy of the electrons in the beam may be used to excite the metal atoms in the target anode, raising their inner electrons to one of several higher energy levels. X-ray photons are emitted when the electrons return to lower energy levels, creating peaks in the X-ray spectrum. Figure 19.54a shows three characteristic peaks as an example. They correspond to the electron transitions shown in Figure 19.54b. Note that the *largest* energy jump corresponds to the characteristic peak with the *smallest* wavelength.

Because atoms of different elements do not have the same energy levels, targets made from different metals do not have the same characteristic peaks, although their continuous spectra are similar (for the same accelerating voltage). The peaks can be used to identify an unknown element.

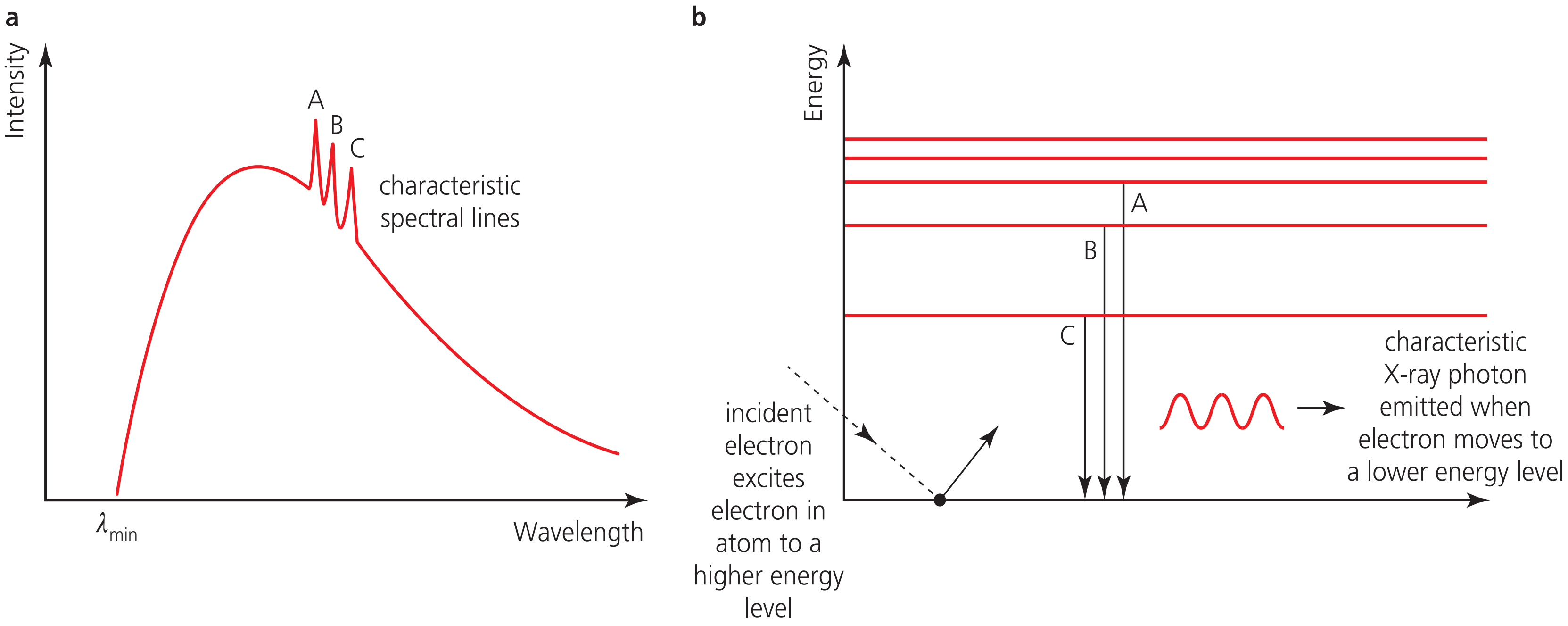


Figure 19.54 Typical X-ray spectrum showing the characteristic peaks and the transitions that produce them

Figure 19.55 compares X-ray spectra from different metal targets (using the same accelerating p.d.) and from the same target using a different accelerating p.d. If there are *no* characteristic peaks on a spectrum, then the maximum kinetic energy of the electrons must be less than the smallest energy transition up from the ground state.

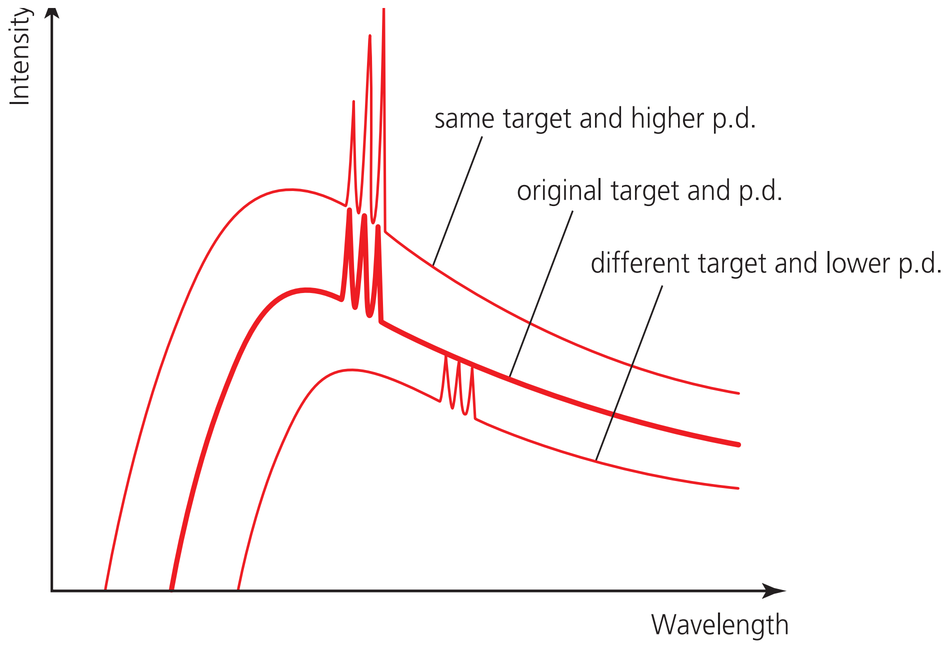


Figure 19.55 Comparing X-ray spectra

G.5.4 Solve problems involving accelerating potential difference and minimum wavelength.

59 a What is the kinetic energy of an electron accelerated by a p.d. of 3.0×10^4 V in:
  i electronvolts
  ii joules?
b What is the final speed of the electron?
c State any assumptions that you made in answering **b**.

60 What is the minimum wavelength of photons produced by a 50 kV X-ray tube?

61 a What voltage is needed across an X-ray tube in order to produce X-rays with a maximum frequency of 6.98×10^{18} Hz?
b What voltage is needed to increase the maximum frequency emitted by 50%?

62 a If the current across a 42 kV X-ray tube is 18 mA and only 1.0% of the electrons' energy is transferred to radiation, calculate the rate at which energy is transferred to internal energy in the target.
b If the target is kept at a constant temperature by water flowing through it at a rate of 245 g min^{-1}, what is the maximum temperature rise of the water? (Specific heat capacity of water = 4180 J kg^{-1} °C^{-1}.)

63 The emitted power of a filtered X-ray beam is 3.82 W. Assuming that all the X-ray photons emitted have a wavelength of 4.17×10^{-11} m, calculate:
a the energy carried by each photon,
b the number of photons emitted every second.

64 a Figure 19.56 shows two energy levels within a tungsten atom. Calculate the wavelength of the X-ray released when an electron in a tungsten atom moves from the higher level to the lower level.
b What could be concluded from an X-ray spectrum that had no characteristic lines?

65 Find out what you can about Henry Moseley's work with characteristic X-ray spectra.

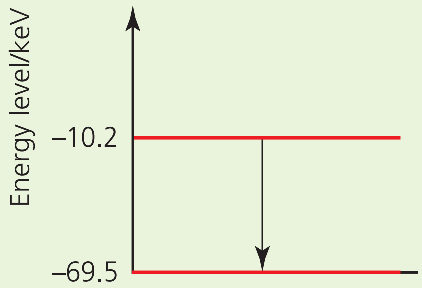


Figure 19.56

X-ray diffraction

G.5.5 Explain how X-ray diffraction arises from the scattering of X-rays in a crystal.

We know that the diffraction of waves is most significant when the wavelength is approximately equal to the size of the object. As we have seen, X-ray wavelengths are typically 10^{-10} m in size which means that atoms (which also have a typical size of 10^{-10} m) and molecules will diffract X-rays well.

X-rays are diffracted into patterns when they pass through a material with atoms, ions or molecules in a regular arrangement. A simple example of a regular structure is the cubic structure of ions in sodium chloride, shown in Figure 19.57. Materials with regularity in their structure are described as **crystalline**. A regular three dimensional arrangement of points in space is called a **lattice** and within a crystal lattice it is possible to identify a number of different sets of *planes* (layers) of ions. In Figure 19.57 the vertical and horizontal planes are easily identified, but Figure 19.58 shows an example of other planes which are not parallel to the 'sides' of a cubic structure.

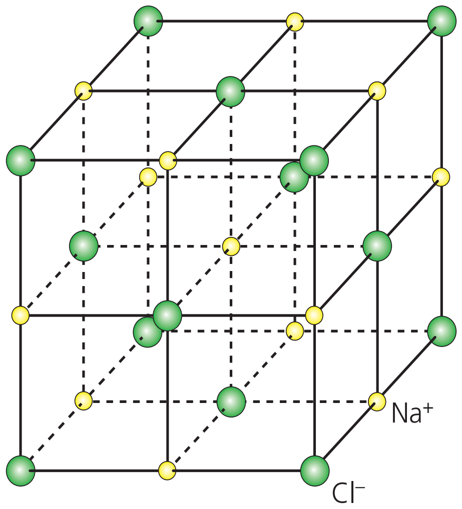


Figure 19.57 Crystalline structure of sodium chloride (Na^+Cl^-) (not to scale)

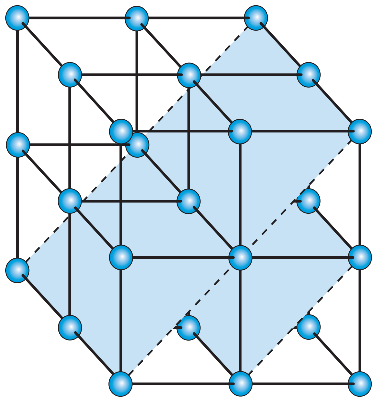


Figure 19.58 Parallel atomic planes in a crystalline structure (not to scale)

To begin our explanation of X-ray diffraction we will first consider how waves of any kind will be diffracted and scattered by a simple regular structure in two dimensions, as shown in Figure 19.59. This shows a series of plane wavefronts passing five objects, each the same size as the wavelength, which are placed in a regular arrangement. The waves are diffracted and scattered from each object into circular wavefronts. The diagram only shows the diffracted waves caused by wavefront W.

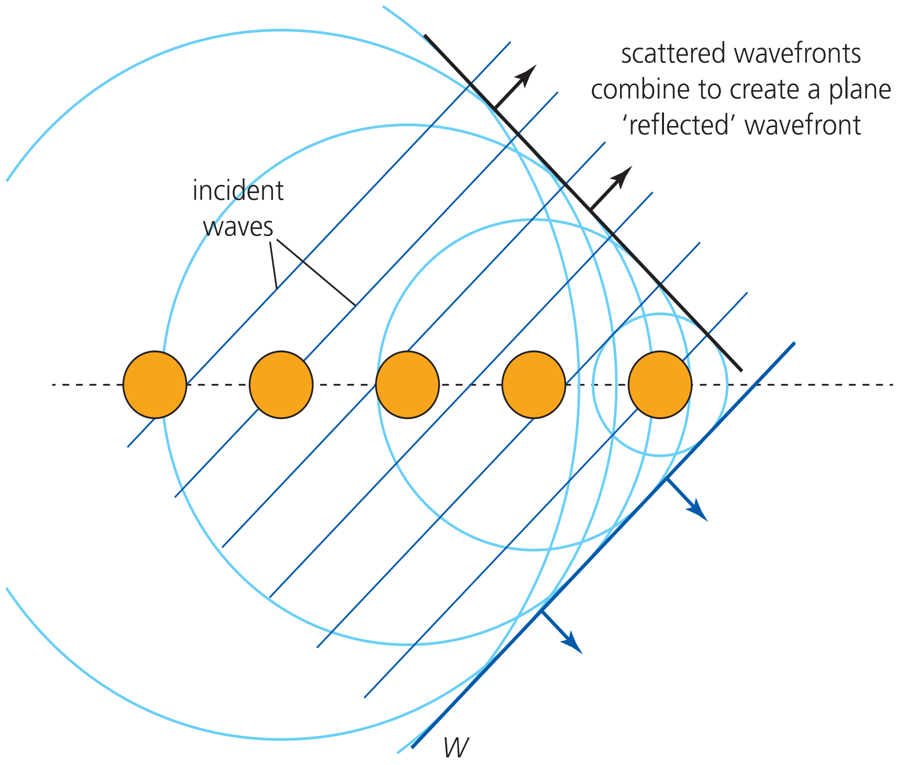


Figure 19.59 Diffraction and scattering of planes waves by a regular structure

The diffracted waves are coherent and will interfere with each other depending on the path differences. The overall effect produces a plane wavefront 'reflected' at the same angle (to a line joining the objects) as the incident wavefronts. In all other directions the overall effect will be destructive interference. Figure 19.60 represents a similar situation using rays. It should be clear that all rays have travelled the same distance, with no path differences between them, so that constructive interference occurs when the angles of incidence and 'reflection' are equal.

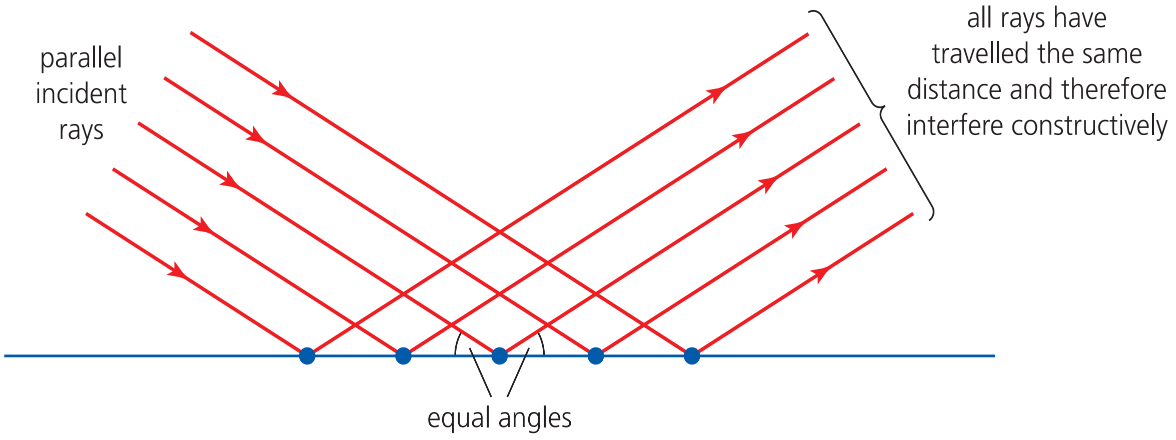


Figure 19.60 'Reflection' of rays by a regular structure

Bragg scattering equation for X-ray diffraction

G.5.6 Derive the Bragg scattering equation.

X-rays are very penetrating and they will be scattered from a large number of planes of atoms, molecules or ions as they pass through a crystalline material. This means that we must consider how X-rays that are scattered from different planes within the crystal lattice interfere with each other. Figure 19.61 (overleaf) shows the 'reflected' rays from the top four (of a very large number of) planes which have a separation of d. The rays are incident at an angle θ to the layers.

Constructive interference will occur if the path difference between rays reflected from adjacent layers is equal to a whole number of wavelength ($n\lambda$). Consider rays reflected off the top two layers: the path difference is the length ACB, and by considering the (similar) triangles

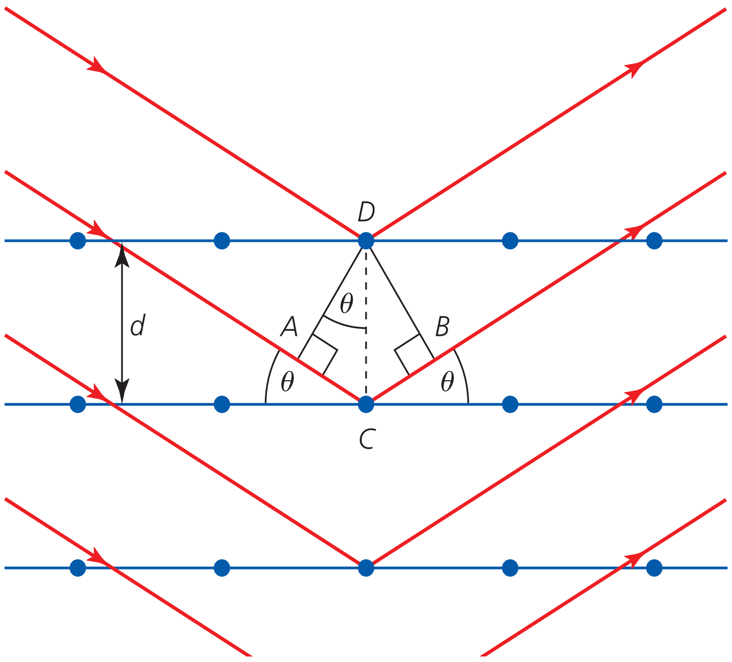


Figure 19.61 Path difference of reflected rays

ACD and BCD, we can determine that $AC = CB = d\sin\theta$. So $ACB = 2d\sin\theta$.

Therefore the condition for constructive interference becomes:

$$2d\sin\theta = n\lambda$$

This equation is given in the IB *Physics data booklet*.

This equation is known as the **Bragg scattering equation**. (William Lawrence Bragg and his son William Henry Bragg were jointly awarded the Nobel prize for their work on X-ray diffraction in 1915.) The equation has been derived by considering only the reflections off adjacent layers, but if there is a whole number of wavelengths difference between adjacent layers, then the same must be true for *all* layers. The very large number of layers involved means that at any angle other than that defined by the Bragg scattering equation, the overall effect will be destructive interference. This reasoning is the same as that used to explain why the large number of lines on a diffraction grating results in intense interference at very precise angles.

X-ray diffraction may be demonstrated in a laboratory analogy using 3 cm microwaves and an array of 3 cm polystyrene spheres stuck together to form a three-dimensional pattern.

G.5.7 Outline how cubic crystals may be used to measure the wavelength of X-rays.

Using X-ray diffraction

In the same way that a diffraction grating can be used to determine an unknown wavelength of light, a cubic crystal (one in which the characteristic pattern of particles is in the shape of a cube) can be used to determine an unknown X-ray wavelength.

Worked example

8 X-rays of unknown wavelength were directed onto a cubic crystal with parallel lattice planes separated by 1.29×10^{-10} m. As the angle between the face of the crystal and the X-ray beam was increased from zero, a maximum of intensity was first detected at an angle of 33.9°. What was the wavelength of the X-rays?

$2d\sin\theta = n\lambda$
$2 \times (1.29 \times 10^{-10}) \times \sin 33.9° = 1 \times \lambda$
$\lambda = 1.44 \times 10^{-10}$ m

9 When X-rays of wavelength 4.7×10^{-11} m were used to obtain a diffraction pattern with a crystalline material, the first-order maximum was formed at an angle of 12°.
a What was the spacing of the ions in the lattice?
b At what angle was the second-order maximum?
c Another set of planes within the same crystal had a spacing of 1.3×10^{-10} m. At what angle was their first-order maximum detected?

a $2d\sin\theta = n\lambda$
$2 \times d\sin 12° = 1 \times (4.7 \times 10^{-11})$
$d = 1.1 \times 10^{-10}$ m

b $2d\sin\theta = n\lambda$
$2 \times (1.1 \times 10^{-10}) \times \sin\theta = 2 \times (4.7 \times 10^{-11})$
$\theta = 25°$

c $2d\sin\theta = n\lambda$
$2 \times (1.3 \times 10^{-10}) \times \sin\theta = 1 \times (4.7 \times 10^{-11})$
$\theta = 10°$

G.5.8 **Outline** how X-rays may be used to determine the structure of crystals.

X-ray crystallography

By measuring all the angles at which X-rays of a known wavelength are reflected, it is possible to determine the spacing of all the planes within a crystal. With this kind of information it is then possible to construct detailed models of crystalline structures. Such methods may be limited to crystalline materials, but most substances have some regularity in their structure. Famously, the double helix structure of the DNA molecule, which was proposed by Watson and Crick in 1953, was discovered following considerable work by many other scientists, including the production of X-ray diffraction images by Rosalind Franklin.

Figure 19.62 shows a simplified representation of a basic apparatus arrangement that can be used. Alternatively, the apparatus may be surrounded by photographic film to record the results. The two slits are used to make sure that the beam is travelling in one precise direction (collimated). The crystal is rotated and when the angle is just right (as represented by the Bragg scattering equation), a strong reflected beam can be detected. Remember that the reflections will occur in three dimensions, not just in the plane of the drawing. Figure 19.63 shows a photograph of typical result from a single simple crystal, with many intensity peaks at precise angles. This view is looking towards the crystal, which was located at the centre of the pattern. Powdered samples of crystalline materials are also commonly used. These produce diffraction *rings* because the sample will contain all possible orientations of a very large number of tiny pieces of the material.

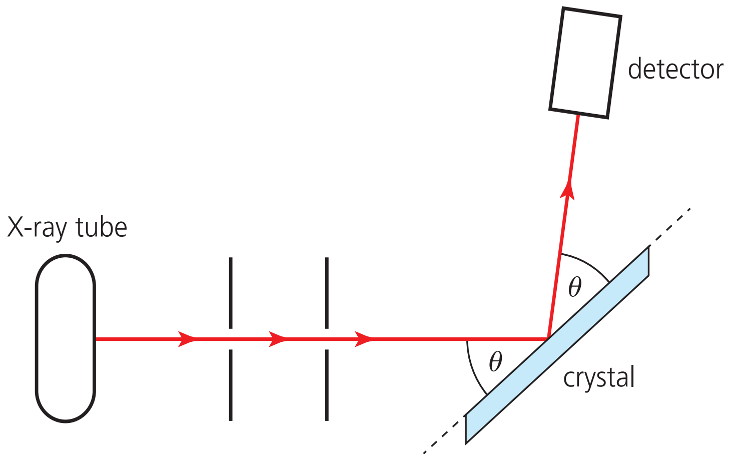


Figure 19.62 Basic practical arrangement for investigating a single crystal

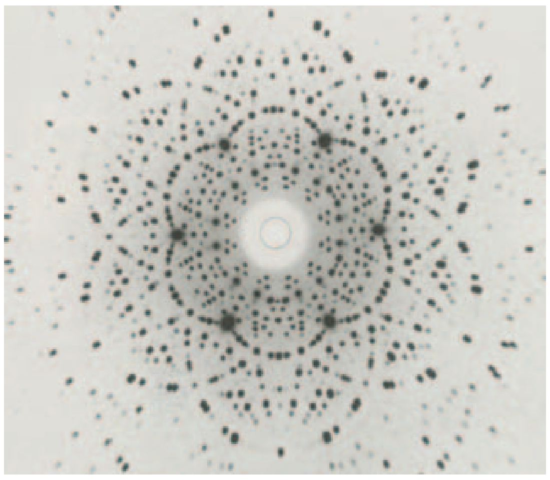


Figure 19.63 X-ray diffraction pattern produced by a single crystal

G.5.9 **Solve** problems involving the Bragg equation.

66 The spacing between certain atomic planes of ions in copper(I) oxide is 3.08×10^{-10} m. Calculate the angles at which there will be reflected maxima when using X-rays of wavelength 1.47×10^{-10} m.

67 Determine the wavelength of X-rays which will produce a second-order interference maximum at an angle of 23.8° when using a crystal with atomic planes separated by 7.68×10^{-11} m.

68 Figure 19.64 shows a simplified two-dimensional representation of a crystal lattice. The planes of atoms shown in red have a separation of d.

a Calculate the spacing of the planes shown in black (in terms of d).

b If the planes shown in red produced a first-order maximum at an angle of 11.5°, at what angle would the first order be for the planes shown in black (with the same X-rays)?

c Make a copy of the lattice and indicate another set of atomic planes with a different spacing.

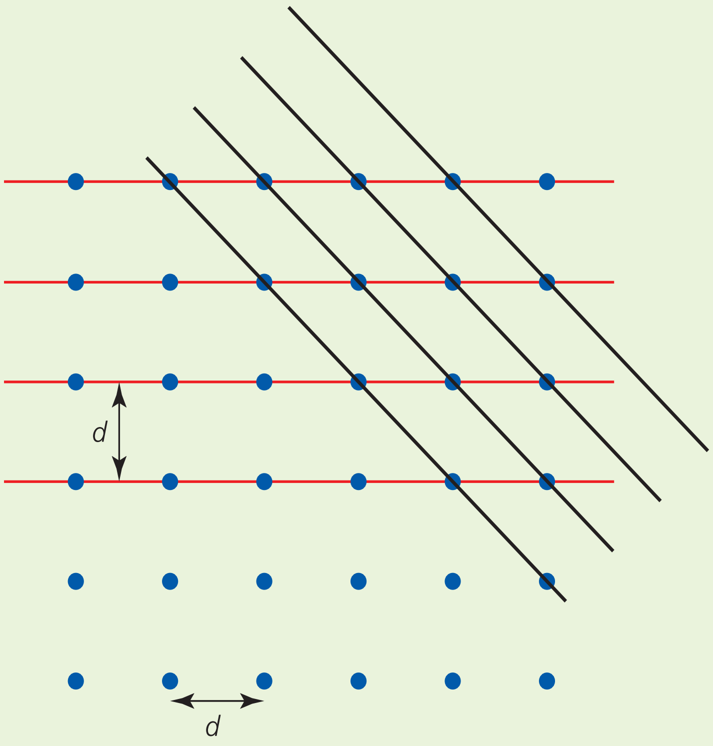


Figure 19.64

69 The smallest X-ray diffraction ring detected photographically when using a crystalline powder had a diameter of 6.4 cm. If the wavelength used was 3.7×10^{-11} m and the distance from the sample to the centre of the pattern on photographic paper was 23 cm, what atomic plane spacing in the sample was responsible for this ring?

70 The diffraction of electron waves was discussed in Chapter 13. Suggest the main differences between using an X-ray beam and an electron beam to investigate crystal structure.

71 Outline how microwaves could be used with a model of a crystalline structure to demonstrate the diffraction of X-rays.

G6 Thin-film interference

We know that coherent light waves are needed to produce any interference pattern, and another way of providing these is by the use of **thin films**. (The word 'film' here means a thin layer.) Figure 19.65 shows light rays being reflected off the top and bottom surfaces of a thin film of a transparent material which, for example, might be water or oil. The dotted lines show other possible directions in which the light could travel.

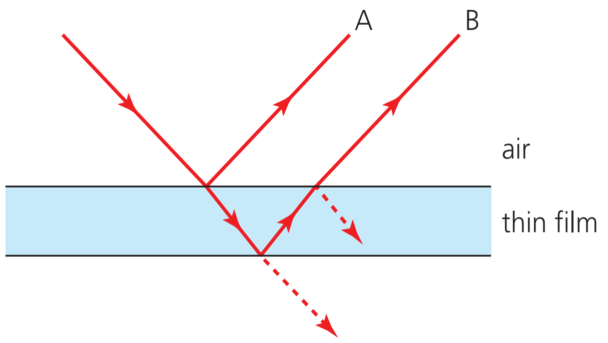


Figure 19.65 Waves reflected off two surfaces of a thin film

If the film is thin enough, the rays A and B will be coherent (because they came from the same source) and they will produce interference effects, depending on their path difference. If the film is not very thin the waves will lose their coherence and no interference pattern will be seen.

It is recommended that students make use of one of the many computer simulations available for thin-film interference.

Wedge films

G.6.1 Explain the production of interference fringes by a thin air wedge.

Interference can be observed using an air **wedge** that is formed between glass plates with a small layer of air between them. Figure 19.66 shows how this can be done with a piece of paper and two microscope slides. The space between the slides forms an air wedge of increasing thickness.

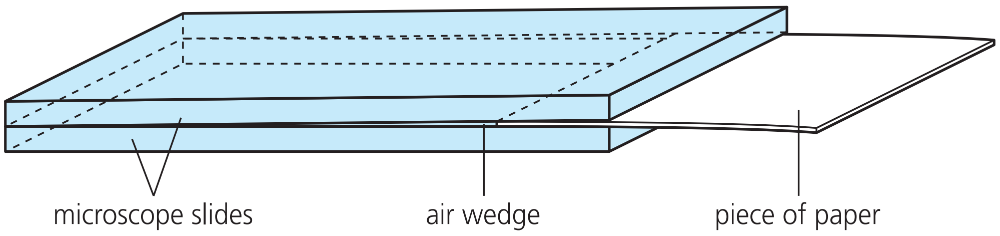


Figure 19.66 Making an air wedge with two glass slides

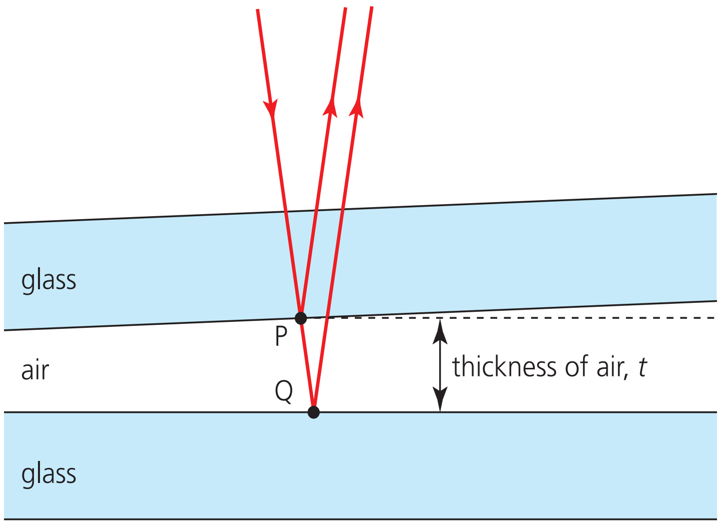


Figure 19.67 Monochromatic light incident normally on an air wedge

Figure 19.67 shows a monochromatic light ray striking an air wedge from above. The size of the angle between the slides has been exaggerated in the diagram. In fact, the angle is so small that we can consider that the rays are incident on all surfaces perpendicularly. This means that we can assume that there is no refraction in the glass. It is important to realize that the thickness of the air wedge compared to the glass has also been greatly exaggerated for the sake of clarity in the diagram.

Rays reflected from the top and bottom surfaces of the air wedge will be coherent and will interfere. The interference will be constructive if the path difference is equal to $m\lambda$. (The symbol m is used here to represent an integer so as to avoid confusion with n, which is used for refractive index.) Reflections off the other two surfaces can be ignored because their separation is relatively large.

Path differences and phase changes

G.6.5 State the condition for light to undergo either a phase change of π, or no phase change, on reflection from an interface.

The path difference is *not* simply twice the thickness, *t*, of the wedge at that point, because there is a *phase change* of π radians at Q, when the light travelling in air reflects off a medium (glass) with a higher refractive index. There is *no* phase change when the light reflects at P because the light is reflecting off a medium (air) with a lower refractive index. This was discussed briefly in Chapter 4.

Equations for interference at wedge films

Generally, we know that the condition for constructive interference is that the path difference equals a whole number of wavelengths, but the phase change introduces a path difference of π, which is equivalent to half a wavelength, $\lambda/2$. This means that the condition for *constructive* interference in air becomes:

$$\text{path difference} = 2t = (m + \tfrac{1}{2})\lambda$$

However, it is possible that the medium in the wedge is not air, in which case the light will have a smaller wavelength in that medium. For example, water has a refractive index of 1.33, which means that the speed of light in water is $c/1.33$ and the wavelength is reduced to $\lambda/1.33$. Therefore, the wavelength used in the equation above (for air) must be divided by the refractive index and the conditions become:

$2nt = (m + \frac{1}{2})\lambda$	for constructive interference
$2nt = m\lambda$	for destructive interference

These equations are given in the IB *Physics data booklet*.

As we move along the wedge, the path difference and the conditions for interference continuously change. With a wedge of **equal inclination** (constant angle) the result will be a series of fringes of **equal thickness**, as shown in Figure 19.68. Because of the phase change caused by reflection, *destructive* interference will occur where the thickness is zero, where the plates join.

Figure 19.68 Equal thickness fringes for a wedge of equal inclination

Measuring small separations using a wedge film

G.6.2 Explain how wedge fringes can be used to measure very small separations.

Using monochromatic light of a known wavelength, *air* wedges can be used to measure small separations, such as that caused by the thickness of the paper shown in Figure 19.66. The same situation is shown again in more detail in Figure 19.69. The fringes are usually observed using a microscope.

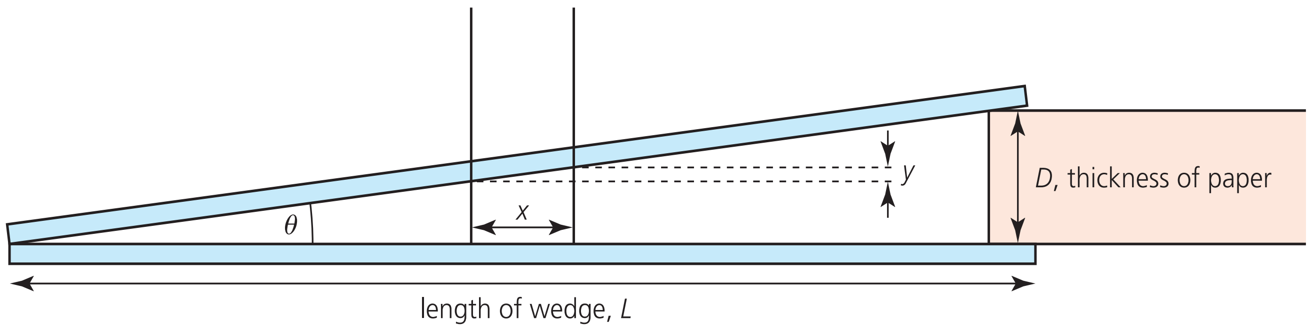


Figure 19.69 Geometry of an air wedge

To calculate the thickness, D, we cannot use the above equations directly. We need to first determine the small change in vertical separation between the plates, y, needed to introduce a path difference of one wavelength, as we move from one bright fringe to the next.

Looking at Figure 19.69 and using trigonometry, we know that:

$$\tan\theta = \frac{D}{L} = \frac{y}{x}$$

But $y = \lambda/2$ because the light travels in both directions, so that a change in path difference of one wavelength is a result of an increase in separation of half a wavelength. Rearranging, we get an equation for the thickness, D:

$$D = \frac{L\lambda}{2x}$$

Worked example

10 a A thin straight wire is placed between the ends of microscope slides to form an air wedge of length 5.89 cm. When illuminated by light of wavelength 6.33×10^{-7} m, the distance between the centres of the first and the eleventh bright fringes was measured to be a horizontal distance of 1.27 cm. What was the diameter of the wire?

b What would be observed if the monochromatic light was replaced with white light?

a $D = \dfrac{L\lambda}{2x}$

$$D = \frac{(5.89 \times 10^{-2}) \times (6.33 \times 10^{-7})}{2 \times (1.27 \times 10^{-2} / 10)}$$

$D = 1.47 \times 10^{-5}$ m

b The fringes would be blurred and edged with colour.

Sometimes the wedges which produce interference patterns are formed by liquids such as oil and water. For example, tears in the eye form liquid wedges above the lower eyelid and measurement of the thickness and shape of the tear film can be of help in diagnosing problems with the eye. In any calculation involving a liquid wedge, the wavelength used must be the wavelength in the liquid, not the wavelength in air.

Optical flatness

G.6.3 Describe how thin-film interference is used to test optical flats.

We have seen that a thin wedge film of equal inclination produces interference fringes of equal thickness. A thin film that varies in thickness but *not* by equal inclination will produce fringes which are not of equal thickness and probably not straight. This can be used to test whether a surface is flat.

The surface of the transparent material to be tested is placed on top of a known **optical flat**. An optical flat is a piece of glass which has been ground and polished so that it is extremely flat, varying by much less than the wavelength of light. If the surface being investigated is not flat, then an air film (wedge) of variable thickness will be formed. A typical result is shown in Figure 19.70. The fringes are effectively lines joining places where the air gap is of equal thickness.

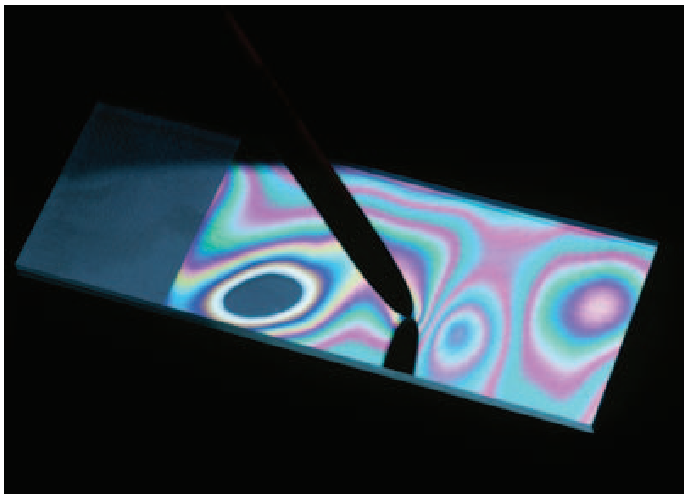


Figure 19.70 Fringes formed by an air film of varying thickness

G.6.4 Solve problems involving wedge films.

72 An air wedge is made using two glass slides joined at one edge, but separated by a piece of paper of thickness 3.7×10^{-3} cm at the other end. When illuminated by monochromatic light, nine equally spaced bright fringes are seen to be separated by a distance of 4.2 mm.

a If the length of the wedge is 6.1 cm, what was the wavelength of the light?

b If water was placed in the space between the slides, what would be the new spacing of the fringes? (Refractive index of water = 1.3.)

73 What thicknesses of air wedge would produce constructive interference for light of wavelength 4.8×10^{-7} m?

74 When observing the interference fringes produced by a thin air wedge with light of wavelength 630 nm, the separation of four bright fringes was found to be 1.8 mm.

a What was the angle of the wedge?

b How would the fringe pattern change if:

i light of a longer wavelength was used

ii the angle was decreased?

75 A plano-convex lens is placed on an optically flat surface and viewed with monochromatic light, as shown in Figure 19.71.

a Make a sketch of the interference pattern observed if the lens has a spherical surface.

b Show in a separate sketch how the observed pattern would be different if the spherical surface of the lens was of poor quality.

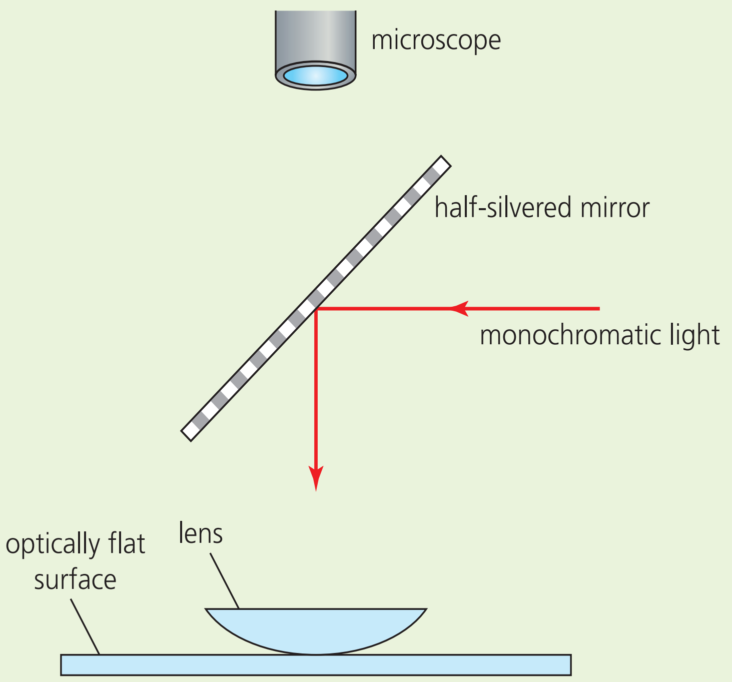


Figure 19.71

Parallel films

Normal incidence

G.6.6 Describe how a source of light gives rise to an interference pattern when the light is reflected at both surfaces of a parallel film.
G.6.7 State the conditions for constructive and destructive interference.
G.6.9 Describe the difference between fringes formed by a parallel film and a wedge film.

In the same way that interference can occur between both surfaces of a wedge, interference can occur between both the parallel surfaces of a thin film of a transparent material, like oil or water (as shown in Figure 19.65).

The conditions for constructive and destructive interference are the same and the equations used in the previous section for wedge films [$2nt = (m + \frac{1}{2})\lambda$ and $2nt = m\lambda$] can also be used for normal incidence on parallel films. However, the film needs to be very thin – interference effects will *not* be seen with, for example, a glass slide because it is too thick.

For normal incidence and observation using monochromatic light reflecting off a parallel film, the path difference will always be the same for any position of the eye. Interference will occur, but it will not change with position and no pattern will be seen. However, the path difference needed for constructive interference depends on the wavelength and, if the film is observed using white light, the thickness of any particular film will be more favourable for constructive (or destructive) interference of some wavelengths than others. This will result in the film appearing one particular colour. If the film is very thin, the physical path difference will be much less than a wavelength, so that phase difference of all light waves will be only that due to the reflection ($\lambda/2$) and the result will be destructive interference for all wavelengths.

Oblique incidence

G.6.8 Explain the formation of coloured fringes when white light is reflected from thin films, such as oil and soap films.

We will now broaden our discussion to include the possibility of observing thin-films from different angles. Figure 19.72 (overleaf) could represent, for example, reflections off an oil film floating on water.

If the thin film is of constant thickness, the pairs of reflections from each of the three incident rays in the diagram will *not* have the same path differences and this means that

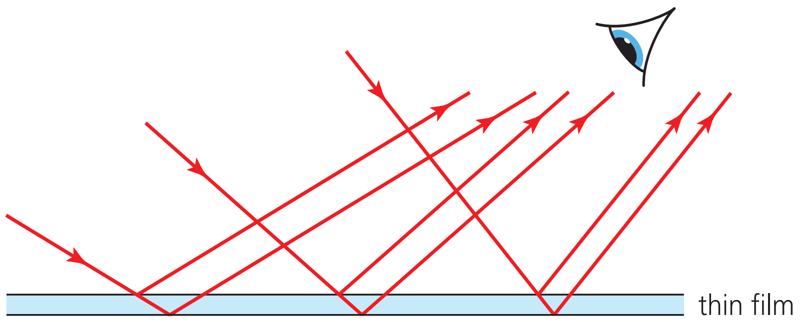


Figure 19.72 Observing a thin film at different angles

constructive interference occurs for different wavelengths at different angles. The eye will see different colours depending on which part of the oil film it is looking at. Figure 19.73 shows the typical appearance of an oil film on water. There is usually no regularity in the patterns observed because the thickness of the oil film is not constant. Similar effects are seen in soap bubbles (Figure 19.74), but their appearances change rapidly as the water moves towards the bottom of the bubble because of gravity and because the water evaporates quickly.

Figure 19.73 Very thin oil film on water

Figure 19.74 Interference effects in soap bubbles

The mathematics of oblique incidence

Figure 19.75 shows in detail how a ray of light striking a parallel film obliquely is divided into two reflected rays; ϕ is the angle of refraction inside the film.

By considering path differences it can be shown that:

Constructive interference will occur if $2nt\cos\phi = (m + \frac{1}{2})\lambda$.

Destructive interference will occur if $2nt\cos\phi = m\lambda$.

For normal incidence, $\cos\phi = 1$, so that these equations become identical to those used earlier.

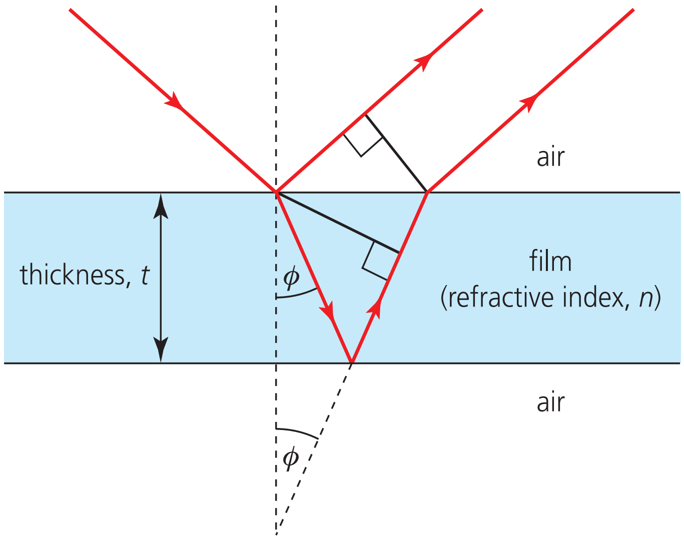


Figure 19.75 Path difference for oblique incidence

Worked example

11 Light of wavelength 589 nm is directed in a narrow beam towards a water film of refractive index 1.33 and thickness 1.37×10^{-7} m. Calculate the minimum (and only) angle of incidence needed in order for constructive interference between reflections off the two water surfaces.

$$2nt \cos\phi = (m + \frac{1}{2})\lambda$$

For the first angle of constructive interference $m = 0$:

$$2 \times 1.33 \times (1.37 \times 10^{-7}) \times \cos\phi = 0.5 \times (589 \times 10^{-9})$$

$\cos\phi = 0.808$ and $\phi = 36.1°$

Since ${}_{\text{air}}n_{\text{water}} = \dfrac{\sin\theta_{\text{air}}}{\sin\theta_{\text{water}}}$ (from Chapter 4):

$1.33 = \sin\theta_{\text{air}}/\sin 36.1°$

$\sin\theta_{\text{air}} = 0.784$, so that the angle of incidence in air is 51.6°.

Uses of parallel thin films

G.6.10 Describe applications of parallel thin films.

1 **Measurement of the thickness of an oil film.**
The equation $2nt = (m + \frac{1}{2})\lambda$ can be used for constructive interference to calculate the thickness, t, of a thin oil film if its refractive index, n, is known. This is illustrated in question **76**.

2 **Non-reflective surfaces.**
Consider a glass surface with a thin film of transparent material ('coated') on top of it, as shown in Figure 19.76. The material of the film has a refractive index *lower* than glass.

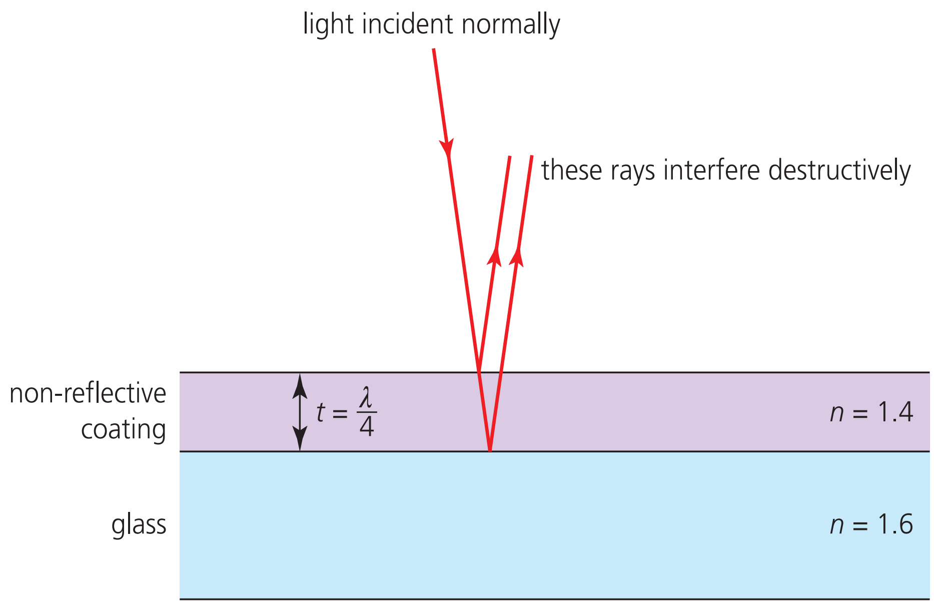


Figure 19.76 Reflections from a non-reflective coating

The incident light is monochromatic, with a wavelength λ and it is incident normally, although in the diagram it has been shown at a slight angle in order for the separate rays to be distinguished.

Because *both* reflections occur when the light is entering a medium with a greater refractive index, we do not need to consider the effect of the phase changes involved. So that the minimum condition for destructive interference becomes:

$$2t = \frac{\lambda}{2n} \quad \text{or} \quad t = \frac{\lambda}{4n}$$

where n is the refractive index of the coating.

A non-reflective coating needs to have a thickness of one quarter of a wavelength (measured in the coating, not air). If light is not reflected, then more useful light energy will be transmitted through the glass. Coating lenses is a way of ensuring the maximum possible light intensity is transmitted through all kinds of optical instruments (including eye glasses), as well as solar cells and solar panels. The process of putting a non-reflective coating on lenses is called **blooming**.

Of course, the thickness of a single layer will only result in perfectly destructive interference for one particular wavelength, but other wavelengths will also undergo destructive interference to some extent. By using multiple coatings of different thickness or refractive index, it is possible to effectively reduce reflections over all visible wavelengths. These effects also depend on the angle of incidence, but most of the time lenses are used with light which is incident approximately normally.

Worked example

12 Calculate the thickness needed for a lens coating designed to eliminate the reflection of green light of frequency 5.77×10^{14} Hz. The material of the coating has a refractive index of 1.46.

We need to calculate the wavelength of the green light (in air):

$$\lambda = \frac{c}{f}$$

$$\lambda = \frac{3.00 \times 10^{8}}{5.77 \times 10^{14}} = 5.20 \times 10^{-7}\,\text{m}$$

then, $t = \dfrac{\lambda}{4n}$

$$t = \frac{5.20 \times 10^{-7}}{4 \times 1.46} = 8.90 \times 10^{-8}\,\text{m thick}$$

3 **Military aircraft.**
Non-reflecting coatings are also used on military aircraft to help them avoid detection by radar. Radar uses a range of possible wavelengths, typically a few centimetres. Figure 19.77 shows a Stealth aircraft which uses a wide variety of techniques to avoid detection.

Figure 19.77 Stealth aircraft

G.6.11 Solve problems involving parallel films.

76 An oil film is observed using normal incidence of light of wavelength 624 nm. If the oil has a refractive index of 1.42, what is the minimum thickness that will produce constructive interference?

77 A transparent plastic film of refractive index 1.51 is observed using monochromatic visible light at normal incidence. If destructive interference occurs when the film is 4.58×10^{-7} m thick, calculate possible values for the wavelength.

78 The blooming on a lens is 8.3×10^{-5} mm thick. What is its approximate refractive index if it is designed to be non-reflective for blue light?

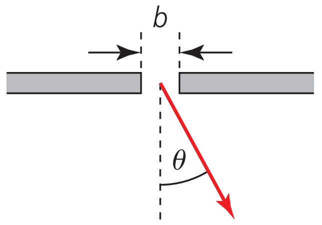
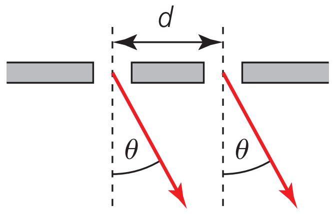
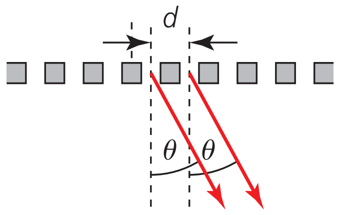
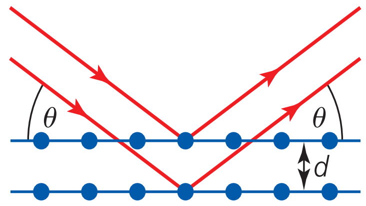
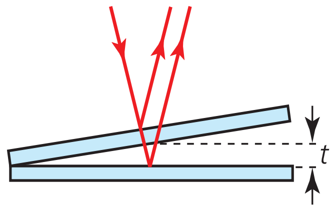
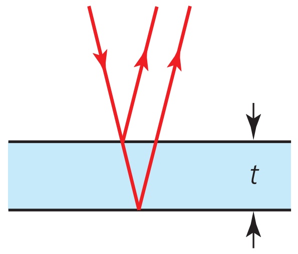
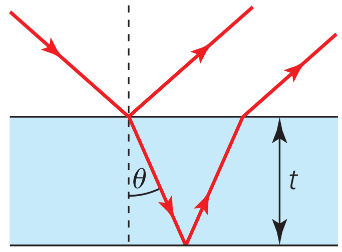
79 Find out the various means by which a military aircraft might avoid detection.

80 By considering the path difference of the rays shown in Figure 19.76 derive the equation for constructive interference: $2nt\cos\phi = (m + \frac{1}{2})\lambda$.

81 White light is incident at an angle of 33.5° to the normal as it passes into an oil film of thickness 1.28×10^{-7} m. The oil has a refractive index of 1.44 (assumed to be the same for all wavelengths).
a What is the angle of refraction in the oil?
b What wavelength of reflected light will interfere constructively for this angle of incidence?
c What colour is that wavelength?

Summary of interference equations

Table 19.2 Interference equations used in this course

Arrangement	Meaning of symbols	Condition for constructive interference	Condition for destructive interference	Chapter
Single slit			For the first minimum, $\theta = \frac{\lambda}{b}$ ($\approx \sin\theta$ for small angles)	Topic 11 (Higher Level only)
Double slits	 n = integer, s = separation of adjacent fringes, x = separation of n fringes, D = slit to screen distance	$\sin\theta = \frac{n\lambda}{d}$ For small angles and adjacent bright fringes on a screen (or adjacent dark fringes): $s = \frac{\lambda D}{d}$ For small angles and measurements across a number of similar fringes on a screen: $\frac{x}{D} = \frac{n\lambda}{d}$	$\sin\theta = \frac{(n + \frac{1}{2})\lambda}{d}$	Option G
Diffraction gratings		$d\sin\theta = n\lambda$		
X-ray diffraction		$2d\sin\theta = n\lambda$		Option G (Higher Level only)
Wedge for normal incidence	 n = refractive index, m = integer (The rays in this diagram are not drawn perpendicularly for the sake of clarity)	$2nt = (m + \frac{1}{2})\lambda$	$2nt = m\lambda$	
Parallel film for normal incidence	 (The rays in this diagram are not drawn perpendicularly for the sake of clarity)	$2nt = (m + \frac{1}{2})\lambda$	$2nt = m\lambda$	
Parallel film for oblique incidence	 n = refractive index, m = integer	$2nt\cos\phi = (m + \frac{1}{2})\lambda$	$2nt\cos\phi = m\lambda$	

Additional Perspectives

Oil spills

The energy-intense lifestyles of modern life demand a continuous supply of crude oil (see Chapter 8). The extraction of crude oil from the under the ocean floor, its transfer to land and its movement around the planet in large oil tankers and through pipes has caused a number of serious accidents. These accidents cause considerable pollution to the seas, and the plants and animals that live in and around them. Figure 19.78 shows a penguin covered with oil from a spill off the coast of Taurana in New Zealand in 2011.

Figure 19.78 Effects of an oil spill on marine life

Oil is less dense than water and therefore it floats on the surface of the sea. An oil spill can be difficult to contain because the oil tends to spread out very thinly on the surface of water (to an extent which varies with the type of oil). The phrase *oil slick* is sometimes used for a smooth layer of oil which has spread over a large area and which has not been broken up by the action of the waves.

An oil spill of about 1000 litres which has spread to cover a square kilometre would be about 10^{-3} mm thick and be noticeable by the coloured interference patterns it produced. An oil film about one tenth of this thickness will have a shiny appearance. A thicker spill will appear much darker, without colours. It is often possible to estimate the volume of an oil spill from its area and appearance.

Obviously the severity of the possible pollution following an oil spill depends on the quantity of oil released into the environment and its location, but there are many other factors involved, including the type of oil and the prevailing weather, currents and waves. The temperature of the water may also be an important factor.

Question

1 Research into the various methods that can be used to contain and clean up a large oil spill on the surface of the sea.

SUMMARY OF KNOWLEDGE

G1 The nature of EM waves and light sources

- Electromagnetic waves are produced whenever a charge is accelerated. EM waves are transverse. They are combined electric and magnetic fields which travel at a speed of $3.00 \times 10^8\,\text{m s}^{-1}$ across free space.
- There are different regions of the electromagnetic spectrum, with a source for each region. An order of magnitude for their frequencies and wavelengths should be known.
- When EM waves pass from air into a different medium their speed decreases. The refractive index of a medium is the ratio of the speed of EM waves in air (or a vacuum) divided by the speed in the medium. Different wavelengths have slightly different speeds in the same medium, therefore the refractive index of a material depends on wavelength.
- Because of this, EM waves may be dispersed (travel in different directions) after they enter a different medium. The dispersion of white light by a triangular glass prism is a common example.
- When EM waves pass through a medium they may be transmitted, absorbed or scattered. Scattering is generally a wavelength-dependent effect. Absorption effects and their possible health hazards depend on the energy carried by individual photons.
- EM waves are absorbed and scattered by the Earth's atmosphere. Examples include the ozone layer, the colour of the sky (including at sunrise and sunset) and the effect of greenhouse gases.
- The light emitted from a laser is monochromatic and coherent. The beam is also intense and very directional because it has a very low divergence.
- Monochromatic means that the light contains only a single wavelength. Coherent means that the waves have a constant phase difference.

- Electrons in atoms are usually in their ground state, but in a material used in a laser, it is possible for many of the atoms to stay in a higher (metastable) energy level for a longer than usual time. This is called a population inversion.
- An atom in a metastable energy level can be stimulated to emit a photon when it interacts with another photon already emitted by the same transition in another atom. When the light is reflected backwards and forwards between parallel mirrors the number of coherent photons emitted in the same direction is increased.
- Lasers are used in the following applications: medicine, digital communications, bar-code scanners, optical data storage, surveying and cutting.

G2 Optical instruments

- When light rays spreading out from a point object are incident upon a lens which is thicker in the middle than at its edges, the rays will be refracted and converged to form an image at the point where the rays cross (unless the object is at the focal point, or nearer to the lens). This kind of lens is called a converging (convex) lens.
- The principal axis of a lens is defined as the imaginary straight line passing through the centre of the lens which is perpendicular to its surfaces.
- The focal point of a convex lens is defined as the point through which all rays parallel to the principal axis converge after passing through the lens.
- The focal length of a lens is defined as the distance between the centre of the lens and the focal point. Its value depends on the refractive index and the curvature of its surfaces.
- The optical power of a lens is defined as 1/focal length. $P = 1/f$. Optical power is measured in dioptres, D. Power (D) = 1/focal length in metres.
- The paths of three rays from the top of any extended object, which pass through the lens and then go to the top of the image, can be predicted. Using these rays, diagrams can be drawn to determine the position and nature of the image formed when objects are placed at various distances from the lens.
- In diagrams and calculations throughout this topic we assume that the lens is thin and that the rays are close to the principal axis. If this is not true, the image will not be formed exactly where predicted.
- If part of a lens is covered, an image will still be formed in the same place.
- Real images are formed where rays actually cross. Virtual images are formed when rays diverge into the eye: the image is formed where the rays appear to have come from.
- The linear magnification of an image is the ratio of the height of the image, h_i, divided by the height of the object, h_o. $m = h_i/h_o = -v/u$.
- If the object is placed further away from the lens than the focal point, the image formed will always be real and inverted.
- The thin lens formula is: $1/f = 1/u + 1/v$. This formula can be used to determine the position and nature of an image. When using this formula it is important to remember that a distance to a virtual image is always negative. A positive magnification indicates that the image is upright; a negative magnification indicates that the image is inverted ($m = -v/u$).
- If the object is placed at the focal point or closer to a convex lens, the image will be magnified, upright and virtual. Used in this way the lens is a simple magnifying glass.
- The nearest point to the human eye at which an object can be clearly focused (without straining) is called the near point. It is accepted to be 25 cm from a normal eye and given the symbol D. The furthest point from the human eye that an object can be clearly focused (without straining) is called the far point, for a normal eye it is at infinity.
- The angular magnification of an image is the angle subtended at the eye by the image divided by the angle subtended at the eye by the object.
- The angular magnification, M, of a simple magnifying glass varies between D/f, for the image at infinity, to $D/f + 1$ for the image at the near point.
- The objective lens of a compound microscope forms a real magnified image of an object which is placed just beyond its focal point. The eyepiece then acts as a magnifying glass to produce a final image which is inverted, magnified and virtual.

- Ray diagrams can be constructed to represent a microscope in normal adjustment with the final image at (or near to) the near point.
- The objective lens of a telescope forms a diminished, real and inverted image of a distant object at its focal point. The eyepiece acts as a magnifying glass to produce a final image at infinity (in normal adjustment) which is inverted, diminished and virtual. The linear magnification is less than one, but the telescope produces an angular magnification, $M = f_o/f_e$.
- Ray diagrams can be constructed to represent a telescope in normal adjustment. The distance between the lenses in normal adjustment is $f_o + f_e$.
- Lens aberrations (especially with higher power lenses) are the principal limitations on the magnification achievable by optical instruments that use lenses.
- Spherical aberration produces distorted images. It is the inability of a lens with spherical surfaces to bring all rays incident upon it (from a point object) to the same focus. It may be reduced by adapting the shape of the lens, or by only using the centre of the lens.
- Chromatic aberration is the inability of a lens to bring rays of different colours (from a point object) to the same focus. It occurs because refractive index varies slightly with colour (wavelength). It can be reduced by combining lenses of different shapes and refractive indices.
- Diagrams can be drawn to represent these aberrations and how they can be reduced.

G3 Two-source interference of waves

- In order to observe an interference pattern it is necessary that the waves which are combining are coherent and approximately the same amplitude. This can be achieved by taking the light from a single source and splitting it into two by passing it through narrow slits that are very close together. Coherent waves passing through slits diffract and then cross over each other.
- The effect at any point can be determined using the principle of superposition. Constructive interference will occur when the path difference between rays from the two slits is a whole number of wavelengths. $\sin\theta = n\lambda/d$. The pattern of fringes is commonly shown on a graph representing the relative intensities at different angles.
- The measurements taken from a double-slit experiment can be used to determine the wavelength(s) of the light used.

G4 Diffraction grating

- As light passes through double slits it produces an interference pattern of maxima and minima which are blurred and of low intensity. If the number of slits is increased (keeping the same spacing), the pattern becomes sharper and more intense.
- A diffraction grating has a very large number of slits (lines) very close together. The diffraction grating formula ($n\lambda = d\sin\theta$) which predicts the angles of constructive interference can be derived and the corresponding intensity distribution drawn.
- Diffraction gratings are used to disperse light into its different wavelengths and certain measurements can be taken in order to determine an unknown wavelength.

G5 X-rays (Higher Level only)

- When electrons are accelerated by a voltage V, the kinetic energy they gain, $E_K = \frac{1}{2}mv^2 = eV$.
- X-rays are produced in an 'X-ray tube' when electrons, which were accelerated by a very large voltage, are then decelerated as they strike a metal target. (Most of the energy of the electrons is transferred to internal energy in the target, which may need to be cooled.)
- Raising the voltage across the cathode will make it hotter and so release more electrons by thermionic emission. This will increase the intensity of the X-ray beam. Raising the accelerating voltage between the anode and the cathode will increase the energy of the individual X-ray photons and decrease their wavelength. They are then described as being 'harder'. (Photon energy $= hf = hc/\lambda$.)
- The maximum photon energy (minimum wavelength) emitted will occur when *all* of an electron's kinetic energy is transferred to a single photon. $\lambda_{min} = hc/eV$.

- The X-ray photons emitted have a range of different energies depending on the nature of their interactions with the metal atoms.
- Some of the energy of a bombarding electron may be transferred to raise an electron in a metal atom to a higher energy level. When it returns to a lower level it will emit an X-ray photon which is characteristic of that particular metal.
- Aspects of typical x-ray spectra can be described as continuous or characteristic.
- X-rays are diffracted well by atoms, ions and molecules in crystals because their wavelength is similar to the size and separation of the particles.
- X-rays which are diffracted and scattered from the regular arrangement of atoms in a crystalline material will interfere constructively at precise angles. This occurs when the path difference of X-rays scattered from adjacent atomic planes equals a whole number of wavelengths.
- The Bragg scattering equation predicts the angles at which constructive interference occurs: $2d\sin\theta = n\lambda$.
- If the angles at which X-rays are scattered from a cubic crystal (with atomic planes of known separation) are measured, then the wavelength of the X-rays can be determined using the Bragg scattering equation.
- X-ray diffraction is widely used to investigate the structure of crystalline materials (for example, the double helix of DNA). A sample of the material is bombarded by X-rays of known wavelength and the angles of the interference maxima are measured. The Bragg scattering equation can then be used to determine the separation of the various lattice planes.

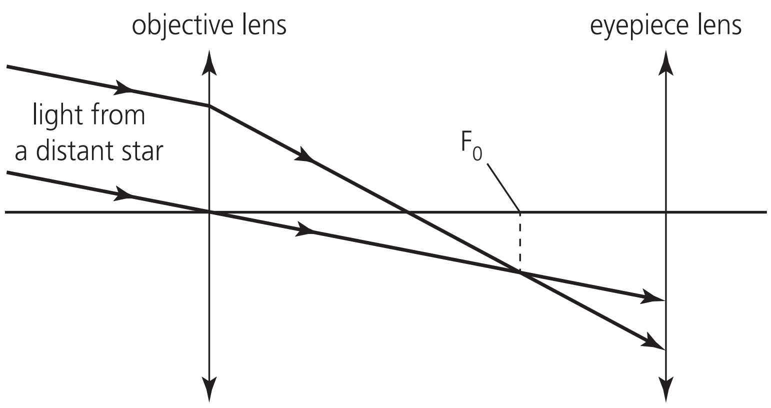
G6 Thin-film interference (Higher Level only)

- Interference effects can also be seen between light rays reflected off the top and bottom surfaces of a thin layer (film). Constructive interference occurs when the optical path difference equals a whole number of wavelengths, but there are two other factors to consider:
  - When light is reflected from the boundary with a medium with a greater refractive index it undergoes a phase change of π, which is equivalent to a path difference of $\lambda/2$.
  - If the medium is not air, the wavelength of the light in air must be divded by n (refractive index) in order to determine the wavelength in that medium.
- Taking these factors into account, the condition for constructive interference for light incident normally becomes $2nt = (m + ½)\lambda$, where m is an integer. The condition for destructive interference is $2nt = m\lambda$.
- The interference of monochromatic light off the top and bottom of oil films and tear films in the eye can be used to measure their thicknesses.
- A very small thickness, D, can be measured by using it to form an air wedge of equal inclination and length L. Using monochromatic light, the fringes observed will have equal thickness and the distance, x, between adjacent fringes is measured. $D = L\lambda/2x$.
- If the variations in the flatness of a surface are much less than a wavelength of light, it is described as optically flat.
- An optically flat surface can be used to determine if another surface is also flat. When they are placed together no interference pattern should be visible if both surfaces are optically flat. If a fringe pattern is seen (which will be neither parallel nor equally spaced) then the second surface is not optically flat.
- Thin coatings on lenses, solar panels and solar cells reduce the amount of light reflected. The coating needs to be $\lambda/4$ thick. Non-reflective radar coatings are also used on military aircraft.
- If monochromatic light is incident obliquely on a parallel film, the condition for constructive interference is $2nt\cos\phi = (m + ½)\lambda$, where ϕ is the angle of refraction inside the film. Destructive interference will occur if $2nt\cos\phi = m\lambda$.
- Coloured effects are seen when oil or soap films are observed in white light because the path difference between rays reflected from the two surfaces depends on the angle of incidence, so that constructive interference occurs for different colours at different angles.

Examination questions – a selection

Paper 3 IB questions and IB style questions

Q1 a The diagram shows two rays of light from a distant star incident on the objective of an astronomical telescope. The paths of the rays are also shown after they pass through the objective lens and are incident on the eyepiece lens of the telescope.



The principal focus of the objective lens is F_o. On a copy of the diagram, mark the position of the:

i principal focus of the eyepiece lens (label this F_e) [1]

ii image of the star formed by the objective lens (label this I). [1]

b State where the final image is formed when the telescope is in normal adjustment. [1]

c Complete the diagram in **a** to show the direction in which the final image of the star is formed for the telescope in normal adjustment. [2]

d The eye ring of an astronomical telescope is a device that is placed outside the eyepiece lens of the telescope at the position where the image of the objective lens is formed by the eyepiece lens. The diameter of the eye ring is the same as the diameter of the image of the objective lens. This ensures that all the light passing through the telescope passes through the eye ring.

A particular astronomical telescope has an objective lens of focal length 98.0 cm and an eyepiece lens of focal length 2.00 cm (i.e. f_o = 98.0 cm, f_e = 2.00 cm). Determine the position of the eye ring. [4]

Standard and **Higher** Level Paper 3, Specimen Paper 09

Q2 a Explain:

i why an astronaut on the Moon will observe that the sky is black during the daytime

ii why the sky on Earth appears blue on a cloudless day.

b Outline the effect that the ozone layers in the upper atmosphere have on incoming radiation from the Sun.

Q3 a Two overlapping beams of light from two flashlights (torches) fall on a screen. Explain why no interference pattern is observed. [3]

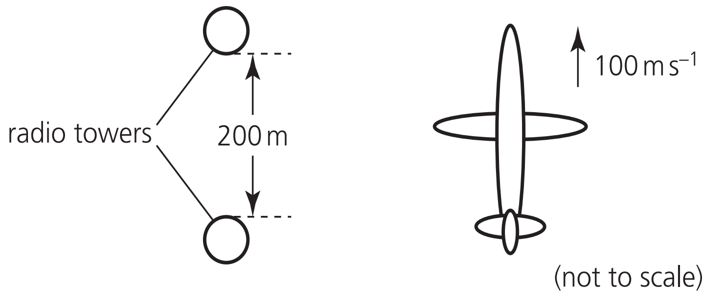
b Light from a laser that passes through a double slit is incident on a screen and produces observable interference.

i Outline how the laser produces light. [2]

ii State the name of the property that enables the laser light to produce observable interference. [1]

c Outline how a laser can be used to read a bar-code. [2]

d A plane is flying at 110 m s^{-1} in a direction parallel to the line joining two identical radio towers, as shown in the diagram.



(not to scale)

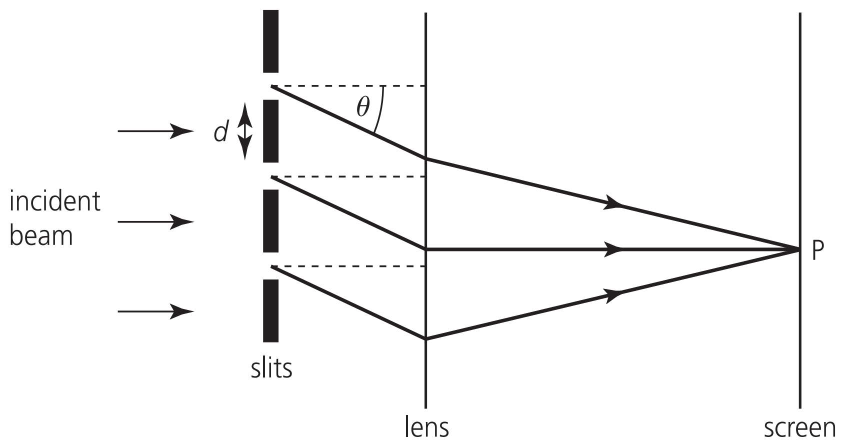
The two towers each emit a coherent radio signal of wavelength of 5.0 m. The separation of the towers is 200 m. To an observer on the plane the intensity of the received signal goes through a maximum every 5.0 s. Determine the distance from the plane to the line joining the radio towers. [3]

Standard and **Higher** Level Paper 3, May 09 TZ2

Q4 a A parallel beam of monochromatic light is incident normally on a diffraction grating. After passing through the grating it is brought to a focus on a screen by a lens. The diagram shows a few of the slits of the diffraction grating and the path of the light that is diffracted at an angle θ to each slit.

The distance between the slits is d and the wavelength of the light is λ.

i On a copy of the diagram, construct a line that enables the path difference between the rays from two adjacent slits to be shown. Label the path distance L. [1]

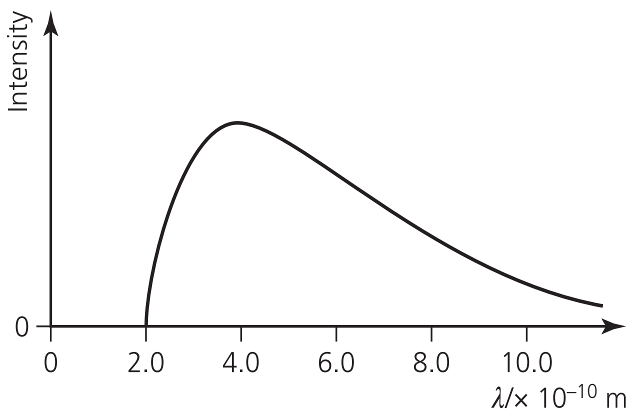


ii Use your answer to **a i** to derive the condition, in terms of d and θ, for there to be a maximum of intensity at the point P on the screen. [2]

b For a particular grating, the distance between adjacent slits is 2.0×10^{-6} m. Determine, for light of wavelength 520 nm, the maximum theoretical order of diffraction. [2]

Standard and **Higher** Level Paper 3, Nov 10

Q5 a Electrons are accelerated from rest by a potential difference. They strike a metal target and the resulting X-ray spectrum is shown in the diagram.

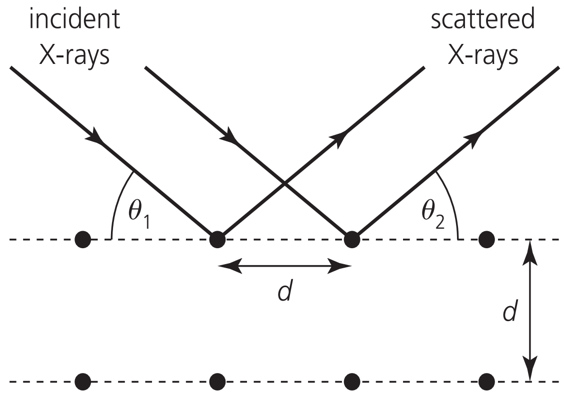


i State and explain what may be deduced about the energy levels of the atoms of the metal from the fact that this spectrum does not contain any characteristic lines. [2]

ii Outline the mechanism by which the photons of wavelength 2.0×10^{-10} m are produced. [2]

iii Calculate the potential difference through which the electrons have been accelerated. [2]

b X-rays are incident on a crystal surface making an angle θ_1 with the surface. The scattered X-rays make an angle θ_2 with the surface. In the diagram the circles, which are separated by a distance d, represent lattice ions of the crystal.



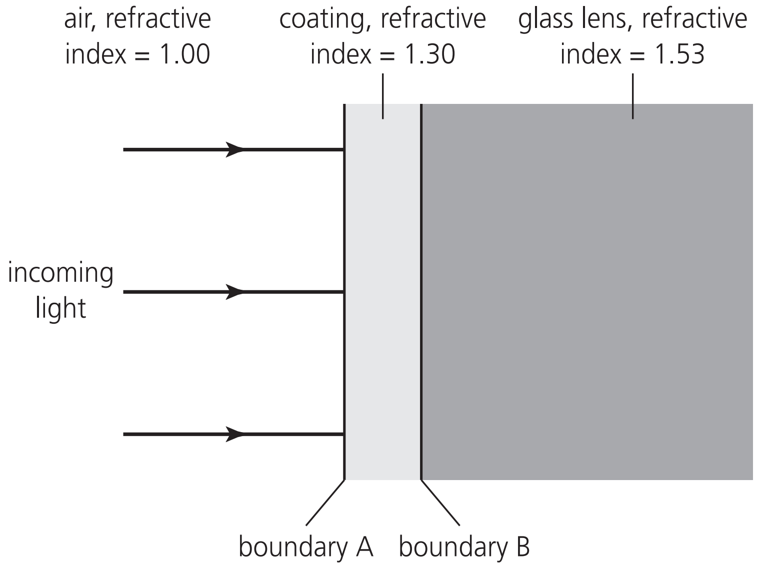
The path difference between the two scattered rays is $d(\cos\theta_1 - \cos\theta_2)$.

i State and explain the condition for constructive interference between the two scattered rays shown. [2]

ii The wavelength of the X-rays is 4.20×10^{-10} m. A maximum in the intensity of the scattered X-rays is first observed at an angle $\theta = 34.5°$. Determine the separation of the atomic planes that give rise to this maximum. [2]

Higher Level Paper 3, May 09 TZ1, QG4

Q6 A transparent thin film is sometimes used to coat spectacle lenses, as shown in the diagram.



a State the phase change which occurs to light that:

i is transmitted at boundary A into the film [1]

ii is reflected at boundary B [1]

iii is transmitted at boundary A from the film into the air. [1]

b Light of wavelength 570 nm in air is incident on the coating. Determine the smallest thickness of the coating required so that the reflection is minimized for normal incidence. [2]

Higher Level Paper 3, Nov 09, QG4

Answers to self-assessment questions in Chapters 15 to 25

Chapter 15 Sight and wave phenomena

1 To reduce the amount of light entering the camera and to increase the range of objects in focus

2 **a** The light passes from water (not air) into the eye, so it does not refract enough to focus the image
b This restores the air in front of the eye, to allow normal focusing

3 More than 10^8

4 Vision is mainly with the rod cells, which are more sensitive to low light levels, but rod cells do not produce coloured images

5 A larger signal can be sent to the brain (for low light levels), but the sharpness of the image is reduced because the location of the signal is not as precise

6 Relative absorptions are approximately 95 (rods), 65 (green cones), 40 (red cones) and 20 (blue cones); the colour is blue–green

7 At night-time vision is mostly by using the rod cells, which are not at the centre of the retina

8 In bright light the object will appear bright, detailed and red because the light is detected using cone cells; under poor lighting the rod cells are used and they cannot respond to red light, so the object cannot be seen, although a dark colourless shape may be apparent

9 **a** Yellow
b Blue
c Red

10 **a** Red
b Yellow and cyan

12 **a** $58.6\,\mathrm{m\,s^{-1}}$
b 71.4 Hz
c 0.492 m

13 **a** **i** 0
ii π
d 21.6 cm; 127 Hz
e $27.5\,\mathrm{m\,s^{-1}}$

14 A stretched string has its own natural frequencies at which it can vibrate freely. If an oscillating force is continually applied to it at one of these frequencies, energy will be transferred into the system and the amplitude will increase

15 **a** The wave speed will increase because of the larger forces in the system
b The fundamental frequency increases because $f = v/\lambda$ and v is higher but λ has not changed
c The oscillating string will accelerate more slowly because it is more massive
d The fundamental frequency decreases because $f = v/\lambda$ and v is smaller but λ has not changed

17 360 Hz

18 $338\,\mathrm{m\,s^{-1}}$

19 0.94 m

20 **a** 1.49 m
b 342 Hz
c 0.745 m
d To produce the same fundamental frequency, they can be half the length of pipes closed at both ends

21 413 Hz

22 $338\,\mathrm{m\,s^{-1}}$

23 $31\,\mathrm{m\,s^{-1}}$

24 **a** The sound will become louder as the train gets closer to P (allowing for a time delay for the sound to reach the observer). A pitch higher than that emitted by the train will be heard, but it will gradually fall as the train approaches (because the component of velocity towards the observer is decreasing). These processes are reversed as the train moves past P
b The pitch and loudness will remain constant

25 59 Hz

26 **a** 3000 Hz
b More easily absorbed and scattered in air; diffract and spread out more; slower speed

27 **a** $260\,\mathrm{m\,s^{-1}}$; 12.6 km
b Make the plane's surface scatter radiation (rather than reflect it); travel close to the ground

29 Moving away with a speed of $8.45 \times 10^6\,\mathrm{m\,s^{-1}}$

30 **a** Ultraviolet
b By fluorescence
c 5.04×10^{-3} rad

31 5.7×10^{-7} m

32 0.085 mm

33 **a** Like Figure 15.40 – minima should occur at angles of $\pm 7.8 \times 10^{-3}$ and $\pm 15.6 \times 10^{-3}$ rad
b The spacing of the diffraction pattern for blue light should be less

34 **a** There is less diffraction with bigger lenses; they receive more light
b It will be more difficult for a larger lens to focus all the light in the right places on the image

35 Blue is near the short wavelength end of the visible spectrum and diffracts less

36 A larger pupil at night means that diffraction is reduced, suggesting that resolution improves; but the much lower light intensity will reduce the quality of the image

37 12 km

38 0.15 m

39 **a** 1.4×10^{14} m
b A line joining the stars is perpendicular to a line joining them to Earth

40 Yes; the angle subtended at the telescope by the writing is 1.6×10^{-5} rad and this is much bigger than $1.22\lambda/b$ (about 5×10^{-7})

41 About 100 km

43 Look through the sunglasses at light reflected from glass or water; if they are Polaroid®, the intensity of the image will change as the glasses are rotated

44 6.7%

45 63°

46 The sky appears blue because blue light is scattered from air molecules; simple scattering, like reflection, can result in polarization

47 **a** 57°
b 53°
c 37°

49 **a** $0.41\,\mathrm{W\,m^{-2}}$
b $0\,\mathrm{W\,m^{-2}}$
c $0.090\,\mathrm{W\,m^{-2}}$

Chapter 16 Quantum physics and nuclear physics

1 1.38×10^{24}

2 2 : 1

3 100

4 **a** 5×10^{-15} J; 3×10^{4} eV

5 **a** 4.91×10^{14} Hz
b 3.26×10^{-19} J
c 3.54×10^{-19} J
d No
f 5.34×10^{14} Hz

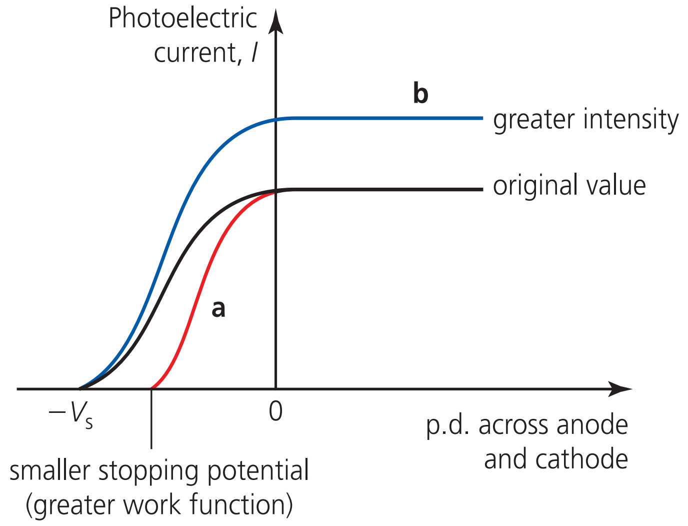
7 2.0×10^{19} J

8 **a** 3.62×10^{-19} J
b 5.5×10^{-7} m; yellow light
c Red

9 1.44×10^{15} Hz

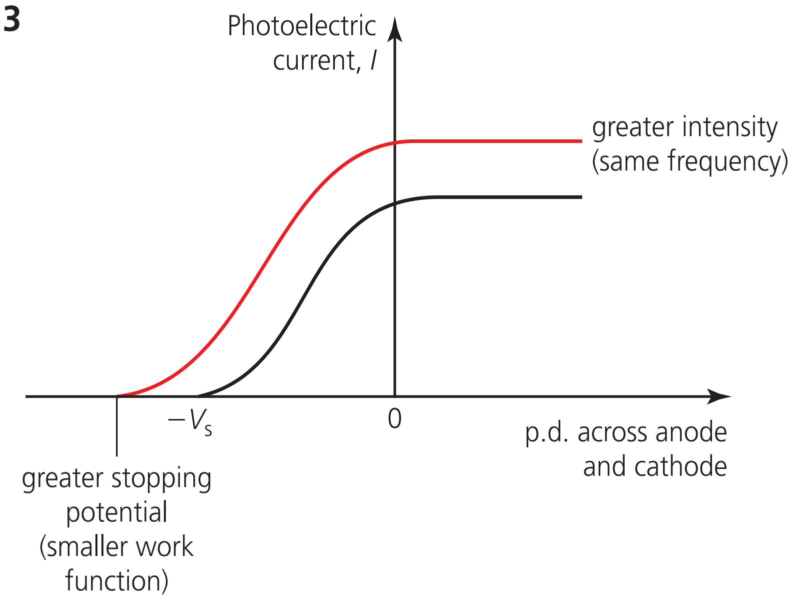
10 3.8×10^{-19} J; 2.4 eV

11

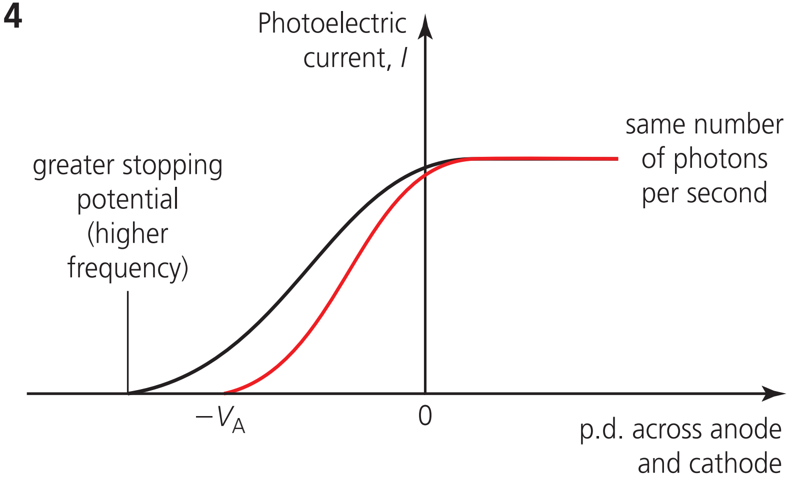


12 6.81×10^{-34} J s

13



14



16 **b** If you double the kinetic energy, the speed increases by the square root of 2. Thus, since the de Broglie wavelength is inversely proportional to speed for a non-relativistic particle, the wavelength will decrease by a factor of square root of 2
c If you double the speed, the de Broglie wavelength will decrease by half

20 3.55×10^{-11} m

21 $4.95 \times 10^{2}\,\mathrm{m\,s^{-1}}$

22 9.0×10^{-7} m

23 105 V

24 a The wavelength of the de Broglie wave associated with a particle is inversely proportional to the mass of the particle. The mass of an electron is less than that of a proton, so the wavelength of the de Broglie waves associated with the electron would be higher
b Electron, neutron, alpha particle, gold nucleus

25 An airplane has a relatively large mass; hence the wavelength of the de Broglie waves associated with it is too small to be observed and measured

29 a 15.3 eV
b $\lambda = 8.13 \times 10^{-8}$ m
c Shorter; there is an inverse relationship between energy and wavelength

30 The energy of an electron at rest outside the atom is taken as zero; when an electron 'falls' into the atom, energy is lost as electromagnetic radiation

32 $n = 1$: 6.0×10^{-18} J
$n = 2$: 2.4×10^{-17} J

34 1.22×10^{-7} m

38 5.8×10^{6} m s^{-1}

39 1.46×10^{-33} m

40 3.7×10^{-19} eV

43 2.0×10^{7} m s^{-1}

46 0.068 m

47 a Nuclear energies are significantly larger than electron energies.
b 1.4×10^{-11} m
c Gamma radiation

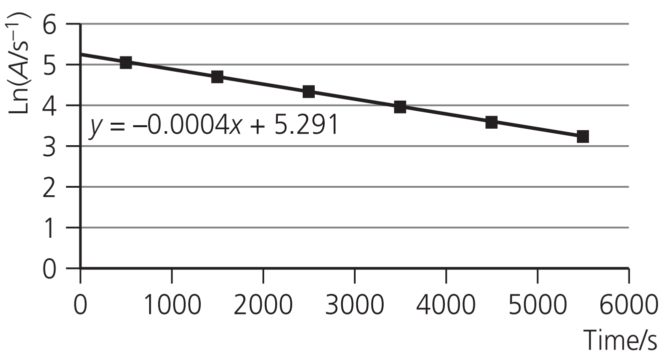
50 6.36×10^{23} atoms

51 a 126 Bq
b 805 nuclei

52 a 1.201×10^{-4} yr^{-1}
b 0.30

53 1620 years

55 λ is 0.0004; $T_{1/2} = 0.693/0.0004 = 1700$ s.



56 Let T_A and T_B be the half-lives of elements A and B.

$$A \xrightarrow{T_A} \tfrac{1}{2}A \xrightarrow{T_A} \tfrac{1}{4}A \xrightarrow{T_A} \tfrac{1}{8}A$$

$$B \xrightarrow{T_B} \tfrac{1}{2}B \xrightarrow{T_B} \tfrac{1}{4}B$$

$3T_A = 2T_B$; therefore $\dfrac{T_A}{T_B} = \dfrac{2}{3}$

Chapter 17 Digital technology

4 a 00001000
b 00001110
c 00010001
d 01000100
e 01111101

5 a 6
b 37
c 50
d 63

7 0100 0011 0100 0001 0100 0010

17 To achieve destructive interference, the path difference between light from the top of the pit the bottom of a land should be $\frac{1}{2}\lambda$. Because the light travels there and back, the pit height must be $\frac{1}{4}\lambda = \dfrac{620\text{ nm}}{4} = 155$ nm

18 If v is constant this means that $f \propto \dfrac{1}{r}$, so if the radius is trebled the frequency is reduced to a third of its original value, namely, 200 rpm

20 a 3.6×10^{9} pits
b 7.1×10^{9} bits
c 850 MB

21 a 3.1×10^{9} bits
b 1.9×10^{10} bits
c 2217 MB

26 0.25 V

27 120 μC

28 8×10^{-10} V

30 25 μF

31 10 μF

33 80%

34 60%

36 a 4.2 mm
b 1.2×10^{5} pixels
c 1.2×10^{-3}

37 a 1.7×10^4
b 12 500

38 0.125 s

39 240 000 pixels

43 a i 7.0 V
ii 12 V
iii –9.5 V
iv –12 V
b i 11 V
ii –6.5 V
iii –12 V

44 1×10^4

45 a 4 V
b 15 V or slightly less
c –15 V or slightly less

46 a From +125 μV to –125 μV
b It is the range of input voltages that the amplifier can amplify without distorting the signal

47 a 8.0 V
b 6.9 V
c 8.0 V

50 14 kΩ

51 a –2.0 V
b +4.0 V
c About +9.0 V (the op-amp is saturated)

52 a 0 V
b 0.25 μA
c 0.25 μA
d 0.30 V

55 a 6.0
b 5.0 V
c The op-amp is not saturated

56 a 21
b 4.2 V
c 0.40 mA
d 4.0 V

59 Such that $R_1/R_2 = \frac{1}{2}$.
For example R_1 = 220 kΩ and R_2 = 400 kΩ

61 So that the alarm only sounds when the light level falls, but not when it rises

62 a i 5.0 V **ii** 2.0 V
b 3.3 V
c 270 Ω

d At 20 °C the output of the op-amp is high, because the voltage from the temperature-sensing circuit at A is higher than the reference voltage at B, and so the LED will be on. As the temperature rises, the voltage at A will be falling (reaching 2 V at 100 °C). When the voltage falls below 3.3 V (the voltage at B) the LED will switch off

64 a 3.25 V
b It would switch at a lower voltage

66 To minimize any possible health hazards; to reduce the size of the battery needed (or extend the time needed between charges)

67 If cells used the same frequency, a phone could be linked to more than one base station at the same time, which would cause confusion. The radio waves would also interfere with each other

68 Few mobile phones per km²; flat land with no tall buildings; powerful, tall base station

69 a 8×10^{-8} W m^{-2}
b The radio waves travel only in straight lines, equally in all directions and they are not absorbed in the air
c The base station is much more sensitive to receiving radio waves

70 The shape of the aerials are designed to spread (diffract) the waves horizontally, but not vertically. This occurs because diffraction effects are greatest if the dimension of the transmitting aerial has the same size as the wavelength

71 b One reason could be due to the reflections of the radio waves from moving objects combining with the main signal to give interference effects

72 When the signal becomes weak because the phone is moving away from a particular base station, the cellular exchange automatically transfers the call to a base station with a stronger signal

Chapter 18 Astrophysics

1 Earth: 5500 kg m^{-3}, Jupiter: 1400 kg m^{-3}. The Earth is solid/liquid, but Jupiter is mainly gaseous

2 a 3.0×10^4 m s^{-1}
b Mercury moves around the Sun with a speed which is about 1.6× faster

3 a About 110 y
b It will not be visible because it is too far from Earth

4 a Mercury, 3.3×10^{23} kg
b The definition of a *planet* is somewhat arbitrary; it is decided by astronomers. Because a large number of objects of similar size to Pluto have been discovered in recent years in the outer reaches of the solar system, it was decided that for any object to be called a *planet*, it must have 'cleared the neighbourhood of its orbit'. Because there are many other large objects orbiting the Sun at a similar distance, Pluto was downgraded from the status of being a planet in 2006

5 Jupiter, 1.38×10^8 m

6 a 10^{24} km in diameter to the nearest order of magnitude
b 10^{10} pc in diameter to the nearest order of magnitude

7 a 4.2 ly
b 3100 km

8 a 0.01 ly
b About 4 km

9 5.0×10^2 s

10 a i About 8 months (using the data from Table 18.1 and assuming the planets are at their closest)
ii About 300 000 years

11 10^{24} (note that the data in Figure 18.8 has been 'rounded' for simplification, making this estimate too large)

12 3.5×10^{17} m

13 3.8×10^{27} W

14 1.5×10^{11} m

15 $L_A/L_B = 130$

16 7.6×10^{5} photons every second

17 3.2×10^{26} W

18 a 8.1×10^{19} m^2
b 2.5×10^{9} m

19 1.8×10^{4} K

20 9.5×10^{-10} W m^{-2}

21 2.8×10^{14} km

22 100/1

23 The star has a radius 2.2 times greater than the Sun

24 5.1×10^{-7} m (green)

25 a 4500 K
b 1.6×10^{22} m^2
c 3.6×10^{10} m

26 a 3.5×10^{-7} m
b 1.0×10^{28} W
c 7.2×10^{-9} W m^{-2}

27 3.6×10^{-9} m

29 a Spectral class G
b 'White light' does not have a precise scientific definition, it is the perception created in our brains when the three different kinds of receptors on the retina of our eyes are stimulated approximately equally. The light *emitted* by the Sun is considered to fit this description, although intensities at different wavelengths vary. The light *received* on the Earth's surface is slightly different because some radiation has been scattered by the atmosphere

30 Red giants: M (also G and K); white dwarfs: A (also B and F)

31 Spectral class B; blue-white

32 All lines of the spectra are shifted to lower frequencies (by equal amounts), compared to the lines of spectra observed from a source which has no relative motion. The shift occurs because the source is moving away from the observer

33 When the white light spectrum emitted from the Sun's core (which is travelling towards Earth) passes through its outer layers, helium atoms absorb photons of particular wavelengths, as they are excited to higher energy levels. The same energy is then re-emitted in *random* directions, so that the radiation of those particular wavelengths travelling towards Earth is greatly reduced

34 a i B to C
ii B to B (or F to F, or A to A)
b The drops in intensity during transit would be equal

35 The Doppler effect may be detected from any binary system except those in which the orbits are in a plane perpendicular to a line from the stars to the observer. The eclipsing effect can only be detected from binary systems in which the orbits are in the same plane as a line from the stars to the observer

36 The centres of both the Earth and the Moon orbit their combined centre of mass, which is about one quarter of the Earth's radius beneath its surface

37 a The elements hydrogen and helium will be present in nearly all stars. The stars will also probably have a comparable origin, size and age, so it is reasonable to assume their compositions may also be similar, so that their absorption spectra are alike
b The observed spectral shifts would be greater

38 Although white dwarfs have low luminosities, they are relatively hot, so that their surface areas must be small in order to reduce the emitted radiation

39 Yes, this term fits their location on the HR diagram

40 It will become cooler, bigger and more luminous

41 12 000 K, 5×10^{28} W

42 a 1.8 pc
b 5.6×10^{16} m
c 5.9 ly

43 a 0.0125 arc-seconds
b 0.41 arc-seconds
c 0.375 arc-seconds

44 a $1/3600 = 2.78 \times 10^{-4}$°
b 4.85×10^{-6} rad

45 4.4 times brighter

46 Yes, because the scale is logarithmic

47 +2.28

48 a Both
b 1.6×10^{-8} W m^{-2}

49 3.7 pc

50 1.2

51 –5.2

52 a 7800 K
b A
c About 40× Sun's luminosity $= 1.5 \times 10^{28}$ W
d 1.5×10^{17} m

53 (1) The measured apparent brightness may have been reduced by the effects on the radiation of travelling through interstellar space. (2) The HR diagram shows that stars which have the same surface temperature have a *range* of different luminosities, so that there is uncertainty in the estimate of luminosity from temperature

54 **a** Towards the top left
b Class B, 20 000 K
c $9.1 \times 10^{-12}\,W\,m^{-2}$

55 **a** $1.7 \times 10^{30}\,W$
b $1.3 \times 10^{-13}\,W\,m^{-2}$

56 **a** $2.3 \times 10^{30}\,W$
b About 20 days

57 The luminosities of different supernovae are always the same

58 (1) The red-shifts of the radiation received from receding galaxies indicates that their speed is proportional to their distance away. (2) The average temperature of a universe that started with a Big Bang would now be 2.7 K; radiation characteristic of this temperature is detected coming from all directions

59 Hydrogen, because it is the most common element in the universe

61 **a** Red-shift
b $2.33 \times 10^{13}\,Hz$
c $9.56 \times 10^{6}\,m\,s^{-1}$

62 **a** **i** Decrease **ii** Increase
b The same as on Earth

64 $2 \times 10^{52}\,kg$

67 At higher temperatures the mean kinetic energy of particles is greater, so that they are able to overcome the greater electric repulsion that occurs between nuclei with a greater number of positively charged protons. More massive stars produce greater gravitational attractions, so that the particles gain greater kinetic energies

68 A red supergiant has sufficient mass that the temperatures can become high enough for the fusion of heavier elements, but because iron has the highest binding energy per nucleon, the fusion of any heavier elements is not possible

69 'Stardust'. Hydrogen was formed throughout the universe at an early stage soon after the Big Bang. Atoms of the other lighter elements were/are formed in the cores of stars. Iron and heavier elements were/are formed in supernovae

71 **a** Less massive
b $1.6 \times 10^{30}\,kg$
c Longer, because although it has less mass, it will be cooler and have a smaller rate of fusion

73 Black holes do not emit radiation and neutron stars are so small that they have very low luminosities

74 From the gravitational effects they have on other stars (which we can detect)

75 **b** The lifetime of star A will be about 0.2 times that of star B

76 If the mass of a main sequence star is known, the Chandrasekhar limit enables astronomers to predict what will happen to it in the future; whether it will become a white dwarf or explode as a supernova

77 **a** $4.8 \times 10^{4}\,km\,s^{-1}$
b $1.1 \times 10^{18}\,kg\,m^{-3}$

79 $6.1 \times 10^{7}\,km\,h^{-1}$

80 The red-shift is 13 nm and the received wavelength is 423 nm

81 $5.87 \times 10^{16}\,Hz$

83 $1800\,km\,s^{-1}$

84 41 Mpc

85 100 Mpc

86 **a** $1.0 \times 10^{-26}\,kg\,m^{-3}$
b About 6

87 **a** $3.8 \times 10^{4}\,km\,s^{-1}$
b 520 Mpc

88 The sub-atomic particles had enough kinetic energy to overcome the attractive forces that hold atoms together

Chapter 19 Electromagnetic waves

1 **a** Ultraviolet
b Radio waves
c Microwaves
d Infrared

2 **a** 7.2×10^{18}
b The diameter of the X-ray circle should be 50 times the diameter of the visible light circle.
c Individual X-ray photons have much more energy than individual light photons

3 Excessive exposure can damage the skin, cause skin cancer and harm the retina of the eye

4 **a** Judging from the size of the aerial and the separation of wavefronts sketched on the diagram, the wavelength may be about 3 m; corresponding frequency is about 100 MHz
b So that waves will not be reflected and diffracted by nearby objects and therefore travel a greater distance over the Earth's surface

5 **a** $7.0 \times 10^{13}\,Hz$
b Infrared
c Greenhouse effect

6 **a** Red, 1.36; violet,1.40
b 1.32°

10 **a** $1.00 \times 10^{3}\,W\,m^{-2}$
b 3.4 mm
c $450\,W\,m^{-2}$
d The intensity from the light bulb is about $0.30\,W\,m^{-2}$; the laser is approximately 1500 times as intense

12 Infrared is significantly less absorbed in the fibre

13 631 nm

14 **a** 40 cm

b The second lens should be 'fatter' in the middle
c Both lenses were made from glass of the same refractive index

15 a 12.5 D
b It is made from a material of lower refractive index

16 5 cm

17 a Image is 17 mm tall and 13 cm from the lens
b 1.7

18 a Image is 14 cm tall and 86 cm from the lens
b 0.71

19 20 cm from lens

20 13 cm

21 a Inverted, diminished, real
c By changing the distance between the lens and where the image is formed

22 a 22 cm from the lens
b –0.5

23 a 22 cm from lens
b –9

24 7.8 cm

25 9.2 D

26 b 2.7

27 An upright, virtual image 20 cm from the lens; linear magnification = 5

28 10 cm

29 a 3.1
b Away from the object
c Decreases

30 12 D

31 0.024 mm

32 a 5.8 cm from the lens
b 4.3

33 The image has a size of 6.0 cm and is 30 cm from the eyepiece

34 a 10 cm
b 3.1
c 2.8 cm
d 2.4
e 7.5

35 a 9.19×10^{-3} rad
b 7.9°

36 a 90 cm
b 5
c Increasing the diameter of the objective will collect more light and decreasing the focal length of the eyepiece will increase the magnification
d Larger lenses, and lenses with shorter focal lengths tend to have more aberrations (faults)

39 The surfaces of more powerful lenses are more curved

40 a 4.5×10^{-3} rad
c 1.4 cm

41 The light that would fall on the slits would not be coherent, monochromatic or intense

42 1.54 mm

43 a 4.8×10^{-7} m
b The wavelength would be smaller than in air, so the fringe separation would decrease

44 a So that the gap was approximately the same size as the wavelength, to achieve maximum diffraction at each slit
b About 1 m

45 a The spacing of the fringes will increase
b The spacing of the fringes will increase
c The spacing of the fringes will increase
d The fringes will have coloured edges

46 260 m s^{-1}

47 16.0°

48 80 lines mm^{-1}

49 1.6 m

50 a About 1 m
b Red light has a greater speed in glass than blue light, so that it is refracted less by a prism; but its greater wavelength means that a greater angle is needed to introduce the path difference of a whole wavelength that is needed for constructive interference

51 Two

52 Because three times the wavelength of violet light is less than two times the wavelength of red light

53 a 3087 nm
b Ultraviolet
c By fluorescence or a suitable photovoltaic cell

54 This would have an effect similar to reducing the line separation, so that the pattern would broaden

55 i The line spacing is not comparable to the wavelength of X-rays
ii X-rays are too penetrating, so that they would pass through all of the grating largely unaffected

56 Maxima at $\sin\theta = 0.272$, 0.544 and 0.816

57 0.267°

59 a i 3.0×10^{4} eV
ii 4.8×10^{-15} J
b 1.0×10^{8} m s^{-1}
c It was assumed that the electrons had zero kinetic energy to begin with. Also, at speeds approaching the speed of light, particles have a relativistic increase in mass. This has been ignored

60 2.5×10^{-11} m

61 a 2.89×10^{4} V
b 4.34×10^{4} V

62 a 750 W
b 44 °C

63 a 4.77×10^{-15} J
b 8.01×10^{14} photons per second

64 a 2.10×10^{-11} m
b The bombarding electrons do not have enough energy to raise electrons in the metal atoms to high enough energy levels

66 13.8°, 28.5°, 45.7°, 72.6°

67 3.10×10^{-11} m

68 **a** $0.71d$
b 16.4°

69 1.3×10^{-10} m

71 A wide parallel beam of microwaves with wavelength of about 3 cm can be directed at a regular arrangement of identical plastic spheres which have a diameter similar to 3 cm. After passing through the model, a greater intensity of microwaves will be detected at some angles than others

72 **a** 6.4×10^{-7} m
b 0.40 mm

73 1.2×10^{-7} m, 3.6×10^{-7} m, 6.0×10^{-7} m, … etc.

74 **a** 5.3×10^{-4} rad or 0.03°
b **i** The spacing would increase
ii The spacing would increase

75 **a** The fringes are concentric circles with the spacing decreasing with increasing distance from the centre
b The fringes would still form closed loops, but would not be perfect circles

76 1.10×10^{-7} m

77 4.61×10^{-7} m or 6.92×10^{-7} m

78 1.4

81 **a** 22.5°
b 6.81×10^{-7} m
c Red

Answers to examination questions in Chapters 15 to 25

15 Sight and wave phenomena

Paper 3

Q1 **a** **i** The ability of the eye to focus objects which are different distances away

ii The range of distances from the eye over which objects are acceptably focused

b The eye muscles gradually relax and the lens in the eye gets thinner

Q2 **a** Magenta

b The red and blue lights were of approximately similar intensities

c When white light is incident on the filter it absorbs the blue end of the spectrum and when the remaining transmitted wavelengths are received by the eye the brain interprets them as the colour yellow

Q3 **a** Rods

b **i** and **ii** Any *one* of the three cone cells curves from Figure 15.8, correctly labelled; the maximum on the vertical scale should be approximately 50

c Responses from the *three* different types of cone cell are needed by the brain to interpret colour. If one (or two) of these types of cells is not functioning correctly, the eye will not be able to see some colours correctly

16 Quantum physics and nuclear physics

Paper 3

Q1 **a** The light emitted from a hydrogen discharge tube is directed onto a diffraction grating (or prism) after it has been made to pass through a slit to make it into a narrow beam. The light emerging from the grating/prism is observed on a screen (or with a telescope).

b 1.90 eV

Q2 **a** The probability of a nucleus decaying per unit time.

b **ii** 2.3×10^5 y

17 Digital technology

Paper 3

Q1 **a** A circuit whose output (signal) is proportional, and opposite, to the input

b **i** Any two from: voltage gain is infinite; infinite input impedance; zero output impedance; infinite open loop gain; infinite bandwidth; infinite slew rate

ii Any one from: the voltage gain is increased; noise in signals is reduced; less distortion of signals

c $G = -10$

d $20\,\mu$A

Q2 **a** A geographical area allocated a specific frequency

b Different frequencies avoid interference, so a large number of frequencies means small cells so small power needed

Q3 **a** The sign of the output voltage is the same as that of the input voltage

b **i** $G = 10$

ii $V_{out} = 20\,$mV

c Op-amp has a high input resistance and so takes little current; (open loop) gain is very large so potential difference between non-inverting input and inverting input is (effectively) zero, i.e. $V_{out} = V_{in}$ so $G = 1$

d **i** 3.0 V

ii The esistance between A and B is smaller than $2\,\text{M}\Omega$ so the voltmeter draws current

iii The voltmeter reads the output voltage of the amplifier and the input voltage is the potential difference to be measured; the two are equal since the gain is 1

Q4 P.d. across $10\,\text{k}\Omega = 10^4 I$; p.d. across $50\,\text{k}\Omega = 5.0 \times 10^4 I$; $I = 1.4 \times 10^{-4}$ A; $V_{in} = 2.4$ V

Alternatively, p.d. across $50\,\text{k}\Omega = \frac{50}{60}$, $(V_{in} + 6.0) = 7.0$ V; $50V_{in} + 300 = 420$; $V_{in} = 2.4$ V

18 Astrophysics

Paper 3

Q1 **a** **i** Nuclear fusion

ii The inwards radiation pressure is balanced by the outwards radiation and thermal gas pressures

b **i** The luminosity of a star is defined as the total power it radiates (in the form of electromagnetic waves)

ii Stars have different masses, resulting in different surface areas and temperatures

c **i** The apparent brightness of a star is defined as the intensity (power/receiving area) on Earth

ii Different luminosities, different distances from Earth

d **i** $7.2 \times 10^{-8}\,\text{W m}^{-2}$

ii Large stars which are relatively cool and therefore yellow/red in colour. They are not on the main sequence and they have a higher luminosity than most other stars, including red giants, because of their size. (Higher Level only: red supergiant stars are stars which have finished their lifetime on the main sequence and will explode as supernovae to become neutron stars or black holes, depending on their mass)

iii 8.5×10^{-7} m

iv M

Q2 **a** A constellation is a group of stars, which may not be close together, but which form a recognizable pattern as seen from Earth. A stellar cluster is a group of stars held relatively close together by gravity

b **i** P_A has the greater apparent brightness (or smaller apparent magnitude), which means that the intensity received on Earth is greater

ii P_A has the smaller absolute magnitude, which means that the emitted power (luminosity) is greater

c Use of $m - M = 5\lg(d/10)$ leads to similar values for distance, d, for the two stars

d The are bound by gravity in the same cluster and are the same distance from Earth

e 1.41×10^4

g White dwarf

Q3 **a** 2.7 K

b The average temperature of the early universe was extremely high. The Big Bang model predicts that, as it has expanded, the average temperature has fallen to the current value

c The Doppler shift of spectral lines indicates that distant galaxies are receding at a rate which is proportional to their distance away, so that in the past they must have been closer together

Q4 **a** Newton's model proposed a universe which was infinite, static and uniform. This would mean that light from stars would come from all directions at all times and the night sky would be bright in all directions. If there are regions without light from stars, then Newton's model cannot be accurate

b The Big Bang model proposes that space and time were created at a point, and that the universe has continually expanded since then. What we can observe of the universe is limited by the distance that light can travel in the time since the Big Bang. Using this model, the observable universe is therefore neither infinite nor static

Q5 **a** Since main sequence stars will conform to the quoted equation, showing that $8 \times 10^4 = 25^{3.5}$ confirms that X is a main sequence star

b The mass of star X is greater than Oppenheimer–Volkoff limit, this means that when it leaves the main sequence it will explode as a supernova and become a black hole.

Q6 **a** 3.8×10^6 m s^{-1}

b **i** Using the variation in luminosity of Cepheid variable stars within the cluster

19 Electromagnetic waves

Paper 3

Q1 **a** and **c** See Figure 19.28

b At infinity

d 2.0 cm from the lens

Q2 **a** **i** The sky is seen because of light scattered by the atmosphere. The Moon does not have an atmosphere, so it does not scatter light

ii Scattering is a wavelength-dependent effect. The blue end of the visible spectrum is more scattered than the red end of the spectrum

b The formation of ozone from oxygen involves the absorption of ultraviolet radiation

Q3 **a** The light waves are emitted randomly so that waves are generally not in phase with each other. Interference may occur in principle, but the effects are not observable

b **i** A population inversion is produced in the lasing material and it is stimulated to emit a large number of in-phase photons (of the same wavelength) at the same time

ii Coherence

c The laser scans across the black and white lines of the bar code and the signal reflected back to the detector from each line is either light or dark

d 2.0×10^4 m

Q4 **a** **i** See Figure 19.43

ii Use geometry to show $n\lambda = d\sin\theta$

b Maximum value of $n = 3.85$, so maximum order is 3

Q5 **a** **i** The maximum kinetic energy of the electrons is less than the difference

in energy level between the ground state and the next highest level
- **ii** Photons are produced when electrons rapidly decelerate. The smallest wavelength corresponds to the greatest possible photon energy, when *all* of the kinetic energy of the electron is given to a *single* photon
- **iii** 6.2 keV

b i Because the angles are equal, the path difference must be zero
- **ii** 3.71×10^{-10} m

Q6 a i Zero
- **ii** π or $\lambda/2$
- **iii** Zero

b 110 nm

24 Graphs and data analysis

Q1 a This is an equation for a straight line, but the graph is curved.

b i seconds, s.
- **ii** 2.35

c ii Line should be straight.

d i 120 m
- **ii** 0.62 s
- **iii** 0.079 (found from a large triangle drawn on the graph).

e $D = 0.62v + 0.079\,v^2$

f i ±0.06
- **ii** Uncertainty in v is unchanged, but the uncertainty in D (or D/V) is less.

Q2 a The lines are not straight and they do not pass through the origin.

b The size (radius) of the petrol can was not known or not taken into account, so that a constant zero-point error was introduced.

c $\lg(R) = n\lg(t) + \lg(k)$; n is gradient = 0.4

d i By taking values of the five lines on the first graph at time t = 20 ms.
- **ii** A thin, smooth line should be drawn through all the error bars. The R intercept is about 5.0 m.
- **iii** No. A straight line cannot be drawn through all the error bars and the origin.

Q3 a The vertical error bars should be 3 divisions long.

b The line should be thin and smooth, and pass within 1.5 divisions of all points.

c A curve is the only line that would pass through all the error bars.

d i 20 °C.
- **ii** 1.05 °C cm^{-1}, found using a large triangle.

e If the rate is proportional to the temperature gradient, 43/1.81 = 1.05/25, which it is (within a few percent). So the relationship is confirmed.

f A graph of $\ln\theta$ against x will have a gradient of $-k$.

Q4 b The best-fit line is not straight and does not pass through the origin

c i ±500 $m^2 s^{-2}$
- **ii** ±270 $m^2 s^{-2}$
- **iii** 25 $kg^{-1/2}\,m^{1/2}$

Glossary for Options chapters

The glossary contains key words from the IB Physics Diploma syllabus. This section takes words from Chapters 15–19. Entries that are IB syllabus-required definitions and statements are coloured in blue.

Chapter 15 – Option A: Sight and wave phenomena

All definitions in this option are standard level.

A

Accommodation The ability of the eye to focus on objects at different distances by changing the shape of the lens.

Antinodes The positions in a standing wave where the amplitude is greatest. *See also* nodes.

B

Blind spot The small circular area at the back of the retina where the optic nerve enters the eyeball. In this area, the number of light-sensitive cells is limited which impedes vision.

Boundary conditions (standing wave system) The conditions at the end of a standing wave system which will encourage either a node or an antinode at those places.

Brewster's angle The angle of incidence for which all reflected light is plane polarized parallel to the reflecting surface.

Brewster's law The tangent of the Brewster angle, ϕ, is equal to the refractive index, n, of the material that the light is entering: $n = \tan\phi$.

C

Ciliary muscles Muscles in the eye that control the shape of the lens, thereby focusing light on the retina.

Colour blindness Inability to see certain colours correctly, usually due to problems with the cone cells in the retina.

Cone cells Light receptors responsible for photopic vision; mostly located near the centre of the retina.

Cornea Curved, transparent surface on the front of the eye which is responsible for most of the refraction which forms images on the retina.

D

Depth of vision Range of distances from the eye over which objects can be focused acceptably as images on the retina.

Dilate To get larger.

Doppler effect When there is relative motion between a source of waves and an observer, the emitted frequency and the received frequency are not the same. This is sometimes called a Doppler shift.

Doppler effect, equation for use with EM waves $\Delta f = \frac{v}{c} f$.

Doppler effect, equations for use with sound waves

For a moving source:

$$f' = f\left(\frac{v}{v \pm u_s}\right).$$

For a moving observer:

$$f' = f\left(\frac{v \pm u_o}{v}\right).$$

F

Far point Furthest point from the eye that an object can be focused clearly; usually accepted to be at infinity.

Filter (colour) Partially transparent material that is used to transmit some wavelengths and absorb others.

First harmonic Another name for the fundamental mode of vibration.

Fundamental mode The simplest mode of vibration, with the fewest number of nodes and antinodes.

H

Humour, aqueous and vitreous Transparent fluids in the eye.

I

Iris Coloured, circular part of the outer eye which controls the size of the pupil and, therefore, the amount of light passing into the inner eye.

L

Lens Transparent material with regularly curved surfaces which can be used to refract light to a focus.

Liquid crystal State of matter with properties between those of a liquid and a solid. Certain kinds of liquid crystal can rotate the plane of polarization of light, dependent on a small p.d. applied across them. Such crystals are widely used in displays and computer screens (LCDs).

M

Malus' law Used for calculating the intensity of light transmitted by a polarizing filter: $I = I_0 \cos^2\theta$, where I_0 is incident intensity and θ is the angle between the polarizer axis and the plane of polarization of the light.

Modes of vibration The different ways in which a standing wave may be set up in a given system.

Monochromatic Composed of only one wavelength/frequency or a very narrow band of wavelengths/ frequencies.

N

Near point Closest point to the human eye at which an object can be focused clearly (without straining); usually accepted to be 25 cm from a normal eye.

Nodes The positions in a standing wave where the amplitude is zero. *See also* antinodes.

O

Optic nerve Nerve which transfers information about images from the retina to the brain.

Optically-active substances Substances which rotate the plane of polarization of light which is passing through them.

P

Perception (visual) Interpretation of the environment by the eye and the brain.

Photopic vision Vision of detailed, coloured images under good lighting conditions, achieved through the action of cone cells in the retina.

Pitch The sensation produced by sounds of different frequency.

Plane of polarization The plane in which all oscillations of a plane-polarized wave are occurring.

Polarimeter Instrument for measuring the angle of rotation of the plane of polarization that occurs when light is transmitted through an optically-active substance.

Polarization (plane) A property of transverse waves in which the oscillations are all in the same plane.

Polarizing filter A filter which transmits light which is polarized in one plane only. A filter used to produce polarized light from unpolarized light is called a polarizer. A polarizing filter which is rotated in order to analyse polarized light is called an analyzer. Crossed filters prevent all light from being transmitted.

Primary colours (additive) The colours red, green and blue. When lights of these three colours are added together in various proportions, it is possible to create the perception of almost any other colour. When added in equal proportions, white light is produced. The combination of two primary colours produces a secondary colour.

Pupil Circular aperture (hole) controlled by the iris, through which light enters into the inner eye.

R

Rayleigh's criterion Two point sources can *just* be resolved if the first minimum of the diffraction pattern of one occurs at the same angle as the central maximum of the other. This means that if the sources are observed through a narrow slit, they will *just* be resolved if they have an angular separation of $\theta = \frac{\lambda}{b}$.
For a circular aperture, $\theta = 1.22\frac{\lambda}{b}$.

Resolution The ability of an instrument (or an eye) to detect separate details.

Retina Light-sensitive inner surface of the eye on which images are formed.

Rod cell Highly sensitive light receptors located away from the centre of the retina; responsible for scotopic vision.

S

Scotopic vision Vision under poor lighting conditions where images are less detailed and less coloured; this is achieved through the action of rod cells.

Secondary colours The colours cyan, magenta and yellow. Secondary colours are formed when two of the three primary colours are added together in approximately equal proportions.

Secondary waves (wavelets) The propagation of waves in three dimensions can be explained by considering that each point on a wavefront is a source of spherical secondary waves.

Standing wave The kind of wave which may be formed by two similar travelling waves moving in opposite directions. The most important examples are formed when waves are reflected back upon themselves. The wave pattern does not move and the waves do not transfer energy. They are also known as stationary waves.

Chapter 16 – Option B: Quantum physics and nuclear physics

All definitions in this option are standard level.

A

Activity (radioactivity) The number of disintegrations per second that occur in a radioactive source. The SI unit of activity is the becquerel (Bq). Activity is given by the formula $A = \lambda N$, where λ represents the radioactive decay constant and N represents the number of unchanged atoms present in the source.

B

Bainbridge mass spectrometer A type of mass spectrometer in which a beam of positive ions is generated and is then subjected to the combined action of perpendicular electric and magnetic fields. For particles with velocity equal to $v = \frac{E}{B}$, the forces due to these two fields are equal and opposite, so they do not experience a resultant force. The particles pass through a slit into another magnetic field, moving in a semi-circular path and striking a detection plate. Since different isotopes have different masses, they trace out different radii.

Becquerel, Bq SI unit of (radio) activity, equal to one nuclear decay every second.

Beta-positive decay Nuclear decay accompanied by the emission of a positron (β^+ particle) from the nucleus.

Bohr model A theory of atomic structure that explains the spectrum of hydrogen atoms. It assumes that the electron orbiting around the nucleus can exist in certain energy states only.

Boundary condition (wave equation) A requirement to be met so that a wave equation can be solved.

C

Conjugate quantities A pair of physical variables describing a quantum-mechanical system such that either of them, but not both, can be specified precisely at the same time.

Coulomb scattering An interaction between two charged particles in which the Coulomb repulsive force is the dominant interaction.

D

Davisson–Germer experiment An experiment which verified the wave properties of matter by showing that a beam of electrons is diffracted by a crystal at an angle dependent upon the velocity of the electrons.

De Broglie's hypothesis and equation All particles exhibit wave-like properties which can be described by the de Broglie equation:

$$\lambda = \frac{h}{mv} = \frac{h}{p}$$

where λ represents the wavelength, h represents Planck's constant, m represents the mass of a particle moving at a velocity v and p represents the momentum of the particle.

Decay law and decay constant, λ The law of radioactive decay states that the number of nuclei that will decay per second ($\frac{\Delta N}{\Delta T}$ = the activity of the source) is proportional to the number of atoms (N) still present that have not yet decayed: $\frac{\Delta N}{\Delta t} = -\lambda N$, where λ represents a constant, known as the decay constant. The decay constant is defined as the probability of decay of a nucleus per unit time. The unit is s^{-1}. The decay constant is linked to the half life by the following equation: $T_{\frac{1}{2}} = \frac{\ln 2}{\lambda}$.

Discrete energy levels Energy which does not vary continuously but that is instead restricted to specific values.

E

'Electron-in-a-box' model A simplified quantum-mechanical system that confines an electron to a one-dimensional region of space. The energy levels of the electron are explained in terms of standing waves that fit into that space.

Exponential decay Activity, $A = \lambda N$, and this can be written as $-\frac{\Delta N}{\Delta T} = \lambda N$. This equation describes radioactive decay and has a mathematical solution in terms of an exponent: $N = N_0 e^{-\lambda t}$. N_0 is the number of undecayed nuclei at the start and N is the number remaining after time t.

H

Heisenberg uncertainty principle A fundamental principle of quantum mechanics which states that it is impossible to measure simultaneously the momentum and the position of a particle with infinite precision, $\Delta x \Delta p \geq \frac{h}{4\pi}$. The principle also applies to measurements of energy and time, $\Delta E \Delta t \geq \frac{h}{4\pi}$.

M

Mass spectrometer A device which can measure the masses and relative abundance of gaseous atoms and molecules.

Matter waves Waves that represent the behaviour of an elementary particle, atom or molecule under certain conditions. *See also* de Broglie's hypothesis.

N

Neutrino A low-mass, neutral and very weakly interacting particle whose emission from the nucleus accompanies the emission of electrons or positrons during beta decay, leading to a continuous distribution of beta-particle energies.

Nuclear energy level Discrete energy levels within the nucleus of an atom whose location and properties are governed by the rules of quantum mechanics.

Nuclear transition A change in nuclear energy level which results in the emission of a high-energy photon.

P

Photoelectric effect Ejection of electrons from a substance by incident electromagnetic radiation, especially by ultraviolet radiation. Sometimes called photo emission. The ejected electrons are called photoelectrons.

Photoelectric equation The maximum kinetic energy of an emitted photoelectron is the difference between the photon's energy and the work function:

$$KE_{max} = hf - \phi \text{ or } \frac{1}{2}mv_{max}^2 = hf - \phi$$

where ϕ represents the work function, hf represents the energy present in the incident photon and KE_{max} represents the maximum kinetic energy of the ejected photoelectrons.

Positron The antiparticle of the electron.

Potential hill The potential in a region in a field where the force exerted on a particle is such as to oppose the movement of the particle through the region.

Probability distribution The square of the absolute value of the wavefunction.

Q

Quantization (atoms) To limit the possible values of a quantity to a discrete set of values by quantum mechanical rules.

Quantum mechanics A branch of physics that describes the behaviour of objects of atomic and subatomic size.

S

Schrödinger wave equation An equation which represents the Schrödinger model of the hydrogen atom mathematically by describing electrons using wavefunctions. The square of the amplitude of the wavefunction gives the probability of finding the electron at a particular point.

Stopping potential The voltage required to reduce the photoelectric current to zero. $V_s = \frac{h}{e}f - \frac{\phi}{e}$, where V_s represents the stopping potential, h represents Planck's constant, f represents the frequency of the incident photons, ϕ represents the work function and e represents the charge on the electron.

T

Threshold frequency The minimum frequency of a photon that can eject an electron from the surface of a metal.

W

Wavefunction Mathematical function of space and time which describes the quantum state of a subatomic particle, such as an electron. It is a solution to the Schrödinger wave equation.

Wave–particle duality The wave–particle duality principle of quantum physics holds that matter and light exhibit the behaviours of both waves and particles, depending upon the circumstances of the experiment. However, it is not possible to observe both properties simultaneously.

Weak nuclear force A short-range force acting within the nucleus, responsible for beta decay.

Work function The minimum amount of energy required to free an electron from the pull of atoms of the metal is called the work function, ϕ. Since the energy of the incident photons is equal to hf, $hf_0 = \phi$, where f_0 represents the threshold frequency and h represents Planck's constant.

Chapter 17 – Option C: Digital technology

All definitions in this option are standard level.

A

Aerial Conductor designed to be used in a transmitter to use oscillating electrical currents efficiently to produce radio waves. (Alternatively, it can be used in a receiver to generate oscillating currents from incident radio waves.) Also called an antenna.

American Standard Code for Information Interchange *See also* ASCII code.

Amplifier A device that increases the amplitude of an electrical signal wave.

Amplifier, differential An amplifier in which the output depends on the difference between the voltages at its two inputs.

Amplifier, inverting and non-inverting The output voltage of an amplifier may have the same sign as the input, or the opposite sign, depending on which of the two inputs is used.

Amplifier, linear Amplifier which produces output voltages which are proportional to the input voltages over a wide range of frequencies.

Amplifier, operational (op-amp) High-gain differential amplifier, usually in the form of an IC (integrated circuit). An *ideal* op-amp has infinite input resistance, zero output resistance and infinite open-loop gain.

Analogue signal A continuous signal that is proportional to the physical property which created it.

Analogue-to-digital converter A device which translates continuous analogue signals into proportional discrete digital signals.

ASCII code An encoding system for the transfer of digital data, in which the letters of the alphabet, numbers and other characters are given an 8-bit binary equivalent.

Attenuation The process by which intensity is reduced during transmission of energy through a medium.

B

Base station That part of the cellular phone system which communicates directly with mobile phones, using a range of frequencies which is different from that of neighbouring base stations.

Binary numbers System that represents all numbers using only two digits: binary one (1) and binary zero (0).

Bit A bit is the smallest unit of data in computing, with a value of either binary 0 or binary 1.

Byte A set of eight binary digits.

C

Capacitance The ratio of charge on an electrically charged, isolated conductor to the potential difference across the conductor. For a parallel plate capacitor this is equal to the ratio of the electric charge transferred between the plates to the resulting potential difference:

$$\text{capacitance} = \frac{\text{charge stored on one plate}}{\text{potential difference between plates}}$$

Capacitor An electric circuit component used to store small amounts of energy temporarily. It often consists of two metallic plates separated and insulated from each other.

Capacity (data storage) The amount of data that can be stored on a storage device. In a communications system it is the amount of data delivered (with no error) from one device to another.

Cassette tape A compact case containing a length of magnetic tape that runs between two small reels; used for recording or playing in a tape recorder and by computer systems to back up data.

CD-ROM (Compact Disc Read-Only Memory) An optical disc used to store computer data.

Cell (mobile phone) Relatively small radio-wave reception area for mobile phones which is controlled by one base station.

Cellular exchange Control centre for a group of base stations, linking them to the public switched telephone network (PSTN).

Charge-coupled device (CCD) A semi-conducting device where incident light causes the build up of electric charge in individual pixels producing an image of the object. The amount of charge is proportional to the intensity of light.

Comparator Circuit in which the output voltage switches on or off (high or low) depending on how one variable-input voltage compares to the other fixed-input voltage.

D

Digital signal A coded signal that can have one of two values (binary 0 or binary 1).

Digital Versatile Disc (DVD) An optical disc used to store computer data. A DVD can contain audio, video or data.

Digitize Convert an analogue or continuous signal into digital format.

E

Earth connection Connection made between a point in a circuit and the earth, ensuring that the point remains at 0 V. Also called ground connection.

Earth, virtual Point in a circuit which is assumed, for the sake of simplifying circuit analysis, to be at 0 V.

F

Farad The unit of capacitance in the SI system equal to the capacitance of a capacitor having an equal and opposite charge of one coulomb on each plate and a potential difference of one volt between the plates.

Feedback, negative and positive Technique in which the output of a process is used to alter the input, so affecting future outcomes.

Floppy disc A removable data storage disc which uses a flexible magnetic medium in a plastic case.

G

Gain (voltage), G Ratio of output voltage to input voltage for an amplifier.

Gain: closed loop Gain of an amplifier using feedback. $G = \frac{V_{out}}{V_{in}} = -\frac{R_F}{R}$ for an inverting amplifier and $G = \frac{V_{out}}{V_{in}} = 1 + \frac{R_F}{R}$ for a non-inverting amplifier.

Gain: open loop Gain of an amplifier without feedback.

Ground connection *See also* earth connection.

H

Hard disc A flat, circular, rigid plate with a magnetizable surface on one or both sides; used to store data.

K

Kilobyte 1024 bytes. This is approximately 1000 bytes.

L

Land A reflective area of an optical medium. It reflects the laser into a sensor to register it as a binary digit one.

Least-significant bit (LSB) The bit is a binary number that is the least important or has the least weight. It is the right-most bit of a binary number.

Logic gate A device, usually an electrical circuit, that performs a logical operation on one or more input signals (1s or 0s). Logic gates are the building blocks of digital technology.

Long-playing disc (LP) An analogue audio recording pressed in vinyl plastic that rotates at 33.3 revolutions per minute. The sound is encoded in a spiral groove, starting at the outer edge of the disc.

M

Magnification (of a CCD) The ratio of the length of the image on the CCD to the length of the object.

Megabyte A megabyte contains 1 048 576 bytes (1024 × 1024 bytes). This is approximately 1 million bytes.

Most-significant bit (MSB) In a binary number, the MSB is the most weighted or most significant bit. It is the left-most bit of a binary number.

N

Noise Random, unwanted electrical disturbances that reduce the integrity of electronic signals. Noise can come from a variety of sources, including heat, other users of the same radio frequency, nearby electrical wires, lightning and bad connections.

O

Optical media (storage) Types of data storage media that use light, usually in the form of a laser, to read and/or write data. Most commonly involve patterns on discs, such as DVDs.

P

Pit A succession of pits form a track on a CD. The pits are usually one-quarter of the laser wavelength in depth and cause cancellation of the beam by interference.

Pixel Picture element; the smallest element in a CCD (or visual display).

Portability Stored data is portable if it can be transported easily on the storage device or medium.

Public switched telephone network (PSTN) System through which the world's phones are all interconnected.

Pulse amplitude modulation (PAM) The transmission of data by varying the amplitudes (voltage or power levels) of the individual pulses in a regularly timed sequence of electrical pulses.

Q

Quality of data A measure used to define the maximum allowable number of bits in error in a digital communication system, such that if the number of bits rose above this threshold, the quality of data would be unacceptable.

Quantum efficiency (of a CCD) The ratio of the number of electrons emitted to the number of incident photons on a pixel.

R

Register (CCD) The area of the CCD which receives the binary information for processing and storage before passing it to the analogue-to-digital converter.

Relay Electrically operated switch.

Reproducibility (data) The ability of a computer system or electronic device to produce the same value or result, given the same input conditions and operating in the same environment.

Resolution (CCD) The ability of an instrument (or an eye) to detect separate details.

Retrieve (data) To recover or make newly available (stored digital information) from a computer system. The retrieval speed is an important measure of how quickly a device can gain access to stored data.

S

Sampling Digitizing an analogue waveform by measuring its amplitude fluctuations at some precisely timed intervals.

Schmitt trigger Circuit using an op-amp with feedback to regenerate distorted digital signals.

T

Tracking (mobile phones) Continuous monitoring of the location of mobile phones.

W

Word The largest number of bits the processor (CPU) of an electronic device can process at one time.

Chapter 18 – Option E: Astrophysics

Standard and Higher level

(see page 810 for Higher level only)

A

Apparent brightness, b Intensity (power/area) of radiation received on Earth from a star (unit: W m^{-2}). Related to luminosity by the following equation: $b = \frac{L}{4\pi d^2}$.

Arc-second (arcsec) 1/3600 of a degree.

Asteroids Rocky objects of various sizes, but smaller than planets, which orbit the Sun in the inner solar system. Most asteroids are located in the *asteroid belt* between the orbits of Mars and Jupiter.

Astronomical unit (AU) Unit of distance used by astronomers equal to an agreed average distance between the Sun and the Earth.

B

Big Bang model Currently accepted model of the universe, in which the universe began at a point 13.7 billion years ago and has expanded ever since.

Binary star systems Two stars orbiting their common centre of mass. They can be classified as *visual*, *spectroscopic* or *eclipsing* binaries, depending on how they are detected.

Blue-shift The spectra of radiation received from stars and (the relatively few) galaxies which are moving towards Earth are shifted towards shorter wavelengths.

C

Cepheid variable star Type of star which is very useful in determining the distance to galaxies. The luminosity (absolute magnitude) of a Cepheid variable changes in a predictable way and its magnitude can be estimated from a measurement of its time period.

Comet Relatively small object of ice, dust and rock which moves around the Sun, usually with a very elliptical orbit and long period. Some have 'tails' which are visible from Earth with the unaided eye when they are close to the Sun.

Constellation An area of the night sky defined and named by the pattern of visible stars it contains. The stars may appear relatively close together, but in practice they are probably a long way apart and unconnected. *Compare with* stellar cluster.

Cosmic microwave background (CMB) radiation Spectrum of electromagnetic radiation received almost equally from all directions and characteristic of a temperature of 2.7 K. CMB radiation is evidence in support of the Big Bang model.

Cosmology Study of the universe (cosmos).

Critical density (of the universe) The average density which would result in a *flat* universe.

D

Dark energy Unknown form of energy whose existence has been postulated to explain the accelerating expansion of the universe. It is believed to account for about three-quarters of the total energy in the universe.

Dark matter Matter which has not been detected directly, yet is believed to make up more than 80% of the total matter in the universe.

Doppler effect (shift) When there is relative motion between the source of waves and an observer, the received frequency is different from the emitted frequency. *See also* red-shift and blue-shift.

E

Elliptical In the shape of an ellipse (oval). An ellipse has two foci on its major axis.

Empirical Based on observation or experiment.

G

Galaxy A very large number of stars (and other matter) held together in a group by the forces of gravity.

Gravitational pressure (in a star) Pressure acting inwards in a star due to gravitational forces.

H

Hertzsprung–Russell (HR) diagram Diagram which displays order in the apparent diversity of stars by plotting the luminosity (or absolute magnitudes) of stars against their surface temperatures (or spectral class).

L

Light year, ly Unit of distance used by astronomers equal to the distance travelled by light in a vacuum in one year.

Luminosity, *L* Total power radiated by a star (unit: W). $L = \sigma A T^4$, where A is the surface area, T is the temperature (K) and σ is the Stefan–Boltzmann constant.

M

MACHOs (Massive Astronomical Compact Halo Objects) This is a general term for any kind of massive astronomical body that might explain the apparent presence of dark matter in the universe.

Magnitude, absolute, *M* An arbitrary scale widely used by astronomers to represent the luminosity of a star. It is defined as the apparent magnitude, *m*, that a star would have if it was observed from a distance of 10 parsec (pc). Mathematical relationship: $m - M = 5\lg\left(\frac{d}{10}\right)$, where d is the distance to the star in pc.

Magnitude, apparent, *m* An arbitrary scale widely used by astronomers to represent the apparent brightness of a star.

Main sequence The band of stable stars which runs from top left to bottom right on the Hertzsprung–Russell diagram. Most stars are located in the main sequence.

Milky Way The galaxy in which our solar system is located.

Moons Massive objects which orbit planets.

N

Newton's model of the universe An infinite, uniform and static universe.

Nuclear fusion Process in which lighter nuclei join to make a heavier nucleus, with the release of energy. Nuclear fusion is the main energy source of stars.

O

Olbers' paradox If the universe is infinite, with an infinite number of stars (Newton's model), then there should be no such thing as a dark sky.

P

Parallax angle (stellar) Largest possible angle subtended between imaginary lines drawn from a nearby star to the Earth and to the Sun.

Parsec, pc Unit of distance used by astronomers; equal to the distance to an object which has a parallax angle of one arc-second.

Period, *T* Time taken for one complete orbit (or other regularly repeating event).

Period–luminosity relationship Graph used with Cepheid variables to determine their luminosity from knowledge of the period of the oscillations of their luminosity. This enables their distance from Earth to be determined.

Planet A massive, approximately spherical object which orbits a star without any significant influence from other objects.

R

Radiation pressure (in a star) Pressure in a star due to radiation emitted.

Red giant (and red supergiant) stars Relatively cool stars which are yellow/red in colour; their luminosity is high because of their large size.

Red-shift The spectra of radiation received from stars and galaxies moving away from Earth are shifted towards longer wavelengths. Red-shift is evidence in support of theories of an expanding universe (including the Big Bang model).

S

Solar system The Sun and all the objects which orbit around it.

Spectral class A way of classifying stars from the visual appearance of their spectra. It is related to their surface temperature and colour.

Spectroscopic parallax Method of calculating the distance to a star by using the H–R diagram to determine a star's luminosity from its spectrum. Limited to distances less than about 10 kpc.

Standard candle Term used by astronomers to describe the fact that the distance to a galaxy can be estimated from a knowledge of the luminosity of a certain kind of star within it (such as a Cepheid variable).

Star Massive sphere of plasma held together by the force of gravity. Because of the high temperatures, thermonuclear fusion occurs and radiation is emitted.

Stellar cluster A group of stars which are relatively close because they are bound together by the forces of gravity. *Compare with* constellation.

Stellar equilibrium Main sequence stars are in equilibrium under the effects of thermal gas pressure and radiation pressure outwards, acting against gravitational pressure inwards.

Stellar parallax Method of determining the distance, d, to a nearby star from measurement of its parallax angle, p:

$$d\ (\text{parsec}) = \frac{1}{p\ (\text{arc-second})}.$$

Sun The object around which the Earth orbits. A main sequence star.

T

Thermal gas pressure (in a star) Pressure in a star due to the motion of the particles within it.

U

Universe All existing space, matter and energy; also called the cosmos. There may be many universes.

Universe (observable) That part of our universe that we are theoretically able to observe from Earth at this time. What we can observe is limited by the age of the universe and the speed of light.

Universe (open, flat, closed) Three different possibilities for the future of the universe. Current scientific evidence suggests that the universe may be open, although it is close to being flat.

W

White dwarf stars Relatively hot stars which are blue/white in colour, but their luminosity is low because of their small size.

Wien's (displacement) law Law which connects the wavelength at which the greatest intensity is emitted from a star to its surface temperature: $\lambda_{max}T$ = constant.

WIMPs (Weakly Interacting Massive Particles) This is a general term for any currently undetected fundamental particles that might explain the apparent presence of dark matter in the universe.

Higher level only

B

Black hole After a supernova, the remaining core of a red supergiant, which is too massive to form a neutron star, will become a black hole, with the forces of gravity so high that light cannot escape.

C

Chandrasekhar limit Maximum mass of a star that can become a stable white dwarf. More massive stars will become neutron stars or black holes.

Clusters of galaxies Groups of galaxies mutually bound together by the forces of gravity.

E

Electron degeneracy pressure Process occurring in a white dwarf star that enables it to remain stable for a long time.

Evolutionary path (of a star) The track of a star across the Hertzsprung–Russell diagram after it leaves the main sequence.

G

Galactic cluster Group of stars within a single galaxy. *Compare with* clusters of galaxies.

H

Hubble's law The current velocity of recession (the speed at which a galaxy appears to be moving directly away from Earth), v, of a galaxy is proportional to its distance away (from Earth), d: $v = H_0 d$, where H_0 is the Hubble constant.

M

Mass–luminosity relationship Equation which links the luminosity of a main sequence star to its mass: $L \propto m^n$ (where $3 < n < 4$).

N

Nebula Cloud of mainly hydrogen and dust (plus some helium and other gases) in interstellar space. (Plural: nebulae.)

Neutron degeneracy pressure Process occurring in a neutron star that enables it to remain stable for a long time.

Neutron star After a supernova, the remaining core of a red supergiant which is not massive enough to form a black hole, will form a neutron star.

Nucleosynthesis Formation of heavier nuclei from the fusion of lighter nuclei.

O

Oppenheimer–Volkoff limit Maximum mass of a star that can become a neutron star. More massive stars will become black holes.

P

Planetary nebula Material ejected from the outer layers of a red giant star when its nuclear fuel has run out. (No connection with planets.) The remaining core is called a white dwarf.

Plasma State of matter containing a high proportion of separated charged particles (ions and electrons).

Pulsar Neutron star which rotates quickly and continuously emits a directional beam of electromagnetic radiation. The radiation received on Earth appears to be pulsed because it is only detected once in every rotation, when the beam is pointing in a suitable direction.

R

Recession speed and red-shift Measurement of red-shift can be used to calculate the recession speed from the following equation:

$$\frac{\Delta\lambda}{\lambda} \approx \frac{v}{c}.$$

S

Stellar mass and the results of fusion Stars with greater mass will have greater temperatures and will therefore be able to fuse heavier elements.

Superclusters (of galaxies) Groups of clusters of galaxies. Probably the largest structures in the universe.

Supernova A relatively brief, but enormous, explosion of a massive red supergiant star that occurs when its nuclear fuel has run out.

Chapter 19 – Option G: Electromagnetic waves

Standard and Higher level

(see page 812 for Higher level only)

A

Aberration, chromatic Inability of a lens to bring light of different colours (coming from the same place) to the same focus.

Aberration, spherical Inability of a lens with spherical surfaces to produce a sharp focus.

Absorption (EM waves) Transfer of wave/photon energy to other forms within a medium, so that it is not transmitted or scattered.

Aerial Conductor designed to be used in a transmitter to use oscillating electrical currents efficiently to produce radio waves. (Alternatively, it can be used in a receiver to generate oscillating currents from incident radio waves.) Also called an antenna.

Amplify Increase the amplitude of a wave/signal.

Astronomical telescope Two (or more) convex lenses used to produce an angular magnification of a distant object. The image is inverted.

Attenuation Gradual loss of intensity due to absorption and scattering as electromagnetic waves pass through a medium.

C

Conditions necessary to observe interference between two sources Sources must produce waves which are coherent, in other words having the same frequency/wavelength and a constant phase difference. Constructive interference occurs at angles, θ, such that $\sin\theta = \frac{n\lambda}{d}$, where d is the separation of the sources and n is an integer.

Converging (convex) lens Lens which is thicker at the centre than at the edges, which can bring parallel rays to a focus to form a real image.

D

Diffraction grating A large number of very narrow parallel slits/lines which are very close together that can be used to disperse light into spectra. Constructive interference occurs at angles, θ, such that $d \sin\theta = n\lambda$, where d is the separation of the lines on the grating and n is an integer.

Diffraction order Numbering of the spectra produced by a diffraction grating, starting from the centre of the pattern.

Dioptre, D Unit for the power of a lens. One dioptre is the power of a lens which has a focal length of 1 m.

Diverging (concave) lens Lens which is thinner at the centre than at the edges, which diverges rays to form a virtual image.

E

Eyepiece lens Lens in an optical instrument which is closest to the eye.

F

Far point Furthest point from the human eye that an object can be focused clearly; usually accepted to be at infinity for normal vision.

Focal length, *f* Defined as the distance between the centre of a lens and the focal point.

Focal point (of a convex lens) Defined as the point through which all rays parallel to the principal axis converge after passing through the lens; sometimes called the principal focus.

Focus To cause light rays/waves to converge to a point or appear to diverge from a point.

Fringes (interference) Pattern of light and dark (usually regularly shaped) caused by constructive and destructive interference.

I

Image distance, *v* Distance between an image and the centre of a lens.

Image (optical) Representation of an object that our eyes and brain 'see'. *See also* image, real and image, virtual.

Image, real Image formed at a place where light rays/waves converge.

Image, virtual Image formed at a place from which light rays/waves appear to diverge.

L

Laser (Light Amplification by Stimulated Emission of Radiation) Device which produces a coherent, highly directional beam of monochromatic radiation.

Lasing material The medium in which stimulated emission occurs in a laser.

M

Magnification, angular, M Defined as the angle subtended at the eye by the image/angle subtended at the eye by the object: $M = \frac{\theta_i}{\theta_o}$.

Magnification, angular, of an astronomical telescope

$M = \frac{f_o}{f_e}$ for normal adjustment with the image at infinity, where f_e is the focal length of the eyepiece and f_o is the focal length of the objective.

Magnification, angular, of simple magnifying glass

$M = \frac{D}{f} + 1$ for an image at the near point (normal adjustment) and $M = \frac{D}{f}$ for an image at infinity, where D is distance to near point and f is the focal length.

Magnification, linear, *m* Defined as height of image/height of object:

$m = \frac{h_i}{h_o} \left(= -\frac{v}{u}\right)$.

Magnifying glass (simple) Single convex lens used to magnify a close object.

Metastable energy state A higher and less-stable energy level than the ground state, in which atoms may remain for longer times than usual.

Microscope, compound Two (or more) convex lenses used to produce a greater magnification of a close object than is possible with a simple magnifying glass.

Monochromatic Composed of only one wavelength/frequency or a very narrow band of wavelengths/frequencies.

N

Near point Nearest point to the eye at which an object can be focused clearly (without straining). Usually accepted to be 25 cm from a normal eye. This distance is sometimes given the symbol D.

Normal adjustment For a compound microscope: image formed at the near point. For an astronomical telescope: image formed at infinity.

O

Object (optical) The thing which emits or reflects light, such that an image is formed.

Object distance, u Distance between an object and the centre of a lens.

Objective lens Lens in an optical instrument which is closest to the object.

Opaque A description of a medium in which electromagnetic waves/photons are absorbed and/or scattered so that none are transmitted.

Ozone layer(s) Part of the upper atmosphere formed by the interaction of ultraviolet radiation with oxygen.

P

Parallel rays Used to represent plane wavefronts, often originating from a distant source.

Population inversion More atoms are in a higher energy state than a lower energy state.

Power (of lens), P Defined as $\frac{1}{\text{focal length}}$: $P = \frac{1}{f}$.

Principal axis Defined as the (imaginary) straight line passing through the centre of a lens which is perpendicular to the surfaces.

Propagate (EM waves) Pass through a medium or vacuum.

Pump source (of laser) That part of a laser that supplies the energy.

R

Ray diagram Scale drawing showing the paths of rays of light from an object, through one or more lenses, to an image. A ray diagram can be used to predict the properties of the image.

'Real is positive, virtual is negative' convention The distances to virtual images are given negative values, so that $m = -\left(\frac{v}{u}\right)$ using the equation upright virtual images will always have positive magnifications, and inverted real images will always have negative magnifications.

Retina Light-sensitive inner surface of the eye on which images are formed.

S

Scattering (EM waves) Various processes in which the directions of waves are changed while they are passing through a medium.

T

Thin lens formula Formula that can be used for light striking a thin lens close to the principal axis. It relates the object distance, the image distance and the focal length: $\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$.

Transmission (EM waves) Sending waves from one place to another without absorption and/or scattering.

Transparent A description of a medium through which electromagnetic waves are transmitted without absorption and/or scattering.

Higher level only

B

Blooming Coating a glass surface with a very thin layer of a transparent material in order to reduce reflections off its surface.

Bragg scattering equation Equation predicting the angles, θ, for constructive interference in an X-ray diffraction pattern: $2d \sin\theta = n\lambda$, where d is the separation of planes.

C

Conditions for constructive and destructive interference in parallel films: normal incidence $2nt = (m + \frac{1}{2})\lambda$ for constructive interference; $2nt = m\lambda$ for destructive interference, where n is the refractive index of the medium, m is an integer and t is the thickness of the film.

Conditions for constructive and destructive interference in parallel films: oblique incidence Constructive interference will occur if $2nt \cos\phi = (m + \frac{1}{2})\lambda$. Destructive interference will occur if $2nt \cos\phi = m\lambda$, where n is the refractive index of the medium, m is an integer, ϕ is the angle of refraction and t is the thickness of the film.

Conditions for reflected light to undergo a phase change Light which reflects off the surface of a medium with a greater refractive index undergoes a phase change of Ω. There is no phase change if the reflection is off a medium with lower refractive index.

Crystalline Atoms, ions or molecules of a solid arranged, to some extent, in ordered patterns.

H

Hardness of X-rays Ability of X-rays to penetrate matter.

L

Lattice Three-dimensional arrangement of points.

O

Optical flat Piece of glass which has been ground and polished so that it is extremely flat, varying by much less than the wavelength of light.

T

Thermionic emission Release of charged particles (usually electrons) from a hot surface.

Thin film interference Light reflecting off two parallel surfaces of a very thin layer of a transparent substance may interfere constructively or destructively, depending on the thickness of the film and the angle of incidence.

W

Wedge film Thin film of variable thickness. With a wedge of equal inclination (constant angle), the result will be a series of fringes of equal thickness.

X

X-ray spectrum, characteristic Sharp peaks of intensity on an X-ray spectrum which are characteristic of the energy levels in the particular metal used in the anode of the X-ray tube.

X-ray spectrum, continuous Broad range of wavelengths produced from an X-ray tube, usually represented by a graph of intensity–wavelength.

X-ray spectrum: minimum wavelength The smallest wavelength of a continuous spectrum, corresponding to the maximum possible photon energy: $\lambda_{min} = hc/eV$.

X-ray tube Apparatus designed to produce X-rays. In a Coolidge tube the X-rays are emitted when high-energy electrons are rapidly decelerated.

Index

Note: entries in blue refer to material in the Options chapters.

20 Teaching and learning IB physics

Whether you are a teacher or a student, you can feel both privileged and excited to be part of the International Baccalaureate Organization (IBO) and its Diploma programme. This model of education is rapidly gaining popularity throughout the world and is already the preferred access/entry course for many universities. It is easy to see why the programme has gained such respect when one appreciates its philosophical roots. The IB Diploma furthers the education of the whole person, not just by requiring knowledge and skills, but also by engendering personal attributes and creating **lifelong learners**.

This approach is encapsulated in a document available from the IBO called the **IB *learner profile***. While this document initially seems to be aimed at students, it is clear that the best IB teachers are those who consider *themselves* to be lifelong learners, so the IB *learner profile* is also useful to teachers. If you are a teacher who is new to IB, then prepare to be a learner too.

20.1 To all IB learners – both teachers and students

This chapter is designed to help students and teachers to develop an overview of the IB Diploma physics course and to appreciate the necessary approaches for success. It enables you to anticipate what the IB Diploma experience will entail and helps you to make the most of it by:

- briefly summarizing the philosophy and approach of the IB Diploma programme and the place of IB Diploma physics within it
- describing the components of the IB Diploma physics course as experienced by students and teachers
- providing straightforward practical advice to ensure your success.

The sections below are intentionally brief and to the point. Further details of procedures can be obtained directly from the IBO via the school's IB Coordinator.

What distinguishes the IB physics approach?

This physics course is one important component within a much greater scheme: the IB Diploma.

- A philosophy of **integration** binds all aspects of the programme. Physics will not stand alone, but will be an important vehicle for exploring and developing the various aspects of the IB *learner profile*.
- There will be a clearly international outlook, considering global issues as well as local case studies.
- Central to all subjects will be the work done on the **IB Theory of knowledge** (TOK) course, which creates a philosophical backdrop against which to discuss issues raised within specific subjects. References to TOK issues have been included within the chapters of the book to help students make these links. For example, within Chapter 13 (Quantum physics and nuclear physics) students are introduced to the wave–particle duality model of light; the wave model is necessary to explain diffraction, but only a particle model can explain the photoelectric effect. Students will also be introduced to the idea of a paradigm shift – a change in thinking – for example, from Newtonian classical mechanics to Einstein's theory of special relativity.
- This **broad** and **transdisciplinary** approach, together with the IBO's periodic syllabus review cycle (usually six or seven years), ensures that the syllabus is both **contemporary and relevant** by including important issues of the time. This is illustrated by the inclusion of topics on digital technology, and energy, power and climate change.

Figure 20.1 Second year IB Diploma students at the United World College of South East Asia brainstorm topics and approaches for their Group 4 project

- Following this course is not simply a matter of learning the facts and principles of physics. Students will find that developing informed viewpoints, especially on environmental and ethical issues, engaging in discussion, and evaluating hypotheses and experimental evidence are all needed for success at IB Diploma physics (Figure 20.1). Students will also come to appreciate that definite, clear answers are not always possible, for example, in the hazards related to high voltage transmission lines, mobile phones and nuclear technology.

It is these special features of the IB Diploma that make the course so engaging for both students and teachers.

What will I experience as a student or teacher of IB Diploma physics?

The student's subject-specific experience of IB Diploma physics will include the following four components:

- **Theory** – developing knowledge and understanding of the facts and theories underlying the topics as dictated by the syllabus
- **Practical** – developing the methodological (practical) skills needed to be successful in the internal assessment of practical work
- **Group 4 project** – engaging in collaborative, constructivist research with science students from other disciplines
- **Extended essay** – engaging in an independent research project, possibly on a physics topic (see Chapter 21).

The most successful IB physics courses are those where these elements are integrated as far as is practically possible.

What will each of these components involve?

Studying physical theory and aiming for success in the external examination

This exciting syllabus leads student to an understanding of physics topics, from basic mechanics and kinematics to quantum and nuclear physics.

The subject matter is divided into:

- the **Core syllabus** which is studied by all students, giving both breadth and depth of understanding
- the **Additional Higher Level** (AHL) syllabus which is studied only by **Higher Level** (HL) students, providing further depth
- two **options topics** that are studied in great depth. There are four options available only to Standard Level students; three options available to both Standard Level and Higher Level students and three options available only to Higher Level Students.

The IB *Physics subject guide*, available from the IBO, provides excellent guidance and includes an extremely detailed syllabus with assessment statements, time allocations and teachers' notes, which are useful to both teacher and students throughout the course. The assessment statements in particular give a clear indication of the material which will be assessed in the final examinations.

These statements can be used as a valuable checklist during examination preparations. Students who cover the syllabus closely and understand it well are likely to be successful – developing a routine of reviewing recent material and asking promptly for help when uncertain is good practice.

The chapters of this book are designed to follow the syllabus very closely and have self-assessment and examination-style questions throughout. Students can expect to be tested periodically by their teacher, using various styles of IB questions to check understanding and gain practice. Past IB exam papers can also be obtained from the IBO.

Developing methodological (practical) skills for success in internal assessment (IA)

Alongside the syllabus content, students are required to develop methodological (practical) skills which are assessed formatively and then summatively at the final assessment (Figure 20.2). This work leads to the internal assessment (IA) component of the course which contributes 24% towards the final grade.

The work will involve a mixture of open-ended and traditional laboratory experiments. These are used to assess students on the IB Group 4 (science) internal assessment criteria. The IA criteria cover skills of:

- designing (planning) an investigation
- collecting, processing and presenting raw data
- concluding and evaluating
- manipulation (carrying out an investigation competently and safely)
- personal skills (perseverance, ethical work, team work and reflection).

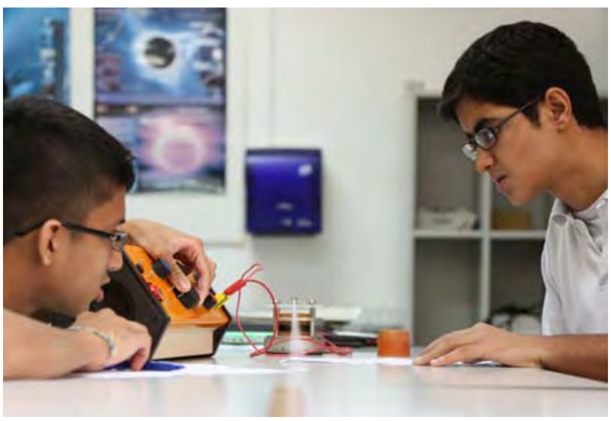


Figure 20.2 IB Diploma physics Higher level students at the United World College of South East Asia engaged in group practical work

Personal skills will only be assessed during the Group 4 project.

Each student is expected to spend a certain amount of time on practical work (regardless of number of practical activities performed). Currently this is 60 hours for Higher Level students and 40 hours for Standard Level students. This time includes all active practical work, but not the time to write the reports. Standard Level and Higher Level students are assessed in exactly the same way and to the same level, the only difference being the amount of time spent. The school's practical Scheme of Work should be designed to spread the practical activities over the main part of the course and should involve the majority of the syllabus topics, including the options.

It is a good idea to start by developing each skill separately using dedicated practical activities. These can be regarded as *practices* but may still contribute to the overall time spent on practical work. In this respect, 'designing an investigation' may not be the best skill to begin with as it is the one that most students find the hardest. The more familiar skills of 'manipulation' and 'data collection and processing' make a better introduction to the use of the assessment criteria. Once the skills have been developed and students are confident, more demanding practical activities can allow students to score well.

Not all skills will be assessed at each practical. One practical activity might be used to assess manipulation, data collection, data processing and presentation (for example, the measurement of *g* from the oscillation of a simple pendulum is an activity that can be given as a prescribed procedure). A different practical activity might be used to assess planning, conclusion and evaluation (for example, in an investigation into the generation of standing sound waves in air columns over water students would be asked to generate their own research question, probably focused on wavelength).

Eventually, by the end of the course, a small number of scores (currently, the best two for each skill) will be used to calculate the summative (final) grade. This implies that students are usually assessed many more times than is minimally necessary. In essence, it is possible to do less well on some early practical activities, but still achieve a good final score provided skills develop during the course.

To drive this improvement, students should:

- be aware of which skills are being assessed during each practical activity
- be given the assessment criteria to work towards
- be given specific feedback on ways to develop skills and get better scores in future.

This is a formative approach, allowing students to receive specific feedback from teachers and learn from their mistakes. To this end, teachers are encouraged to include criteria-specific comments and grade(s) on students' work, both for the students' benefit and to aid the IB moderator.

Examples of such comments might be:

- 'Design skills, Aspect 2 is partially complete – you did not control the temperature of the resistance wire.'
- 'Data collection and processing skills, Aspect 1 is partially complete – you did not include random uncertainties in your width measurements measured by the micrometer screw gauge.'

Students should keep a record of their own grades, set themselves specific targets, and review and track their own progress. However, all evidence of the completion of the required number of hours of practical work (marked practical reports with student instruction sheets attached) must be kept securely in the student's practical portfolio, which may later be required for moderation purposes by the IBO. Teachers may prefer to keep portfolios safely in school. Only those lab reports which are to be remarked are sent away.

Similarly, a record of all internal assessment scores must be kept by the teacher from the very beginning of the course. The IB will require the final internal assessment grades and moderation samples one to two months before the IB examinations begin (by the 20th of April or October); therefore, schools normally complete and internally moderate practical work two or three months before the exams. Further details of the current internal assessment procedures, skills and assessment criteria can be found in the current IB *Physics subject guide* available from the IBO and in the procedures manual for the IB Diploma Coordinator.

The Group 4 (Science) project

Group 4 projects are often carried out midway through the course, after students have developed their methodological skills and become familiar with a range of practical techniques, but before the pressures of finalizing coursework and preparing for external examinations.

A typical organization plan is:

1. Group 4 students from all subjects brainstorm together to select a theme for study (for example, science in sport).
2. The group generates individual research questions to be studied by the separate subject teams comprising biologists, chemists, physicists, students studying Sports, exercise and health science or Design technology).

   These research questions should be interlinked in some way under the overall theme (e.g. 'physiological response to exercise at different altitudes' for the biologists, 'reflective or thermal properties of different sports clothing material' for the physicists and 'determining the concentration of salt in an isotonic sports drink' for the chemists).
3. Subject teams work on their individual investigations over a number of days.
4. Subject teams report findings to the group, emphasizing links to the other teams' studies and often making a group presentation or display (Figure 20.3).

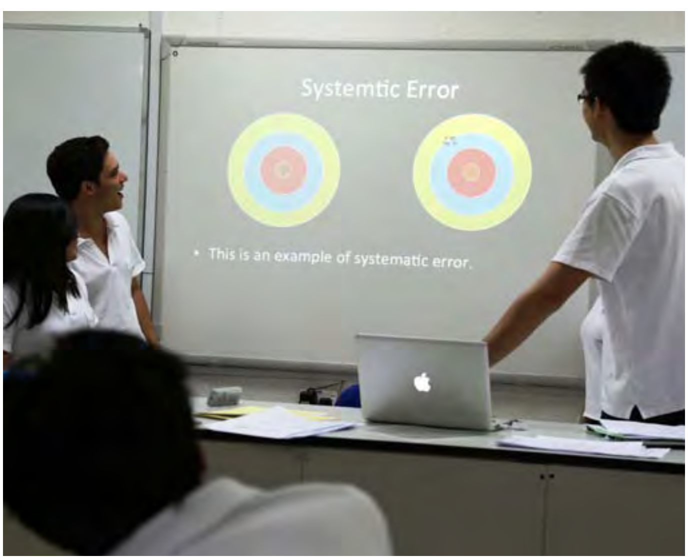


Figure 20.3 First year IB Diploma students at the United World College of South East Asia give a presentation and lead a class discussion on planning a good experiment and writing the report

The extended essay

Each IB Diploma student is required to carry out an individual piece of research in a chosen subject area, such as physics, working under the guidance of a teacher supervisor. This culminates in the production of a 4000-word essay. It is an opportunity to follow an area of personal interest, which might well be related to the student's plans for tertiary education.

Students who choose to write the extended essay in physics must first ensure that the topic is clearly from the physics discipline. For example, although a title such as 'A survey of people's attitudes to the future use of mobile phones' is clearly related to digital technology issues, the overall focus of the essay is more likely to be sociological than from physics. Secondly, to fulfil the criteria for gaining a good grade, the essay should be based on a practical investigation, rather than on a literature survey or internet search. A description of the touch technology used in the iPad® and iPhone®, while eminently worthwhile in itself, would not allow for scoring on the physics components of the extended essay assessment criteria.

The study must start with a well-focused research question or title and a realistic plan generated by the student. It must use either novel techniques of the student's own design or modified standard practical procedures that can be applied to the particular study. Undertaking a simple, well known textbook investigation will not gain a good grade. For example, 'Determining the refractive indexes of a range of transparent plastics' is unlikely to yield the opportunities needed for success. A more suitable research title might be 'Investigating the properties of the zoom lens of an SLR camera'.

Another important (though not physical) element in the final grade is the student's ability to present the essay clearly and in accordance with the conventions for research papers (for example, referencing). Students and teachers should refer to the *Extended essay guide* from the IBO; students should have access to this *before* planning their extended essay.

Many students find their extended essay in physics engaging and rewarding. However, some discover that the nature of practical work can be excessively time consuming unless the research question is chosen very carefully indeed. Often students design and perform a trial run of the project, find that it does not generate suitable raw data, and adapt their approach accordingly. (See Chapter 21 for more information about the extended essay.)

How will the course be assessed?

The final grade awarded for IB Diploma Physics (7 highest, 1 lowest) is based on the internal assessment of practical work (24%) and on the final external examinations (76%). The final examinations currently include multiple-choice questions, data-analysis questions, short-answer questions and paragraph answers:

- Paper 1 is a multiple-choice paper that tests all topics except the options.
- Paper 2 includes data analysis and shorter answer questions (often quite long in reality and requiring in-depth understanding).
- Paper 3 is dedicated to the option topics.

Students who practise past examination questions, and study model answers and mark-schemes become familiar with the format, which contributes to their success.

The extended essay is graded separately, on a scale from A (excellent) to E (elementary), and the mark for this is combined with that from the TOK essay to give the student up to three bonus points.

20.2 To new IB teachers

This section is designed to provide further guidance to teachers who are new to IB Diploma physics. It assumes that the previous section has already been assimilated.

You are probably very keen to obtain all the important details of the syllabus, internal assessment procedures and external examinations, and start planning your course. In this section, we will first look at ways of obtaining the detailed information needed and then at ideas for developing an IB Diploma physics course appropriate for your particular circumstances.

How do I find out more about IB physics now and in the future?

You will quickly come to understand that the IBO is a dynamic organization which continually evolves according to a regular development review cycle. As such, it is important that you are familiar with ways of finding out more now and of keeping informed in the future:

- The first point of contact for subject teachers is the school's IB Coordinator, who will probably act as the sole route of communication with the IBO. He or she will obtain all the necessary documents and will also receive periodic IB *Coordinator notes*, which will include news about developments in Group 4 subjects, including physics.
- The main document to first refer to is the IB *Physics subject guide*. This specifies all current details about the subject.
- To underline their commitment to lifelong learning and continuing professional development, the IBO organizes workshops that run throughout the year. Details of the IBO workshops are found on the ibo.org website.
- The IBO provides excellent online help through its online curriculum centre (OCC) which includes all IB documentation, subject forums and resource-sharing facilities. Access details are obtainable from your IB Coordinator. The OCC includes teacher support material such as sample marked lab reports.
- Schools new to IB often make links with an established IB school nearby. Contacting experienced IB physics teachers can prove to be very useful.

How should we organize and deliver IB physics in our school?

Schools all over the world operate different course-delivery models as dictated by local circumstances, such as size of cohort or responsibility to another course (e.g. a national curriculum). The IB requires that IB Diploma physics students complete 240 hours of study for Higher Level and 150 hours of study for Standard Level. This includes all theory and practical work. How these hours are organized is for each school to decide.

The delivery model

Different schools organize these hours in different ways, varying mainly with regard to the degree of integration (Higher Level students with Standard Level students; theory work with practical) deemed desirable or possible. One popular model is to have Higher Level and Standard Level students timetabled separately over a 2-year course within laboratories so that practical work can be integrated seamlessly with the relevant theory. However, some schools are unable to organize in this way and other delivery models are used:

- Theory is delivered in classrooms with separate, weekly laboratory-based practical sessions.
- Both Higher Level and Standard Level students are timetabled together to cover the core material, with Higher Level students attending extra lessons to cover the AHL material. The options can also be delivered in this way since the majority involve a shared Standard/Higher Level section followed by an additional Higher Level section. This arrangement seems to be the most common one globally.
- The core syllabus, Standard Level option material and 40 hours of internal assessment is covered in year 1 so that Standard Level students sit the examination at the end of the first year. Higher Level students continue into year 2 to cover the AHL material, complete the option topics and finish their internal assessment work. This arrangement is not especially common and requires double the usual number of teaching periods for Standard Level in year 1.

The options

Once the delivery model has been decided on, the next important decision is to select from the option topics offered by the IB (if your model allows for this). Options can be chosen:

- by the physics teacher
- by consensus among students, following which *all* students study the chosen options
- by students individually, who then work independently on their options.

In most schools, the physics teacher decides on the two options to be studied and delivers the material in a similar way to other sections of the syllabus. The more aware students will be quick to point out that, since they are examined on both options in Paper 3, the 'options' are, in fact, not optional at all!

The scheme of work

Once the options have been chosen, the scheme of work and teaching order can be developed. The IB Physics syllabus has *not* been designed as a scheme of work; each physics team will wish to develop a scheme that suits their situation.

There are clear links between the syllabus areas, and a linearity in understanding, which would lead many teachers to start with physics and physical measurement, proceeding to mechanics. It would, however, be a mistake to expose students who are new to IB to an initial topic that is unfamiliar and very conceptual, such as, for example, thermodynamics or forces and fields.

Measurement and data processing, error propagation and graphical techniques should be taught early to allow students to develop the necessary skills for assessed practical work. Many teachers leave atomic physics, nuclear physics and digital technology to the second year of the IB Diploma.

Table 20.1 shows two example sequences in which the chapters of this book might be studied, one for Standard Level students and one for Higher Level students. These are in no way prescriptive – each physics team will wish to pioneer an individual approach, and to modify the sequence in the light of experience. Another common approach is to teach the Standard Level topics and then to teach the Higher Level topics. In this approach the majority topics are taught twice, once at an introductory level and then at a more advanced level.

Table 20.1 Two teaching schemes for IB Diploma physics

A scheme for teaching Standard Level IB Diploma physics	A scheme for teaching Higher Level IB Diploma physics
1 Physics and physical measurement	1 Physics and physical measurement
2 Mechanics	2 Mechanics
5 Electric currents	5 Electric currents
6 Fields and forces	6/9 Fields and forces/Motion in Fields
3 Thermal physics	12 Electromagnetic induction
7 Atomic and nuclear physics	3/10 Thermal physics
8 Energy, power and climate change	7/13 Atomic and nuclear physics/ Quantum physics and nuclear physics
4 Oscillations and waves	14 Digital technology
	8 Energy, power and climate change
	4/11 Oscillations and waves/Wave phenomena
Option E: Astrophysics	Option G: Electromagnetic waves
Option B: Quantum physics and nuclear physics	Option E: Astrophysics

Careers in physics

Students should be informed of the wide range of career paths in the science and engineering fields that require a physics qualification. These include: aerospace, astronomy, astrophysics, atomic and molecular physics, biomedical engineering, computer science, electrical engineering, geophysics, nuclear physics, semiconductors, photonics, ocean science, particle physics, medical physics, nanotechnology and material sciences, to name a selection.

21 Theory of knowledge

21.1 Introduction

In the theory of knowledge (TOK) course you will be asked to analyse and discuss the different **ways of knowing** (perception, reason, language and emotion) and **areas of knowledge** (natural sciences, human sciences, history, the arts, ethics and mathematics). You will compare and contrast the areas of knowledge (AOKs) and ways of knowing (WOKs), identifying links between them. Another important consideration is how different *knowers* have different *perspectives* of the same AOK or WOK.

Throughout this chapter there are a number of quotes for you to reflect on and some questions to research. This chapter also provides an introduction to the TOK concepts and TOK links which have been provided at suitable points in the chapters of the book.

Consider and discuss these questions in class.

1. How is physics different from two other natural sciences, chemistry and biology, and how is it different to the arts and human sciences?
2. What are the roles of creativity, imagination and intuition in physics?
3. Can equations in physics have beauty?
4. Is there any place for emotion (as a WOK) in physics?

21.2 The scientific method

Physics is not a body of unchanging facts, but a process of generating new knowledge, **theories** and **laws**, using the **scientific method**. There is no single agreed 'scientific method' but there are a number of variations, all of which can be used to generate new scientific knowledge. However, the scientific method with which you will be familiar from your practical work (investigations) is known as the **Baconian** or **inductive scientific method**.

> '*Science is nothing but trained and organized common sense.*' **T. H. Huxley**

> '*It seems to us that it would be nearer to the truth to say that science is sharply contrasted with common sense.*' **J. J. Thomson**

> '*Nature never gives up her secrets under torture.*' **Francis Bacon**

Observation

Physics is concerned with formulating mathematical models of physical phenomena. However, before a model can be developed **observations** and measurements (raw data) must be recorded. In physics, observations often arise from questions in the form, 'How does variable *X* affect variable *Y*?'

> '*Science is measurement.*' **Kelvin**

Hypothesis

A **hypothesis** is a scientific explanation of the event or process that caused the physical phenomenon that was observed. It is assumed that the cause precedes the effect. The hypothesis will identify which **variables** are involved in the cause (**independent**) and effect (**dependent**) and which variables are not involved (the **controlled variables**). A hypothesis allows a prediction to be made of what would happen if the independent variable is changed.

Experimental design

5 Find out about the role of Occam's razor when deciding on a hypothesis to test by investigation.

An investigation is designed to collect data that will be used to test a hypothesis. It is important to ensure that your design involves a **fair test** and that there is only one independent variable. If the results of the investigation do not support the hypothesis then a new hypothesis must be developed that takes into account the data (Figure 21.1), assuming that any uncertainties in the experiment had been addressed.

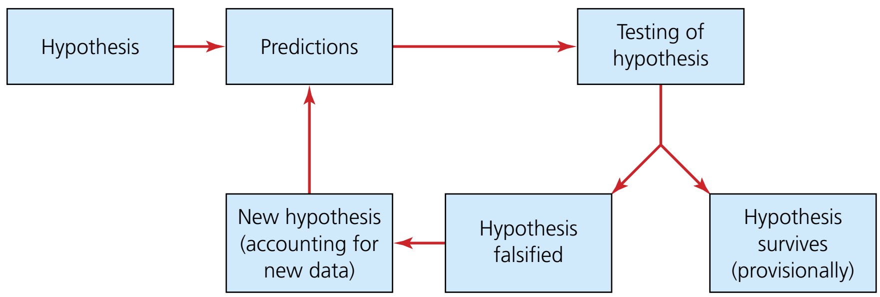


Figure 21.1 The scientific method

You should aim for **reliability** in your experimental results by the control of variables, accurate observation and measurement, representative samples and repeatable investigations.

'Experimentation is criticism.' **Sir Peter Medawar**

21.3 Serendipity

A **serendipitous** discovery is a discovery made by accident when looking for something else. For example, in 1974, Russell Hulse was a graduate student compiling data from the Arecibo Observatory in Puerto Rico. He was trying to detect periodic radio sources that could be interpreted as a pulsar (see Chapter 19). All pulsars known at that time had a stable period with a tendency to increase extremely slowly. One night, Russell noticed a very weak radio wave signal that, had it been slightly fainter, would have passed unnoticed. The period was too short and it was variable. Russell came close to discarding the data during the following weeks, but eventually he interpreted the observation as a binary pulsar. He was awarded the 1993 Nobel Prize in Physics for his work on binary pulsars.

6 Find out about the role of serendipity in the discovery of background microwave radiation from the Big Bang (see Chapter 19) and Röntgen's discovery of X-rays.

'Did you ever observe to whom the accidents happen? Chance favors only the prepared mind.' **Louis Pasteur**

'Thought is only a flash between two long nights but this flash is everything.' **J. H. Poincare**

21.4 Inductive reasoning and science

Inductive reasoning involves forming generalizations from specific examples. Physics uses inductive reasoning – generalizations based on evidence – as the basis of its justification for knowledge. A reliable scientific conclusion will be based on a large number of repeated investigations.

7 Find out how Kepler developed his laws of planetary motion.

However, inductive reasoning can never give *certainty*. We also cannot be sure that the generalizations made in the past will continue to hold into the future. The impossibility of reaching certainty through induction is known as the '**problem of induction**'. Inductive generalizations (theories or laws) may, therefore, be overtaken by new data. A single counter-example falsifies an inductive conclusion.

8 Find out what is meant by the 'Black Swan Effect' and David Hume's views on induction.

21.5 Falsification

The philosopher Karl Popper rejected the idea that science creates new knowledge by inductive steps. He suggested that physicists may work intuitively and creatively to generate a hypothesis before collecting data. This guides the physicist to plan and carry out investigations to collect data to test the hypothesis. The data will then either support the hypothesis or **falsify** the hypothesis. Figure 21.2 shows how a falsification approach would be used to test the hypothesis that water boils at 100°C.

Popper does not claim that after a positive test result we have definite support of the hypothesis. He said that a positive test provided corroboration of the hypothesis in question; but emphasizes that corroboration is simply a report on past performance.

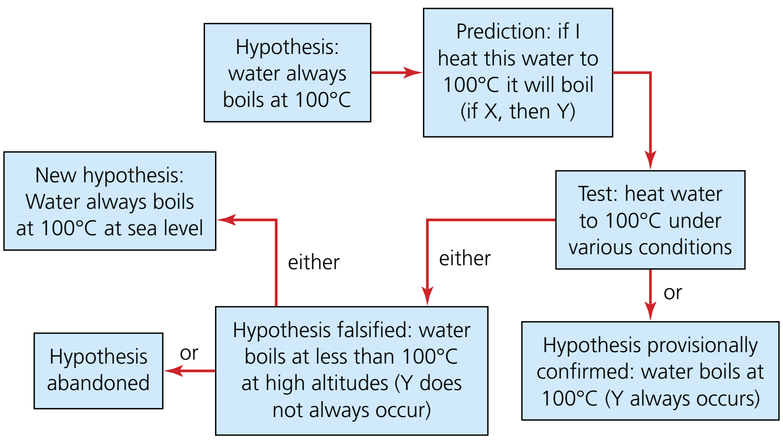


Figure 21.2 Falsification model

Theories and laws in physics can never be **verified** or proved true; scientific knowledge is always provisional. Theories and laws can only be shown to be false (refuted) by experimental results. So, physicists only accept those laws and theories which have been extensively tested and, *so far*, have not been falsified.

According to Karl Popper the difference between science and non-science is that scientific statements are open to being falsified. They must be able to be tested experimentally.

9 Research what is meant by pseudo-science and describe some examples.

10 Describe the caloric theory of heat. How was it falsified?

11 Explain the difference between the terms 'false' and 'falsifiable', and give some examples of statements which are *not* falsifiable.

12 What implications does Popper's falsification approach to science have for time-based ideas, such as the Big Bang (Figure 21.3), where the physicist cannot repeat the experiment in the laboratory?

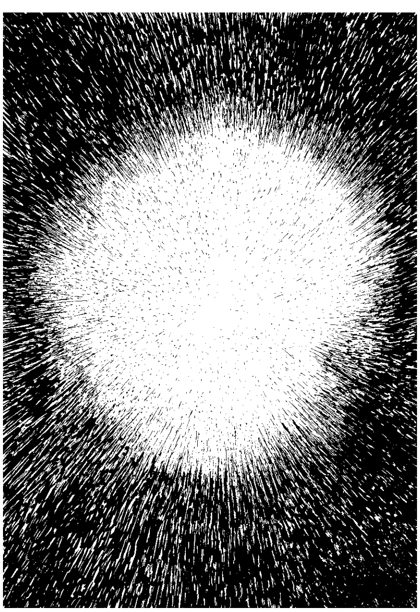


Figure 21.3 An impression of the Big Bang

21.6 Scientific theories

A scientific theory gives the physicist a framework within which to carry out research (investigations) and determines the kinds of questions which are appropriate to ask and test. A new theory in physics replaces an old theory when the new theory seems better supported by current data and when it gives a better explanation of physical phenomena. A new theory may be considered better because it:

- is more effective in giving a framework for problem solving and making predictions
- simply works better and is more useful – this is the '**pragmatic** theory of truth'
- fits better with other existing theories – 'this is the **coherence** theory of truth'
- seems, in terms of evidence, more accurately to reflect what is really there (in an independent external reality) – this is the '**correspondence** theory of truth'.

Thomas Kühn favoured the pragmatic approach and suggested that observations and data collection are made within the framework of an accepted scientific theory or **paradigm** – taught and accepted by the scientific community. He called this '**normal science**'.

During a **scientific revolution**, an accepted theory fails to explain increasingly anomalous data which does not fit in with the predictions of the theory. Revolution occurs when a completely new hypothesis or model is developed which requires that old data and the new anomalous data be interpreted in a totally new way. Kühn called this sudden shift from the old theory to the new theory a **paradigm shift**. It is often met with extreme resistance by the scientific community.

13 Find out what Kühn meant when he said 'paradigms are incommensurable'.

14 Find out about and describe the paradigm shifts that occurred during the following scientific revolutions: geometrical optics to physical optics, Maxwell's theory of electromagnetism, and the development of quantum mechanics which has redefined classical mechanics.

21.7 Comparing science with other areas of knowledge

Table 21.1 shows some comparative questions linking science with the other AOKs and different academic disciplines.

Table 21.1 Comparing science and other disciplines

Science and mathematics	Could physics exist without mathematics?
Science and morality	Does a physicist have a social responsibility? Are physicists too objective about science and too subjective about morality?
Science and religion	Consider the roles of faith and empirical evidence; is science concerned with 'how' and religion 'why'? Why does anything exist at all? Can the scientific view of the world be reconciled with a religious view of the world?
Science and economics	What is the price of scientific knowledge? How can we justify expenditure on pure research in astrophysics and high energy particle physics?
Science and art	Are there differences between creativity in science and creativity in art?
History	How different are the roles and methodology of a historian and a physicist?
Science and ethics	Is it ethical for university research foundations to be funded by tobacco companies or arms manufacturers?

15 Find out about Jacob Bronowski's 'Principle of tolerance'.

21.8 Use of language

To non-scientists, science is a discipline with its own language. Language is used in physics in a very precise way; many IB Diploma physics examination questions test specific definitions. For example, the definition of *quantum efficiency* (for a CCD) is the ratio of the number electrons emitted to the number of photons incident on the pixel. Hence, in the course of studying IB Diploma physics, students must learn the meaning of many scientific words and phrases.
A difficulty may occur when the scientific usage, while related to the everyday meaning, is not identical. For example words such as 'force', 'work', 'energy', 'power', 'ideal' and 'pressure' are used in everyday life but may not be used correctly from a scientific viewpoint. The word 'law', as in Newton's laws of motion (Chapter 2) presents a similar problem. Clearly there are similarities between a scientific law and a legal law, but there are also differences. A legal law may be removed or changed, but this is very unlikely with a well tested physical law, such as the second law of thermodynamics (see Chapter 10).

16 Find out what language is used in particle physics to describe the *flavours* of quarks. Are these terms potentially confusing and misleading?

'When it comes to atoms, language can only be used as in poetry. The poet, too, is not nearly so concerned with describing facts as with creating images.' **Niels Bohr**

21.9 Limits of certainty in physics

The observations and measurements of physics are based on our perception of the physical world. Our perception is shaped by the types of sense organs we have as a species; these allow a limited range of data (from an external reality) to be collected. Our observations become *indirect* when we use instruments to extend our senses. The atomic and subatomic world is beyond our sensory perception.

17 Is it accurate to say that scientific endeavour investigating subatomic particles is beyond our everyday experience of the world? To what extent do these entities exist?

Shortcomings in the method can interfere with the accuracy of results

- Observations are influenced by current theory (current paradigm), both in terms of what the physicist selects for examination and in the interpretation of results.
- Conclusions based on empirical evidence may be falsified, or revised in the face of counter evidence.
- Induction cannot give certainty and Popper's scientific method is characterized by falsifiability.

18 Are all scientific truths provisional?

'We found that the theory did not fit the facts; and we were delighted, because this is how science advances.' **O.R. Frisch**

'Science should leave off making pronouncements; the river of knowledge has too often turned back on itself.' **Sir James Jeans**

The Heisenberg uncertainty principle (see Chapter 13) suggests that physicists must accept ultimate limitations in what we can observe at a subatomic level. This is because the act of measurement actually alters the physical phenomenon being observed. We cannot know the exact position and the exact momentum of a photon or an electron, since the knowledge of one variable can be gained only at the expense of the precision of the other.

The prediction of the future cannot be precise. We cannot have total information and perfect data in the present, and that imprecision gives increasingly unreliable results as we extrapolate into the future. Perfect data of the present would have to take into account every phenomenon; leaving out a single piece of information would distort any prediction.

19 Find out about the concept of *Chaos* and the *Butterfly effect*.

'*True science teaches, above all, to doubt and be ignorant.*' **De Unamuno**

'*Although this may seem a paradox, all exact science is dominated by the idea of approximation.*' **Bertrand Russell**

21.10 Scientific models

When conducting an IB Diploma physics investigation, the first step is to devise a *research question*; the next step is often the generation of a *hypothesis*. It is helpful to think about the underlying **scientific model** that your hypothesis uses.

The term 'model' is potentially misleading since a scientific model is *not* the same as a scale model of, for example, a house or car. A scientific model is a description of an idea that allows scientists to create an explanation of how they think some aspect of the physical world works. There are different types of models including mathematical models, metaphors and teaching models.

An example of a relatively simple *qualitative* teaching model is the nuclear model of the atom (see Chapter 7). The models presented in IB Diploma physics are often *quantitative* and described in the form of equations. For example, the behavior of a simple pendulum can be 'modelled' by the equation, $T = 2\pi\sqrt{l/g}$, where l represents the length of the pendulum, g represents the acceleration of free fall and T represents the period. It should be noted that this equation is a simplification and best describes the behavior at small angles of release.

20 Use a spreadsheet to investigate the relationship between T and l, and T and g for a simple pendulum.

It is important to understand that the scientific model of the atom is not a microscopic version of the description of protons and neutrons surrounded by electrons, as described in Chapter 7. Atoms (Figure 21.4) are not physical structures that can be seen or directly observed. The scientific model is simply the *abstract idea* that subatomic particles exist and show certain behaviours. Such particles are not necessarily real in the way that macroscopic objects that we can see are real. All the models that physicists use to understand natural phenomena are constructions. You should not confuse models with reality – they are useful fictions ('stories').

Following experiments involving electron beams, the idea of electrons as particles (point charges) has been replaced with the idea of wave–particle duality (see Chapter 13). This is the idea that electrons can behave as waves or particles, depending on the type of experiment being performed. We need different models of electrons in order to explain different observations.

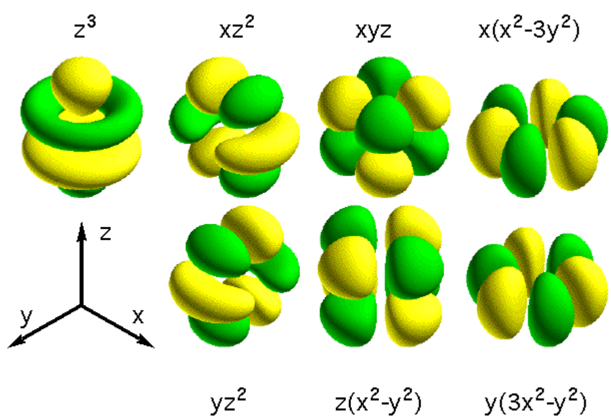


Figure 21.4 Computer generated model of f electron orbitals in an atom

'Those who are not shocked when they first come across quantum theory cannot possibly have understood it.' **Niels Bohr**

One of the most important models used in physics is field theory – the idea that certain particles disturb the space around them so that another particle placed in that space experiences a force (Figure 21.5). Fields are distinguished from one another by the class of particles on which they act (Table 21.1).

Table 21.2 Properties of various fields

Field theory	Source of field	Affected entity
Electrostatics	Charges	Charges
Magnetism	Moving charges	Moving charges
Gravity	Masses	Masses

Field theory was developed by Faraday and Maxwell to explain the influence of one particle on another particle. Hence, fields are only revealed in the motions of these particles. The important property of a field is that it has magnitude and direction at each point in space. Fields are therefore described as vectors (see Chapter 1).

21 Figure 21.6 is a map showing the underground railway lines in Singapore. Is the map 'true'? Is it a useful analogy for the idea of a scientific model?

Figure 21.5 Woolsthorpe Manor, birth place of Isaac Newton, who developed the first scientific model of gravitational attraction

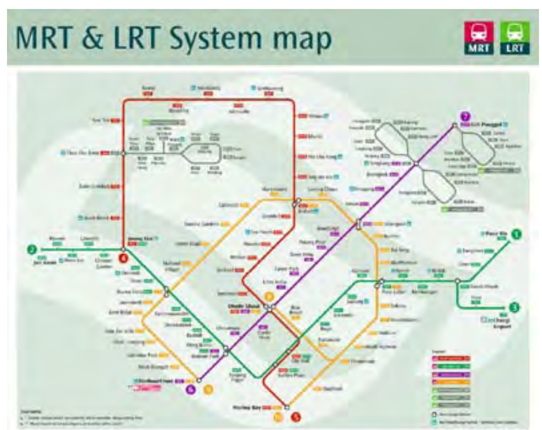


Figure 21.6 Underground and light railway map in Singapore

21.11 Summary

- This chapter has shown that physics uses all four ways of knowing. Physicists use:
  1 sense perception;
  2 enhanced by technology and instrumentation;
  3 combined with reason;
  4 to communicate with other physicists using language and mathematics.
- There are a number of scientific methods, but all rely upon empirical data from observations and experiments where variables are carefully controlled.
- The explanation of the results is usually explained in terms of a theory or mathematical model, often with entities that cannot be directly observed.
- Imagination and creativity play an important role in generating theories and associated models.
- All scientific models have their limitations and are approximate descriptions of physical reality.

22 The extended essay

22.1 Introduction

The IB Diploma's extended essay is a comprehensive study of a specific and focused topic which is intended to promote high level scientific research and scientific writing skills. It will provide you with an opportunity to engage with a personal research topic of your choice, under the guidance of your IB physics teacher. He or she will act as your supervisor and will meet with you a number of times during the course of the extended essay. The completion of the extended essay will be followed by a short interview, known as a *viva voce*, with your supervisor.

The extended essay is a compulsory component of the IB Diploma and takes about 40 hours to complete. The upper word limit is 4000 words, but this excludes the abstract, bibliography, diagrams and tables, etc. The extended essay should be completed within your own time and is not part of the taught curriculum.

Some schools conduct the extended essay in the first year of the diploma; others conduct it in the second year, some across the two. The extended essay is one of the most intellectually challenging parts of the IB Diploma. Full IB Diploma students are usually encouraged to select an extended essay topic from one of their three higher subjects. An extended essay in physics is especially recommended for students who intend studying physics or engineering at college or university level. It can be an excellent point of discussion at an admissions interview.

An extended essay may be entirely based on the physical literature: research papers and reviews. For example, you could write an extended essay about rotating black holes. In this type of extended essay there is no experimental work and, hence, it may be difficult to demonstrate *personal input*. You may be able to find primary data to analyse or, perhaps, be able to simulate some aspect of a rotating black hole on a spreadsheet. However, most schools recommend that you carry out practical work in a laboratory for your own extended essay.

If you do not do your own laboratory work, you must obtain raw data from primary sources (such as scientific journals), analysing it yourself to attempt to address a research question that you have devised. This will most likely involve the use of complex mathematics or analysis with statistics; so you must consider where you will obtain your data and whether your mathematical skills are strong enough.

The majority of extended essays in physics involve some form of practical work. Students in some schools may be involved in collaborative projects with universities where a variety of analytical equipment may be available, for example, atomic force microscope, mass spectrometer, X-ray diffraction apparatus, a nuclear magnetic resonance spectrometer or even a wind tunnel. The use of specialist equipment in 'cutting edge' projects does not necessarily generate high scoring extended essays. Often such equipment can only be handled by skilled technicians and thus becomes a 'black box' for students.

The best extended essays in physics usually involve using simple school-based laboratory equipment, equipment built from bought components or a commercial kit which has to be assembled and perhaps modified. However, you should avoid building equipment that is very complex and takes too long to assemble.

Suitable apparatus could include a homemade ferrofluid, a bubble raft, a parabolic solar heater, a flame probe, a pinhole camera, a radio antenna, a crystal set radio, a phonograph or a vibration detector (using a hacksaw blade, a coil and a magnet). However, before spending time building a particular piece of apparatus think hard about whether it can be used to generate the right data to test a suitable research question.

22.2 The research question

Devising a focused research question is the first step in planning a successful extended essay in physics. The question must be sharply focused, and capable of being addressed and explored within the word limit and time limitation of an extended essay. You may need some help from your supervisor to help formulate a research question from a topic you are interested in. Some suggested dos and don'ts are listed below in Table 22.1.

Table 22.1 Dos and don'ts for Criterion A (research question)

Do	Don't
■ Choose the topic and, if possible, decide on the research question yourself. ■ Clearly write the research question in the early part of the essay. It must also be present in the abstract. ■ Check with your supervisor that your research question is focused, and can be answered within the word count. ■ Make sure that the research question clearly states the variables that are to be manipulated, measured and analysed.	■ Choose a research question that mixes the science subjects or has a strong focus on material science (e.g. focusing on the chemical reactions of a rechargeable battery). ■ Choose a research question that can be easily answered by looking in a physics textbook. ■ Choose a research question that will not allow you to apply IB Diploma physics syllabus theory in a personal way. ■ Choose a research question that goes far beyond the current IB Diploma physics syllabus. ■ Choose a research question at the frontiers of science, for example, dark matter or quantum teleportation of DNA, that may degenerate into science fiction.

Deciding on a topic and establishing a suitably focused research question may be a lengthy process. You will need to read your textbook around the topic and then consult more specialized books, perhaps aimed at college level. You may also want to consult research papers or review papers, perhaps in on-line journals. Your local university or college may be able to give you access to physics literature. You may also need to carry out some relevant background reading if you intend to use theory, such as fluid mechanics, which is outside the current syllabus.

Here are some invaluable sources of ideas or topics for extended essays:

- *Physics Review* (published by Philip Allan; http://www.philipallan.co.uk/physicsreview/)
- *Physics Education* (published by the Institute of Physics; http://iopscience.iop.org/0031-9120/)
- *New Zealand Institute of Physics* (http://nzip.org.nz/index.php)
- *Journal of Chemical Education*
- *School Science Review* (published by the Association of Science Education; http://www.ase.org.uk/)

EBSCO and Science Direct (produced by Elsevier) are on-line databases subscribed to by many International schools. Talk to your school librarian to find out whether your school has access. There are also two CDs available from the IB shop, called '50 Excellent Extended Essays', which contain useful exemplars for teachers and students. Your school library should obtain copies of these.

It is essential that your topic falls clearly within one of the science subjects. If your topic falls between two subjects, such in the case of materials science, you will have to decide which subject you are focusing on. For example, if you choose a material sciences topic and register it as physics, you can only be assessed on the physics content; not on any chemistry which it might contain.

Some topics are unsuitable because the outcome is already well known and documented in textbooks. An example would be the relationship between potential difference applied to a metal wire (at constant temperature) and the current that flows (see Ohm's law, Chapter 5). Your extended essay should **not** just illustrate *known* scientific relationships.

Table 22.2 lists some of the apparatus available in a well equipped school laboratory or university laboratory, with examples of general topics they could be used to investigate (these topic areas would need to be focused more sharply to qualify as extended essay research questions).

Table 22.2

Equipment	Potential area of research
Photoelectric unit (with photocell)	Estimate of Planck's constant
Thyratron	Electron collision experiments
Polarimeter	The effect of temperature on the optical activity of sugars
Flame probe (and electroscope)	Photoelectric effect and electric fields
Light dependent resistor (LDR)	Response of LDR to light of different wavelengths or intensities
Transparent film and crossed polaroid filters	Photoelastic stress analysis
Electric motor	Effect of load on current, speed, power output or efficiency
d.c. generator (dynamo)	Effect of rotation speed and field current on the induced e.m.f.
Strain gauge	Strain in, for example, a model bridge
Thermistor	Effect of temperature on the resistance of NTC and PTC thermistors
Ultraviolet lamp and detector	Absorption of ultraviolet light by, for example, glass or washing powders; viability of UV-sensitive beads
Vibration generator	Powered by the output of a signal generator, can be used to simulate earthquakes
Tesla meter	Measuring the strength of magnetic fields
Gyroscope	Investigating precession and torque
Infrared thermometer	Measuring the surface temperature of an object
Data-logging equipment	Ultrasonic 'ranger' to measure the distance to an object which is oscillating; measuring apparent weight changes in a lift; Faraday's law
Digital laser tachometer	Measuring the rate of rotation of a wheel
Pressure sensor	Measurement of small pressure variations in, for example, soap bubbles
Ticker tape timer and signal generator	Resonance
Consistometer, Redwood viscometer or falling ball viscometer	Measuring viscosity

Most of the apparatus listed in Table 22.2 is not essential for the IB Physics syllabus, but may be available for you to use. The IB *learner profile* encourages 'risk taking' and the moderator may take this into account when grading your extended essay.

Once you have decided on a topic and perhaps an experimental approach, you need to devise a focused research question. It is advisable that your research question clearly states the variables that you wish to study. Actually, the research question does not have to be a question; it could be a statement which outlines what you are trying to find out.

You might be interested in the polarization of light, planning to construct a polarimeter to study polarization in an aqueous solution of molecular solute. A focused research question might be, '*To determine the relationship between the angle of rotation (of sodium light) and temperature of aqueous glucose solution of fixed concentration*'.

You might be interested in why your guitar string breaks more frequently in winter. A focused research question could be '*How does the breaking strength of a treble nylon acoustic guitar string vary with temperature?*'. Your research question may be prompted by demonstrations that you may have seen in class, such as Nitinol memory wire or a demonstration of a syllabus topic, such as Lenz's law, involving dropping a magnet and a non-magnet down a copper pipe and timing them.

As discussed previously, your extended essay should not just attempt to confirm theory that is already documented. It is, of course, necessary to explain the physical theory that forms part of the syllabus of your subject. For example, an extended essay that is focused on investigating a charged-coupled device (CCD) would need some relevant background information showing you understand the principles of how a CCD detects photons. However, a successful extended essay will most likely apply the syllabus theory in a personal way or will explore some theory that is outside the current IB physics syllabus, for example, looking at the dark current (the small current that flows in a CCD when no light is incident).

You should also indicate what problems you encountered while carrying out your experiment – and how you overcame them.

22.3 Assessment of the extended essay

Your extended essay is not just a long internal assessment lab report. In fact, the IB recommends that you do *not* set it out like a normal investigation report (Design, Data Collection and Processing, Concluding and Evaluating). Remember, the Internal Assessment is intended to test your understanding of the IB Diploma physics syllabus theory, whereas the extended essay should apply theory in a personal way, or use theory from outside or beyond the syllabus. However, some aspects of the Internal Assessment, for example the skills involved in Design, do carry over well to the extended essay.

All extended essays, regardless of the subject group, are graded according to the same assessment criteria, although physics and the other subjects have their own particular interpretation of them. Physics extended essays that are written addressing the Group 4 Internal Assessment Criteria are unlikely to score highly, because some issues may not be addressed.

There are 11 separate criteria worth a total of 36 marks. Your extended essay supervisor will give you a copy of the criteria showing the total marks and the descriptors. Table 22.3 is a guide to interpreting criteria B to K. It takes the same approach as Table 22.1, listing dos and don'ts for students.

Table 22.3 A summary of recommendations for extended essay criteria B to K

	Do	Do not
Criterion B: Introduction	■ Set your research question in the context of existing scientific knowledge. ■ Change the research question during the investigation if necessary – this is acceptable. ■ Outline the scientific theory and reasoning that leads to your research question. ■ If you need background information that is not directly scientific, keep it to a minimum and place it in an appendix. ■ State the 'scope' of your extended essay – this means identifying the limits of what you want to find out.	■ Include large amounts of information that is not relevant to your subject (only the scientific knowledge can be credited). ■ Write detailed theory that has come directly from the syllabus – keep it relevant to your research question.
Criterion C: Investigation	■ If your extended essay does not involve laboratory work, you should state how you have obtained and selected the data. ■ Raw data should come from primary sources (e.g. scientific journals), not secondary sources (e.g. textbooks). ■ Discuss the reliability of your data sources. ■ For experimental extended essays, provide enough information about procedures so that another researcher could reproduce the results from your experiment. ■ State whether you designed the experiment or if you obtained or modified the procedure from another source. ■ Evaluate your procedures carefully so that you know the limitations of the data.	■ Use secondary sources (e.g. textbooks, websites) as your only source of data. ■ Give excessively detailed procedures for standard techniques.
Criterion D: Knowledge and Understanding of the Topic Studied	■ Apply the theory in the particular context of your extended essay. ■ Use theory from outside the syllabus (but not too far outside!). ■ Where you have used information from other sources, ensure that it is referenced.	■ Explain the very basic theory that can be found in the IB Diploma physics syllabus.

	Do	Do not
Criterion E: Reasoned Argument	■ Refer back to your research question often to make sure you do not lose focus on what you want to find out. ■ Refer to any model, preferably mathematical, or hypothesis that you have suggested and describe the extent to which it has been supported. ■ Compare different approaches and methods and say how they have helped you draw the conclusion. ■ State any assumptions, e.g. the liquid is assumed to show Newtonian flow.	■ Try to deal with too many variables – it is likely that you will lose the focus of the extended essay. ■ Use phrases such as, 'We can see from the graph that…'. You need to persuade the reader of your conclusions – give data to support your relationship. ■ State personal views or opinions without supporting them with data.
Criterion F: Application of Analytical and Evaluative Skills	■ Calculate the magnitude of random uncertainties in your raw data and propagate them correctly through calculations. ■ Comment on the reliability of the data, and the suitability of equipment used (even if you have used advanced data-logging equipment). ■ Question the validity of assumptions you have used. ■ Use secondary sources or calculations to assess the quality of data.	■ Assume that computer equipment such as a data-logger has no random uncertainties or limitations – the experimental design is still likely to be subject to errors or incorrect assumptions. ■ Take data obtained from the Internet at face value – it may be unreliable. ■ Be afraid to point out weaknesses in your own work – this actually improves the score of your extended essay.
Criterion G: Use of Language Appropriate to the Subject	■ Use correct scientific terminology and symbols throughout. ■ Use mathematical equations and formulae where appropriate. ■ Define and explain technical terms, especially if they are outside the IB Diploma physics syllabus. ■ Refer to Chapters 23 and 24 which deal with drawing and interpreting graphs and internal assessment (IA).	■ Use undefined jargon and acronyms – you are aiming for precision in your writing. ■ Copy passages of text into your essay from other sources. If you do not acknowledge them then you have committed academic dishonesty (i.e. plagiarism). You could lose your chance of being awarded the Diploma.
Criterion H: Conclusion	■ Make sure your conclusion is consistent with your argument and analysis, even if it does not support your hypothesis. ■ If the outcome is unexpected, then state this. ■ Suggest unresolved questions or further work you could do to extend the investigation.	■ Introduce new variables or points to your argument at this stage. ■ Misinterpret your data so that it supports your hypothesis.
Criterion I: Formal Presentation	■ Include a full bibliography (MLA style is recommended). If this is omitted, you score 0 for presentation. ■ Ensure that all your sources are referenced in the text (including diagrams or illustrations). ■ Label diagrams and refer to them in the text. ■ Include a title page, table of contents and page numbers. If you miss out one of these, you score 2 maximum; If you miss out two of them, you score 1. ■ Ensure that your table of contents and your section headings are consistent with each other. ■ Include one example of your calculations, and then summarize processed data in a table. ■ Put large quantities of raw data in an appendix. ■ Processed data that will directly support your argument should be in the main body. ■ Draw graphs or diagrams by hand if you are not able to do it properly on a computer.	■ Exceed 4000 words (graphs, diagrams, calculations, formulas and equations are not included). ■ Include pages of repetitive calculations. ■ Include digital photos that do not show anything useful (e.g. a picture of you doing a straightforward procedure such as recording a potential difference). ■ Allow data tables to spread over more than one page.
Criterion J: Abstract	■ Include three items in your abstract; the research question, the scope and the conclusion. The 'scope' refers to how the research was conducted and how the limits of the research were chosen.	■ Exceed 300 words for the abstract.
Criterion K: Holistic Judgment	■ Devise a novel or innovative research question or approach that is personally interesting to you. ■ Research your extended essay thoroughly, and constantly make sure you are addressing the research question. ■ Display originality or creativity in the way you carry out your research and experimental work.	■ Miss appointments with your supervisor. ■ Expect your supervisor to provide the research question and direct your research. ■ Change your research question without discussing it with your supervisor. ■ Leave your experimental work until the last minute.

22.4 The abstract

Writing an abstract is a new task for many students. Though the abstract is placed at the beginning of the extended essay, it should not be written until the extended essay is finished. One approach to writing your own abstract is to familiarize yourself with some abstracts of physics research papers. An example of an abstract for a physics extended essay is provided below, followed by some critical comments.

A sample student abstract

Heat capacity and thermal conductivity are phenomena that both involve heat transfer. I wanted to establish whether they were correlated and, if so, what relationship is exhibited. The specific heat capacity of a substance is determined by its molecular, ionic or atomic composition and heat transfer via conduction also depends on the composition of the medium. Hence, both phenomena are dependent on the vibration of particles. The factor that controls particle vibration (if the amount of internal energy present is constant) is the magnitude of the forces acting on them.

Research question: *'Does there exist any relationship between the* **specific heat capacity** *and the* **thermal conductivity** *of a liquid?'*

I used an electrical method approach to determine the specific heat capacity of five liquids: distilled water, sodium chloride solutions with concentration of 10 and 50 $g\,dm^{-3}$, palm oil and ethanol. I used a temperature sensor to determine the relative thermal conductivity of each liquid. On comparing both the orders, I observed that the liquid with the lowest heat capacity (palm oil) has the lowest thermal conductivity and the liquid with the highest specific heat capacity (50 $g\,dm^{-3}$ sodium chloride solution) has the highest thermal conductivity. I found that specific heat capacity and thermal conductivity have a directly proportional relationship.

Teacher comments on student abstract

The student could have plotted specific heat capacity on the horizontal axis versus thermal conductivity on the vertical axis. This would help to show more clearly if the two variables are related in a nearly linear fashion or not.

The student should also try to explain why the concentrated salt solution has the highest thermal conductivity. Does the sodium chloride provide mobile ions that carry energy quickly from higher to lower temperatures?

The student could have investigated whether there was a 'mass effect' for the comparisons of specific heat capacity and thermal conductivity.

By extending the work to solids, the student could check their hypothesis for metals and non-metals, which would be interesting – metals are generally good conductors because their conduction electrons transfer energy rapidly. In contrast, non-metals do not share this property.

22.5 Citing sources

As mentioned in Table 22.3, it is very important that you provide a full bibliography of all the sources you used in writing your essay. Many IB schools prefer the use of the MLA (Modern Language Association) referencing style. Examples of references to the different types of source you will use are given in Table 22.4. The University of Northampton has an online guide to the MLA referencing system that you may find useful (available from http://library.northampton.ac.uk/pages/mla).

Table 22.4 Referencing different types of sources

Source	Guide to reference style	Example
Book, single author	Author's name (family name, given name). Title. Place of publication: Publisher, Year of Publication. Titles can be underlined or put into italics.	Duncan, Tom. Advanced Physics (Fifth Edition). London: John Murray, 2000.
Book, two authors	Note order of names for second author	Nelkon, Michael and Philip Parker. Advanced Level Physics (Seventh Edition). Oxford: Heinemann, 1995.
Book, three or more authors		Mee, Chris *et al.* International A/As Level Physics. London: Hodder Education, 2008.
Encyclopedia	'Article title'. Title of Encyclopedia. Year of Publication. Put the title of the article in speech marks.	'High energy particle Physics'. World Book Encyclopedia. 2011.
Interview	Name of the person interviewed (family name, given name). The kind of interview (personal, telephone, email). Date or dates of interview.	Taylor, Bernard. Personal interview (at Department of Physics, National University of Singapore). 8–12 May 2011.
Magazine or journal article	Author (family name, given name). 'Article title'. Magazine or journal title. Date of magazine or journal: Pages. Note the use of speech marks and underlining.	Mussard, P. 'An experiment to determine the factors affecting the frequency of vibrations of a stretched wire'. School Science Review. December 1993: (75) 271, 109.
Internet websites	Author (if available). 'Title of the article.' Title of whole site. Date of visit to site. <URL of page>	'Using MLA Format.' Purdue University Online Writing Lab. 23 January 2006. <http://owl.english.purdue.edu/handouts/research/r_mla.html>
Online encyclopedia article	'Title of article.' Title of Encyclopedia online. Date of visit to site. <URL of the source>	'The use of CCDs in radio-astronomy.' Smith & Wells New World Encyclopedia online. 20 February 2011. <http://www/epnet.com/ehost/login.html>
Online magazine or newspaper article	Author (family name, given name). 'Title of article'. Magazine title. Date of magazine: Page numbers. Product name. Date of visit to site. <URL of the source>	Redden, Stewart. 'Measurement of *h* using an LED'. Physics Review. March 2009: 28–31. MasterFILE Premier on-line. EBSCO Publishing. 30 Feb 2004. <http://www/epnet.com/ehost/login.html>

23 Internal assessment

The overall mark for a student on the IB Diploma physics course (the mark sent to universities or colleges) is based upon two kinds of assessment:

- **external assessment**: performance in written external IB Diploma physics examinations (making up 76% of total mark).
- **internal assessment (IA)**: performance related to laboratory work (making up 24% of total mark).

Internal assessment of laboratory work is based on a *selection* of the *best* practical work that a student completes during the IB Diploma physics course. The marks for the investigations will be summarized by the physics teacher on form 4/PSOW (Group 4 – Experimental Science, Practical Scheme of Work). Higher Level students are required to undertake 60 hours of laboratory investigations over the two years of the course. Standard Level students are required to undertake 40 hours over two years. These durations exclude any time spent writing laboratory reports.

Laboratory work and investigation reports ('write-ups') are assessed or graded using the detailed IB assessment criteria that are listed in this chapter. All IB physics teachers around the world must use the same assessment criteria and apply them in the same way. To ensure that IB physics teachers follow the marking instructions correctly, schools must send samples of assessed student investigation reports to the International Baccalaureate Organization (IBO) for monitoring. Then, if an IB physics teacher is found to be too harsh or too lenient, *all* the marks of that teacher will be adjusted accordingly by the IBO.

IB Diploma physics investigations are graded against five assessment criteria:

- Design (D)
- Data collection and processing (DCP)
- Conclusion and evaluation (CE)
- Manipulative skills (MS)
- Personal skills (PS).

Each criterion is given a maximum mark of six. The first three criteria must be assessed on *two* occasions, producing a maximum mark of $(3 \times 2 \times 6) = 36$. Students will undertake many more than two investigations, but the final marks chosen for their assessment in each criterion will be their *best* two marks. Manipulative skills and personal skills each have a single assessment $(2 \times 6) = 12$. Therefore, internal assessment has a maximum mark of 48.

The assessment of the first three criteria is based largely on the student's written work. Manipulative skills (MS) are assessed summatively (as a whole) over the two year course. Personal skills (PS) are only assessed once, in relation to the Group 4 project. MS and PS assessments are *not* based on written reports from students.

Each of the first four criteria is divided into three parts called *aspects*. When a student's work is graded, a teacher must assess whether each aspect has been met *completely*, *partially*, or *not at all* (c, p or n). This will then determine what total mark is given for that section of the assessment. Up to six marks can be given for each assessment criterion: a 'complete' is worth two marks, a 'partial' is worth one mark, and a 'not at all' is worth zero marks toward that total score.

In the rest of this chapter we will look more closely at each of the assessment criteria. The tables contain the official IB statements, but individual schools may provide their students with more detailed advice.

It is important to realize that the assessment statements are just a guide, and a student's performance may not exactly match any particular description. The teacher must decide which statement most closely describes the student's work. A student's work does *not* need to be faultless to be given a grade of 'complete/2'.

Chapters 1 and 24 also contain much material that is relevant to this chapter.

23.1 Assessment criteria and aspects

Design (D)

Table 23.1 IB statements for the assessment of the Design criterion

	Aspect 1	Aspect 2	Aspect 3
Level/mark	**Defining the problem and selecting variables**	**Controlling variables**	**Developing a method for collection of data**
Complete/2	Formulates a focused problem/ research question and identifies the relevant variables.	Designs a method for the effective control of the variables.	Develops a method that allows for the collection of sufficient relevant data.
Partial/1	Formulates a problem/research question that is incomplete **or** identifies only some relevant variables.	Designs a method that makes some attempt to control the variables.	Develops a method that allows for the collection of insufficient relevant data.
Not at all/0	Does not identify a problem/research question **and** does not identify any relevant variables.	Designs a method that does not control the variables.	Develops a method that does not allow for any relevant data to be collected.

Aspect 1: Defining the problem and selecting variables to investigate

To test a student's ability to design an investigation, the teacher must give an *open-ended* task, such as 'investigate a factor that affects X'. Within this topic there will be many different possible investigations, and each student must decide on a *focused* problem (research question) to investigate, which can usually be expressed in the form of a single sentence that must clearly state the objective of the investigation.

Experiments to determine the value of a specific quantity (for example a density, a resistance or a specific heat capacity) are *never* appropriate as the aim of investigations which are designed to assess the Design criterion, because they do not provide students with sufficient challenges.

In general, students are well-advised *not* to suggest investigations which involve mostly *non-quantitative* observations (for example, investigating different surfaces on which motion is taking place). Such investigations may be interesting, but they do not give students the opportunity to show a full range of skills.

A teacher might set a very wide-ranging task, such as asking students to design an investigation into *any* aspect of forces and dynamics (see Chapter 2). Consider this poor research proposal (which is not expressed in the form of a question) on this topic: 'To investigate the motion of an object'. Will this investigation involve studying displacement, velocity, speed, momentum, kinetic energy or acceleration? What object will be studied? What medium (air, water etc.) will the object be travelling through? The research proposal is too wide and too vague.

Here is a much better proposal, now in the form of a research question: 'What are the factors that affect the acceleration of a cart (on a low friction air track)?' But this investigation can be even more specific. Establishing a relationship between just *two* quantitative variables always makes for the best investigations. One variable (called the *independent variable*) can be varied deliberately and the effect on the other (the *dependent variable*) is measured. We may now confirm a well-focused aim for an investigation: 'What is the relationship between the mass of a cart and its acceleration on a low-friction air track (with a constant resultant force)?'

However, this research question is only a verification of Newton's second law of motion, which is a part of the core IB Diploma physics course, and probably understood by the student. A better research question would not lead to a well known relationship, and would be more open-ended. For example, if a cardboard sail was placed on the cart, perpendicular to the direction of motion, the research question could be: 'What is the relationship between the area of the sail and the acceleration of a cart on a low-friction air track (with a constant mass and constant resultant force)?'

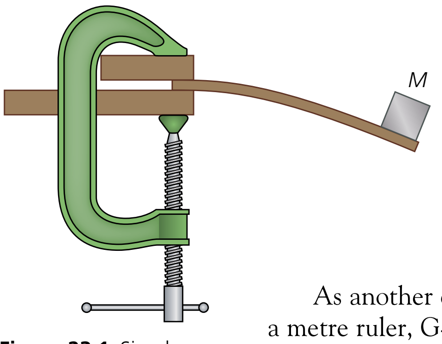


Figure 23.1 Simple cantilever

It is good practice to write some introductory material either before or after the research question. In the last example, it would be helpful to include background information about air resistance and Newton's second law of motion, and then to make a prediction (if possible). For example, the student may think that the acceleration of the cart would be inversely related to the area of the sail. But it is not necessary to know the probable outcome of a research question. In fact, it is much more interesting if the result is unpredictable! It should be noted that background material and predictions are not part of the assessment.

As another example, consider Figure 23.1 which shows a simple cantilever constructed from a metre ruler, G-clamp and a mass, M. (A cantilever is a beam supported at one end.)

Students could be asked to investigate *any* aspect of this arrangement that interests them. Here are some suggestions for good investigations:

- What is the relationship between the frequency of the vibration of a cantilever (of fixed length) and the mass, M, attached to its free end?
- What is the relationship between the sag (deflection) of the free end of a loaded cantilever (of fixed length) and the mass attached to its free end?
- What is the relationship between the sag of the free end of a loaded cantilever and the length of the cantilever? (This investigation is discussed in Chapter 24.)

Identifying other variables

To fully cover this aspect of the design criterion you must clearly identify not only the *independent* and *dependent* variables, but also any other variables which could affect the results if they changed during the investigation. The identification and control of these variables is an important part of the design criterion. It is important to explain *why* the controlled variables are kept constant during the investigation. Quantities that cannot be controlled, such as the acceleration due to gravity or the density of air in a room, are not regarded as variables; these are constants.

Table 23.2 shows one possible detailed format for the presentation of information about variables. It is *not* prescribed by the IBO. The example shown relates to an investigation stretching a rubber band (see Figure 23.2).

Table 23.2 Possible format outlining the identification and classification of variables

Type of variable	Variable	Method for control	Reason for control
Independent	The load (weight) applied to the end of the rubber band		
Dependent	The length of the rubber band		
Controlled	Type of apparatus	The same retort stand and clamp will be used to support the rubber band in the same way throughout.	Different ways of supporting the band may produce different lengths.
Controlled	Type of rubber band	The same type of rubber band with the same dimensions will be used to record all measurements.	Different rubber bands will stretch by different amounts under the same loads.
Controlled	Temperature of surrounding air	Set the air conditioner in the laboratory to a specific temperature.	Temperature changes may affect the way in which the band stretches, although this is unlikely to be significant.
Controlled	Technique for measuring extension	The length of the rubber band will be measured using a ruler and set square each time (as shown in Figure 25.2). The rubber band should be extended slowly by the weight.	Different techniques have different experimental uncertainties.
Controlled	Masses	The same size masses (100 g) will be attached to the spring each time.	To ensure that the data points obtained are evenly spread out.

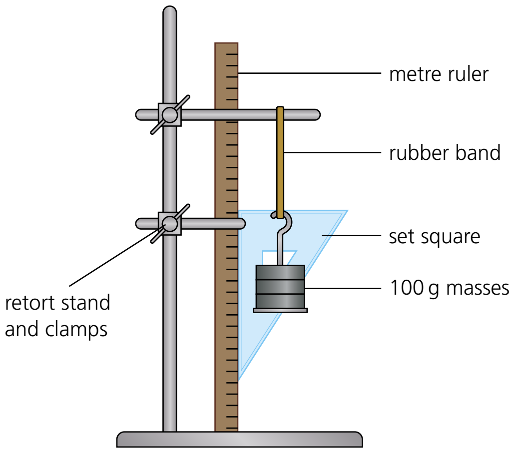


Figure 23.2 Stretching a rubber band

Aspect 2: Controlling the variables

This aspect involves the writing of a detailed experimental *method* with a *labelled diagram* (see Figure 23.2 for an example). The method is not just a list of instructions; it should explain how and why certain actions are performed and it should cover the purpose of the apparatus used. Any procedures designed to improve accuracy by minimizing systematic and random errors in measurements should be clearly described. Anyone reading the method should be able to repeat the investigation from the information provided.

For the proper control of variables the method must contain details about:

- manipulation and measurement of the independent variable
- measurement of the dependent variable
- keeping other variables constant.

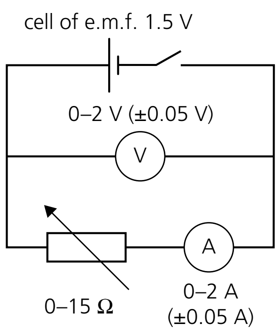


Figure 23.3 Circuit for investigating internal resistance

Consider another example. When designing an experiment to investigate if the internal resistance of a cell varies with the size of the current, it would be necessary to use an ammeter, voltmeter and rheostat. A circuit diagram (see Figure 23.3) should be drawn, using standard symbols and including the ranges and uncertainties of the meters.

It should be explained how the investigation will be carried out. For example, 'With the rheostat set to its maximum resistance, the current and potential difference will be measured. The resistance will then be progressively decreased and new readings taken. About ten pairs of readings will be taken, spread equally over the full range of the rheostat. To ensure that the experiment is as accurate as possible the meters will be checked to ensure they do not have zero offset errors; parallax errors will be avoided when taking readings and the current will be switched off between measurements.'

An indication of how the data will be processed and what graph(s) may be drawn can be given, but this is *not* required in the assessment of design.

Aspect 3: Developing a method for collection of data

The methods described in the previous section must facilitate collection of '*sufficient relevant quantitative*' data. This *usually* means that:

- measurements should be taken with the independent variable having as many different values as possible in the time available. At least *six* data points are needed to draw a reasonable graph.
- any measurements with significant random uncertainties should be repeated at least three times and mean values calculated; however, sometimes (as in the last electricity experiment) it is sensible to take more measurements for different values rather than to repeat the same readings.
- the independent variable should be varied over as wide a range as possible.
- measurements should be equally spaced across the chosen range of the independent variable (occasionally there will be exceptions to this).

The design of the investigation should *specify exactly* what measurements will be taken. For example, it is much better to write '100 g masses will be added to the load on the rubber band, one by one, until a total mass of 1000 g is stretching the band' rather than '10 readings will be taken'. This investigation is a good example of an experiment in which repeating measurements is not sensible. To improve its overall accuracy, 50 g masses could be used (over the same range) instead of 100 g masses.

It is common for an investigation to be modified *after* the plan has been written and an experiment has begun – this is good practice and should be encouraged. For example, it may be realized only during an experiment with a rubber band, that it would be interesting to keep increasing the load until it breaks!

Accurate measurements

It is important to select the correct apparatus for the investigation. Consider the size of the scale divisions on the instrument – they should be much smaller than the quantity you are trying to measure. There are a variety of instruments that allow precise measurements to be recorded in different circumstances. Some of these are shown in Table 23.3.

Table 23.3 Instruments for precise measurement

Instrument	Typical use
Vernier calipers	Lengths and diameters
Micrometer screw gauge	Small diameters and thicknesses
Travelling microscope	Very small distances, depths etc.
Spectrometer	Wavelengths of light

For example, to measure the diameter of a ball-bearing (approximately 1 cm) vernier calipers could be used, but for smaller objects (approximately 1 mm in diameter or less) a micrometer screw gauge would be better.

One useful method is to measure a large number of objects together. For example, if the mass of one ball bearing needs to be known, it would be better to weigh ten (and then divide by ten). The random percentage uncertainty in the measurement of the mass of a single ball bearing is ten times greater than the random percentage error in the measurement of ten (with the same instrument). In general, percentage uncertainties are reduced by making measurements as large as possible.

The most significant sources of error should be identified and the investigation designed to reduce them. For example, when using a manually operated stopwatch to time a single period of a pendulum, the greatest source of error is probably human reaction times when starting and stopping the stopwatch. Uncertainties in the measurement of the amplitude and/or the length of pendulum are probably significantly less. The uncertainty in timing can be reduced by timing 10, 20 or more oscillations, or by using light gates and a data-logger.

Data collection and processing (DCP)

Table 23.4 IB statements for the assessment of the Data collection and processing criterion

	Aspect 1	Aspect 2	Aspect 3
Level/Mark	**Recording raw data**	**Processing raw data**	**Presenting processed data**
Complete/2	Records appropriate quantitative and associated qualitative raw data, including units and uncertainties where relevant.	Processes the quantitative raw data correctly.	Presents processed data appropriately and, where relevant, includes errors and uncertainties.
Partial/1	Records appropriate quantitative and associated qualitative raw data, but with some mistakes or omissions.	Processes quantitative raw data, but with some mistakes and/or omissions.	Presents processed data appropriately, but with some mistakes and/or omissions.
Not at all/0	Does not record any appropriate quantitative raw data **or** raw data is incomprehensible.	No processing of quantitative raw data is carried out **or** major mistakes are made in processing.	Presents processed data inappropriately **or** incomprehensibly.

Aspect 1: Recording raw data

Generally, raw numerical data (the actual measurements made during an investigation) should be recorded in a table, with neat columns and rows. Each student must design their own table of results and this aspect cannot be assessed if the table has been provided by a teacher.

A table of results for recording raw data from an investigation to determine how the time period of oscillation of a vibrating cantilever varies with its length (see Figure 23.1) is shown in Table 23.5. For simplicity, only the first set of results is shown and the three blank columns will be used for processed data – see later.

Table 23.5 Sample results table with raw data for a cantilever investigation

Length of vibrating cantilever, L / ±0.05 cm	**Time for 50 oscillations, t / ±0.2 s**			Columns for processed data		
9.50	25.4	25.3	25.5			

The heading for a column of measurements should include the name of the physical quantity and its symbol (if appropriate), as well as the symbol for the unit of measurement and its uncertainty. Raw data may be recorded in the most convenient units, but data is usually converted into SI units before any calculations are performed.

The independent variable is usually placed in the left-hand column. Processed data should be presented on the right-hand side of the table. The number of decimal places should be the same for all values in a column (unless the range of the measuring instrument was changed during the experiment).

The following simplified rules are often used to determine the random uncertainty in a single measurement from a scale or digital measuring device. Random uncertainties are nearly always recorded as *one* significant digit. Reasons for the choice of uncertainty should be given in the report.

- The random uncertainty when using an analogue instrument is usually equal to the smallest scale division divided by two. For example, for a ruler that measures to the nearest millimetre, the random uncertainty is ± 0.5 mm.
- The random uncertainty when using a digital meter is usually equal to the smallest increment shown on the display. For example, for an electronic balance with the mass reading 5.7513 kg, the random uncertainty is 0.0001 kg.
- In some experiments, significant uncertainty may be introduced by the nature of the techniques used, rather than the equipment itself. For example, uncertainties could be due to human reaction times when using a manual stopwatch, or to the difficulty of judging the position of a fast-moving object. In such examples, the uncertainty should be estimated from experience and the spread of measurements made. In Table 23.5 the stopwatches probably recorded the time to within ±0.01 s, but the results have been given an uncertainty of ±0.2 s to allow for the reaction times of the user. For this reason the measurements have only been recorded to one decimal place.

When giving any experimental measurement, the uncertainty should usually be stated clearly. However, if it is not, an uncertainty is still suggested by the last significant digit of the data. Typically, a measurement of 34.7, for example, suggests an uncertainty of ±0.05 or ±0.1, whereas a measurement of 347 suggests an uncertainty of ±0.5 or ±1.

All measurements made in an investigation should be recorded. For example, if the difference between two levels (for example, of a liquid in a U-tube) is required, *both* levels should be recorded, not just the difference between them – this is processed data, not raw data. Any measurements that would be needed to repeat an experiment should be recorded, even if they were kept constant. For example, in an investigation into the height of bounces of a ball, the original drop height should be recorded.

Any unusual (anomalous) measurements should be repeated. Good experimental technique sometimes involves responding to unexpected results by modifying the original plan. Occasionally, it may become clear only after starting an experiment that the original plan was not sensible; then the experiment has to be completely redesigned!

Relevant qualitative data (observations) should also be noted, for example, 'the resistance wire became hot during the investigation'. This information may be useful when evaluating the method and considering the control of variables.

If raw data comes from a data logger, then a print-out of the results could be included in the investigation report.

Aspect 2: Processing raw data

'Processing raw data' usually means making calculations from measurements. This process also involves determining the uncertainties in calculated results. Wherever possible, processed data and raw data should be presented in the same table, although sometimes it may be helpful to provide an explanation of the calculation separately (as below).

The first step in processing data is often calculating the mean values of some measurements, with their uncertainties. Consider the raw data first shown in Table 23.5 and repeated in Table 23.6 below.

Table 23.6 Processing the results for a cantilever investigation

Length of vibrating cantilever, L / ±0.05 cm	Time for 50 oscillations, t / ± 0.2 s			Mean time for 50 oscillations, t / ± 0.2 s	Time for one oscillation $\left(=\frac{t}{50}\right)$ T / ±0.004 s	T^2 / s^2
9.50	25.4	25.3	25.5	25.4	0.508	0.258 ±0.004

- The uncertainty in the time for one oscillation is the uncertainty in the time for 50 oscillations divided by 50.
- Values of T^2 are calculated because the student thinks that T^2 may be proportional to the length, L.
- The percentage uncertainty in T (for the value shown) is 0.787%, so that the percentage uncertainty in $T^2 = 2 \times 0.787 = 1.575\%$, which means that the absolute uncertainty in $T^2 = 0.01575 \times 0.258 = 0.004$ (rounded sensibly to one significant digit). This uncertainty needs to be calculated separately for each result.

The principles for uncertainty calculations are provided in Chapter 1, including a copy of an example computer spreadsheet. Spreadsheets are often the best way of processing the many similar calculations required.

As another more detailed example of a calculation of uncertainty, consider the determination of the maximum kinetic energy of a trolley under the following circumstances: a trolley of mass 0.68 ± 0.01 kg accelerated uniformly from rest and moved 1.05 ± 0.01 m in 3.2 ± 0.1 s. Note that the calculations below are only rounded off to a sensible two significant figures *at the end*.

- The percentage uncertainties are: mass ± 1.471%, distance ± 0.952%, time ± 3.125%
- Final speed, v, = average speed × 2 = (1.05/3.2) × 2 = 0.6563 m s^{-1}
- The percentage uncertainty in the final speed is the sum of the percentage uncertainties in distance and time = 0.952 + 3.125 = 4.077%
- Kinetic energy $= \frac{1}{2}mv^2 = \frac{1}{2} \times 0.68 \times 0.6563^2 = 0.1464$ J
- The percentage uncertainty in the kinetic energy is the sum of the percentage uncertainties in mass and speed (× 2) = 1.471 + 4.077 + 4.077 = 9.625%.
- 9.625% of 0.1464 is 0.01409
- So, the maximum kinetic energy of the trolley can be quoted to a sensible number of significant figures as 0.15 ± 0.01 J.

Aspect 3: Presenting processed data

'Presenting processed data' usually means drawing one or more graphs. Drawing and interpreting all aspects of graphs are very important skills in physics. Students are recommended to consult Chapter 24 for a detailed examination of graphical techniques, as well as Chapter 1 which contains an explanation of the use of *error bars*.

Concluding and evaluating (CE)

Table 23.7 IB statements for the assessment of the Conclusion and evaluation criterion

	Aspect 1	**Aspect 2**	**Aspect 3**
Level/mark	**Concluding**	**Evaluating procedure(s)**	**Improving the investigation**
Complete/2	States a conclusion, with justification, based on a reasonable interpretation of the data.	Evaluates weaknesses and limitations.	Suggests realistic improvements in respect of identified weaknesses and limitations.
Partial/1	States a conclusion based on a reasonable interpretation of the data.	Identifies some weaknesses and limitations, but the evaluation is weak or missing.	Suggests only superficial improvements.
Not at all/0	States no conclusion **or** the conclusion is based on an unreasonable interpretation of the data.	Identifies irrelevant weaknesses and limitations.	Suggests unrealistic improvements.

Aspect 1: Concluding

Any conclusion(s) should be related to the original research question (the aim of the investigation) and include a clear explanation of *how* the conclusion was reached from the results.

For example, an investigation into the relationship between the pressure and volume of a gas (at constant temperature) produced the graph of raw data shown in Figure 23.4a. Because an inverse relationship was predicted, the same data was processed to produce the graph shown in Figure 23.4b. Note that the error bars in the second graph are *not* all the same size: there are greater uncertainties for larger values of 1/volume, even though the uncertainties in volume are constant.

Because Figure 23.4b is a straight line which can be extended through the origin, it may be concluded that, under these circumstances, the pressure of the gas was inversely proportional to its volume. It is possible to reach the same conclusion by numerically processing the data to show that (pressure × volume) = constant, but the graphical method is better because the accuracy of the conclusion is easier to assess.

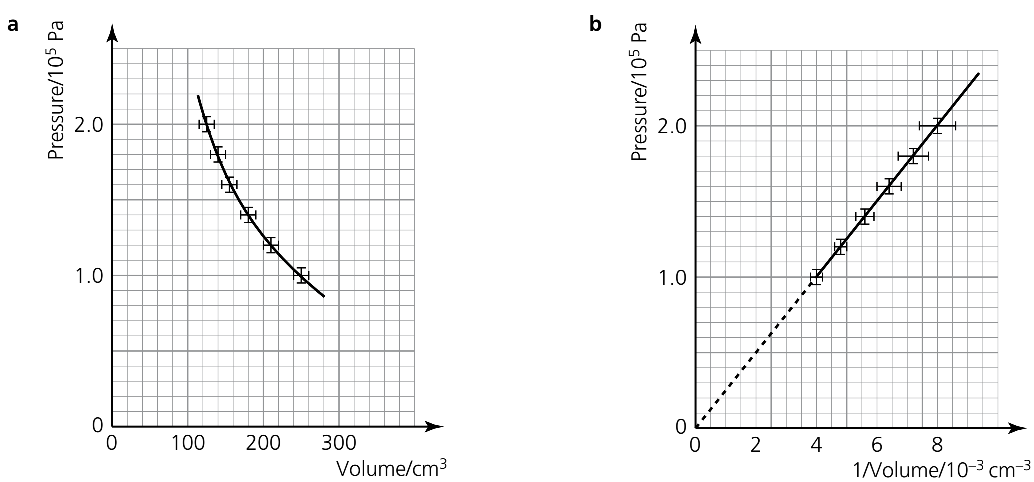


Figure 23.4 The relationship between pressure and volume of a gas

If the conclusion to an investigation is a numerical value of a quantity, it should be stated with its correct unit, to a degree of accuracy consistent with the uncertainty of the various measurements on which it is based. It should also be compared with any accepted value that can be obtained from other sources. For example, if the speed of sound in air was found by experiment to be $329 \pm 5\,m\,s^{-1}$ at 298 K and 101 kPa, it could then be compared to a value of $331.0\,m\,s^{-1}$ found on the Internet and assumed to be correct (source should be quoted). The measured speed is consistent with the accepted value, because the correct value of $331\,m\,s^{-1}$ lies within the experimental range of $324–334\,m\,s^{-1}$. However, if the measured speed was found to be $329 \pm 1\,m\,s^{-1}$ then it would be inconsistent with the correct value, even though the value is the same as in the previous example. Any such inconsistency would suggest the presence of experimental error or, more likely, the underestimation of uncertainties in measurement.

Aspect 2: Evaluating procedure(s)

When the results of an investigation are represented in graphical form, it is possible to learn a lot about the quality of the experimentation from the graph(s) by asking the following questions:

- How many data points are there?
- Can a smooth line of best fit (curved or straight) be drawn through all the error bars?
- Is the position and shape of the best-fit line obvious, or are there alternative possibilities and conclusions?
- Are the data points close to (or on) the line of best fit?
- Are the data points spread out evenly along the line?
- Are the error bars (showing random uncertainties) small or large?
- Are the uncertainties the same for all measurements, or are they greater for certain values?
- Did the collected data cover a wide enough range of the chosen variables?

An evaluation could begin by focusing on the questions in this list. If:

(i) there are six (or preferably more) data points evenly spread over a wide range;
(ii) the error bars are small; and
(iii) a smooth line of best fit can be drawn which passes through all the error bars and close to the points, then it is probable that the investigation was successful and reasonably accurate.

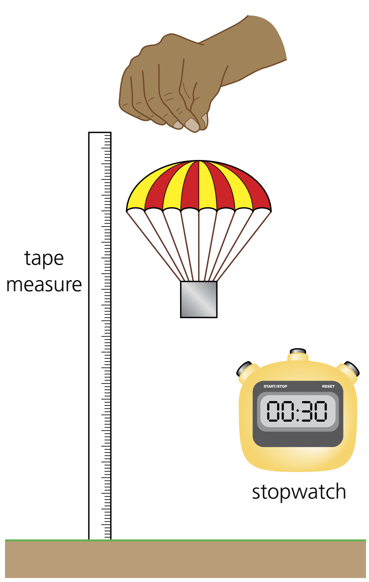


Figure 23.5 Investigating the time of fall of a toy parachute

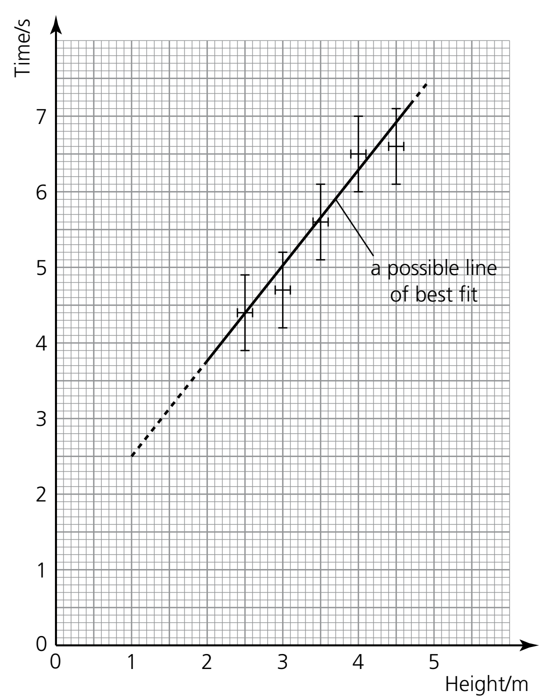


Figure 23.6 Time of fall–height fallen graph for a parachute

It is also possible that a much more challenging and interesting investigation could produce results which did few of these things, but which was still done very well by the student.

Consider Figures 23.5 and Figure 23.6, which show an investigation into a falling toy parachute and a time–height graph of the results.

An evaluation of this graph would conclude that:

- the parachute should have been dropped from more than five different heights.
- although a straight line of best fit has be drawn through the error bars, there are significant uncertainties in the measurements of time. This means that the line of best fit cannot be drawn with any certainty and it may not be straight. For example, a smooth *curve* could also be drawn through all the error bars and the origin. A definite conclusion *cannot* be made from this graph.
- times for smaller heights need to be measured, so that the shape of the graph closer to the origin can be determined. It would also have been interesting to check greater heights.
- an extended line of best fit does not pass through the origin, which means that, either the pattern of results for small heights would not be linear and/or there is a systematic error.

Further evaluation points

The specific techniques and apparatus used during the investigation should be discussed with respect to their precision and accuracy. For example, in the parachute experiment, the large uncertainties associated with using a hand-operated stopwatch should be discussed, as well as the problems associated with releasing the parachute smoothly from a known height. Good experimental techniques in experiments like this (with significant random errors) involve repeating the same measurements, and the number of repeats and the range of values produced should be commented on.

It is possible, through no fault of the student, that the apparatus and/or techniques available were not sufficiently precise for reliable results. This would be described as a *limitation* of the investigation. A simple example would be if the parachute investigation was done in a single storey building, but the student wanted to try greater heights. The same experiment could also be limited by the use of a stopwatch that only measured to the nearest 0.1 s (rather than 0.01 s).

There may be significant variables that were not controlled or monitored during an investigation. For example, the parachute may have been dropped outdoors in an unpredictable air flow due to the wind.

In thermal physics experiments, the flow of thermal energy out of (or into) apparatus from the surroundings is always a possible source of significant error. In mechanics experiments, friction and/or air resistance will usually introduce errors.

An unexpected intercept on a graph should be commented on. It may be because of a zero offset error, or another regular systematic error.

Aspect 3: Improving the investigation

Suggested improvements must be related to the weaknesses or limitations mentioned in the evaluation (see Aspect 2). Suggestions should focus on *specific* pieces of equipment or *specific* experimental techniques. Vague statements like 'take more readings', 'use data gatherers' or 'use more accurate equipment' deserve no credit unless they are more specific and linked to a clearly identified weakness or limitation of the investigation. Suggestions also need to be realistic. For example, carrying out a mechanics experiment in a vacuum or a heat experiment at −100°C is not practicable in a school laboratory. The following is a list of some *specific* suggestions for the parachute experiment.

- Take readings for 10 different heights starting from 0.5 m, increasing in equal steps to 5.0 m (or more).
- Two people would probably be needed to do this experiment accurately. They would need to develop a precise way of coordinating the release of the parachute and the start of the timing (give details).
- Repeat readings more often (perhaps five times each) because of the large random uncertainties in the timings. (Note that the use of automatic timers or data gatherers may not be realistic for this experiment.)
- Design a way of releasing the parachute in exactly the same way each time (give details).
- Make sure that the parachute is released in a location that is free from air currents.

Manipulative skills (MS)

Table 23.8 IB statements for the assessment of the Manipulative skills criterion

	Aspect 1	**Aspect 2**	**Aspect 3**
Level/mark	**Following instructions**	**Carrying out techniques**	**Working safely**
Complete/2	Follows instructions accurately, adapting to new circumstances (seeking assistance when required).	Competent and methodical in the use of a range of techniques and equipment.	Pays attention to safety issues.
Partial/1	Follows instructions but requires assistance.	Usually competent and methodical in the use of a range of techniques and equipment.	Usually pays attention to safety issues.
Not at all/0	Rarely follows instructions **or** requires constant supervision.	Rarely competent and methodical in the use of a range of techniques and equipment.	Rarely pays attention to safety issues.

Manipulative skills are assessed summatively in the IB Diploma physics course. This means that the physics teacher will base the marks for each student on observations made throughout the experimental programme, although the final mark awarded will reflect the levels of performance maintained by the *end* of the course.

Personal skills (PS)

Table 23.9 IB statements for the assessment of the Personal skills criterion

	Aspect 1	**Aspect 2**	**Aspect 3**
Level/mark	**Self motivation and perseverance**	**Working within a team**	**Self-reflection**
Complete/2	Approaches the project with self-motivation and follows it through to completion.	Collaborates and communicates in a group situation and integrates the views of others.	Shows a thorough awareness of their own strengths and weaknesses and gives thoughtful consideration to their learning experience.
Partial/1	Completes the project but sometimes lacks self-motivation.	Exchanges some views but requires guidance to collaborate with others.	Shows limited awareness of their own strengths and weaknesses and gives some consideration to their learning experience.
Not at all/0	Lacks perseverance and motivation.	Makes little or no attempt to collaborate in a group situation.	Shows no awareness of their own strengths and weaknesses and gives no consideration to their learning experience.

Students will only be assessed for personal skills once, which will be during the Group 4 project. Students will not be assessed against any of the other criteria at the same time.

The form of assessment will vary from school to school, but some teachers involve peer assessment and/or self assessment. This means that students are asked for their opinions regarding the participation of other students and on their own contributions to the project. In such reviews, students tend to be very open and honest, but ultimately their teacher will make the final decisions when awarding marks.

During the Group 4 project it is important that students listen to and respect the opinions of other students in their group, and cooperate and work as a team, taking an active role in all activities.

24 Graphs and data analysis

24.1 Representing data graphically

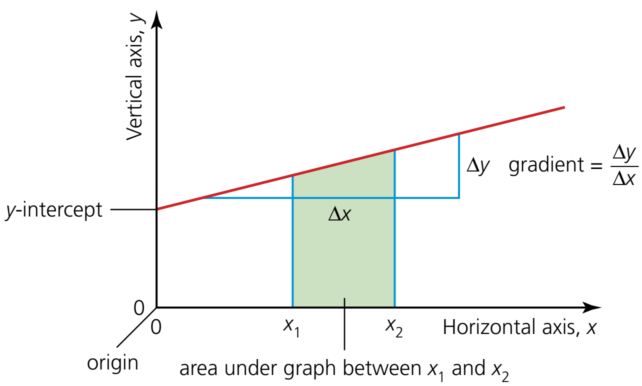


Figure 24.1 Terminology for graphs

There are many quantities that can be measured in a physics experiment. Usually, all but two of them are **controlled** so that they do not change during the course of the experiment. Then one quantity (called the **independent variable**) is deliberately varied or changed and the effect on one other quantity (called the **dependent variable**) investigated.

The best way to analyse the results of such experiments is often to plot (draw) a graph. Looking at a graph is a good way to identify a pattern, or trend, in numerical data. Graphs can also provide extra information: **gradients** (slopes), **intercepts** and the **areas** under graphs often have important meanings. Figure 24.1 illustrates the terminology associated with graphs.

Drawing graphs

The drawing of good quality graphs is a very important skill in physics. The following points should be remembered when plotting a graph.

- The larger a graph is, the more precisely the points can be plotted. A simple rule is that the graph should occupy at least half of the available space (in each direction).
- Each **axis** should be labelled with the quantity and the unit used (e.g. force/N, speed/m s^{-1}). For example, if you wish to record a mass of 5 g as the number five on the axes of a graph, then labelling the axes as mass/g indicates that you have divided 5 g by g to get 5.
- The *independent variable* is usually plotted on the horizontal axis and the *dependent variable* on the vertical axis. Sometimes, the choice of what to plot on each axis is made so that the gradient of the graph has a particular meaning. If time is one of the variables, it is nearly always plotted on the horizontal axis.
- The **scales** chosen should make plotting the points and interpreting the graph easy. For example, five divisions might be used to represent 10 or 20, but not 7 or 12.
- Usually, both scales should start at zero, so that the point (0, 0), the **origin**, is included on the graph. This is often important when interpreting the graph. However, this is not always sensible, especially if it would mean that all the readings were restricted to a small part of the graph. Temperature scales in °C do not usually need to start at zero.
- Data **points** should be neat and small. If points are used (rather than crosses), drawing a small circle around them can make sure that they are not overlooked, especially if the line goes through them.
- The more points that can be plotted, the more precisely the line representing the relationship can be drawn. At *least* six points are usually needed, although this may not always be possible.
- Once all the points have been plotted, a pattern will usually be clear and a **line of best fit** can be drawn (Figure 24.2 shows points represented by error bars, with two correct examples). These lines are sometimes called *trend lines.* Trend lines may be straight (drawn with a ruler) or a curve. (A straight-line graph is described as being **linear**.) The line should be smooth and thin. Bumpy lines that try to pass through, or near, all the points show that the person drawing the line did not understand that points cannot be perfectly placed, and that there is uncertainty in all measurements. Typically there will be about as many points above a line of best fit as there are below the line. Points on a graph should *never* be joined by a series of straight lines.

- Drawing graphs by hand is a skill that all students should practise. However, knowing how to use a computer program to plot graphs is also a very valuable and time-saving skill (especially for investigational work). A graph generated by a computer (or graphic display calculator) must be judged by the same standards as a hand-drawn graph and sometimes their best-fit lines are not well placed.

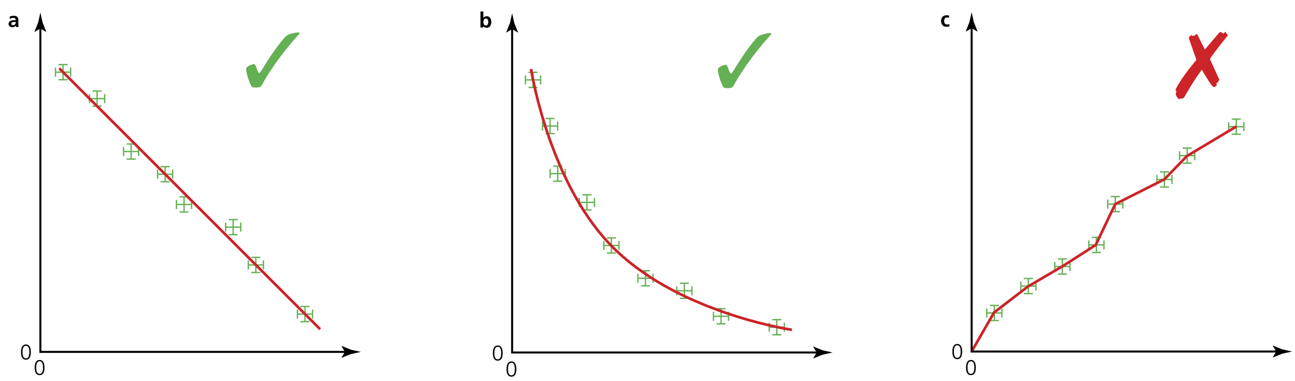


Figure 24.2 Right and wrong ways to draw best-fit lines

Extrapolating and interpolating

A line of best fit is usually drawn to cover a specific *range* of measurements recorded in an experiment, as shown in Figure 24.3. If we want to predict other values *within* that range, it can be done with confidence. The figure indicates how a value for y can be determined for a chosen value of x. This is called **interpolation**.

If we want to predict what would happen *outside* the range of measurements (**extrapolation**) we need to *extend* the line of best fit. Lines are often extrapolated to see if they pass through the origin, or to find an intercept, as shown in Figure 24.3.

Predictions made by extrapolation should be treated with care, because it may be wrong to assume that the behaviour seen within the range of measurements also applies outside that range.

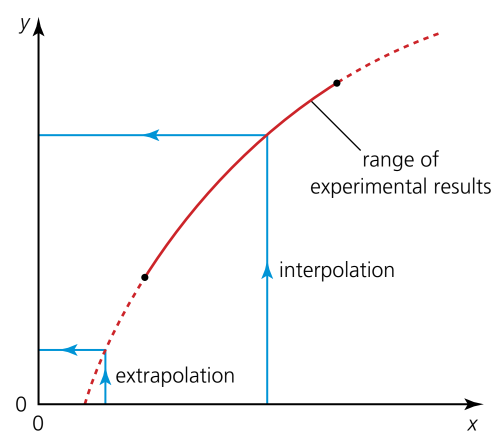


Figure 24.3 Interpolating and extrapolating to find the intercept on the x-axis

Proportionality

The simplest possible relationship between two variables is that they are **proportional** to each other (sometimes called *directly* proportional). This means that if one variable, x, doubles, then the other variable, y, also doubles; if y is divided by five, then x is divided by five; if x is multiplied by 17, then y is multiplied by 17, etc. In other words, the ratio of the two variables (x/y or y/x) is constant. The symbol for proportionality is as follows:

$y \propto x$ means that y and x are proportional to each other.

Many basic experiments are aimed at investigating if there is a proportional relationship between two variables, and this is usually best checked by drawing a graph.

If two variables are (directly) proportional, then their graph will be a *straight line passing through the origin* (see Figure 24.4). It is important to stress that a linear graph that does not pass through the origin *does not* represent proportionality (see Figure 24.5).

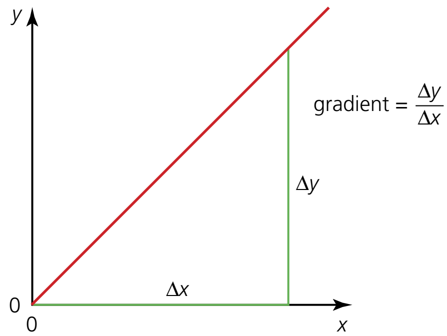


Figure 24.4 A proportional relationship

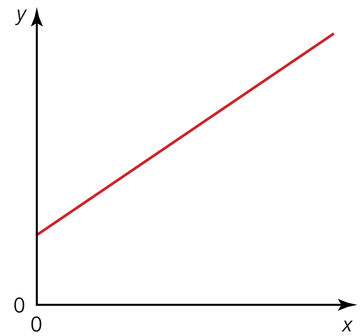


Figure 24.5 A linear relationship that does not pass through the origin is *not* proportional

Gradients of lines

The gradient of a line is given the symbol m and it is calculated by dividing a change in y, Δy, by the corresponding change in x, Δx, as shown in Figure 24.4. (A delta sign, Δ, is used to represent a change of a quantity.)

$$m = \frac{\Delta y}{\Delta x}$$

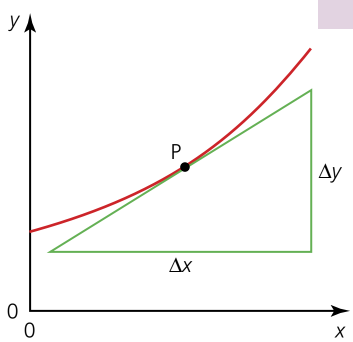


Figure 24.6 Finding the gradient of a curve at point P with a tangent

It is important to note that a *large* triangle should be used when determining the gradient of a line, because the percentage uncertainty will be less when using larger values.

The gradients of many lines have a physical meaning, for example the gradient of a mass–volume graph equals the density of the material.

The gradient of a curved line, such as that shown in Figure 24.6, is constantly changing. The gradient of the curve at any point can be determined from a **tangent** drawn to the curve at that point. (In mathematics, if the equation of a line is known, then the gradient at any point can be determined by a process called *differentiation*.)

Worked examples

1 Figure 24.7 shows a best-fit line produced from an experiment in which the masses and volumes of different pieces of the same metal alloy were measured.

a Calculate a value for the density of the alloy, which is equal to the gradient of the line.

b Suggest why the graph does not pass through the origin.

c Explain why using the gradient to find the density is a much better method than just calculating a value from one pair of readings of mass and volume.

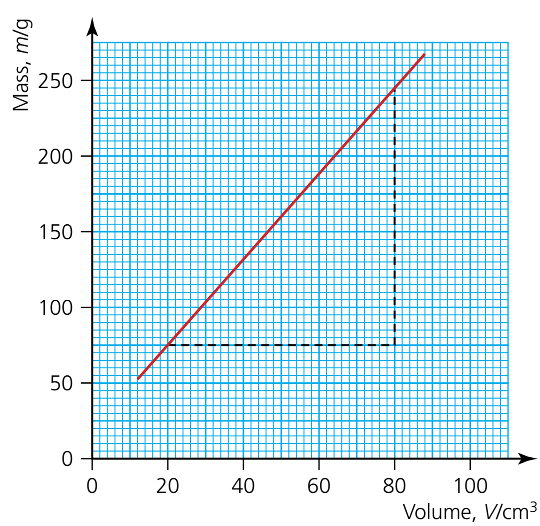


Figure 24.7 Graph of mass m against volume V of different pieces of alloy

a Using the triangle shown on the graph:

$$\text{gradient} = \text{density} = \frac{\Delta m}{\Delta V} = \frac{245 - 75}{80 - 20}$$
$= 2.8\,\text{g}\,\text{cm}^{-3}$

b The instrument used to measure mass had a zero offset error (of about +20 g).

c Individual readings may be inaccurate. The best-fit line reduces the effect of random errors and the zero offset error does not affect the result of the calculation.

2 Figure 24.8 shows a distance–time graph for an accelerating car. Determine the speed of the car after 4 s (equal to the gradient of the line at that time).

The dotted line is a tangent to the curve at $t = 4$ s.

gradient = speed

$$= \frac{\Delta s}{\Delta t}$$

$$= \frac{20 - 0}{5.0 - 2.0} = \frac{20}{3.0}$$

$$= 6.7\,\mathrm{m\,s^{-1}}$$

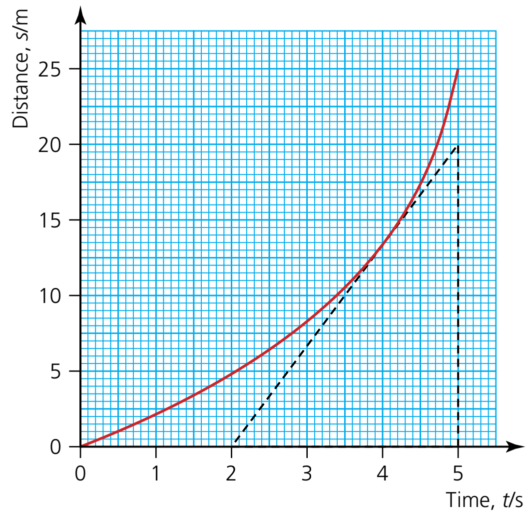


Figure 24.8 Graph of distance s against time t for an accelerating car

Areas under graphs

The areas under many graphs have physical meanings. As an example, consider Figure 24.9a, which shows part of a speed–time graph for a vehicle moving with constant acceleration. The area under the graph (the shaded area) can be calculated from the average speed, given by $\frac{(v_1 + v_2)}{2}$, multiplied by the time, Δt. The area under the graph is therefore equal to the distance travelled in time Δt.

In Figure 24.9b a vehicle is moving with a changing (decreasing) acceleration, so that the graph is curved, but the same rule applies: the area under the graph (shaded) represents the distance travelled in time Δt.

The area in Figure 24.9b can be estimated in a number of different ways, for example by counting small squares, or by drawing a rectangle which appears (as judged by eye) to have the same area. (If the equation of the line is known, it can be calculated using the process of *integration*.)

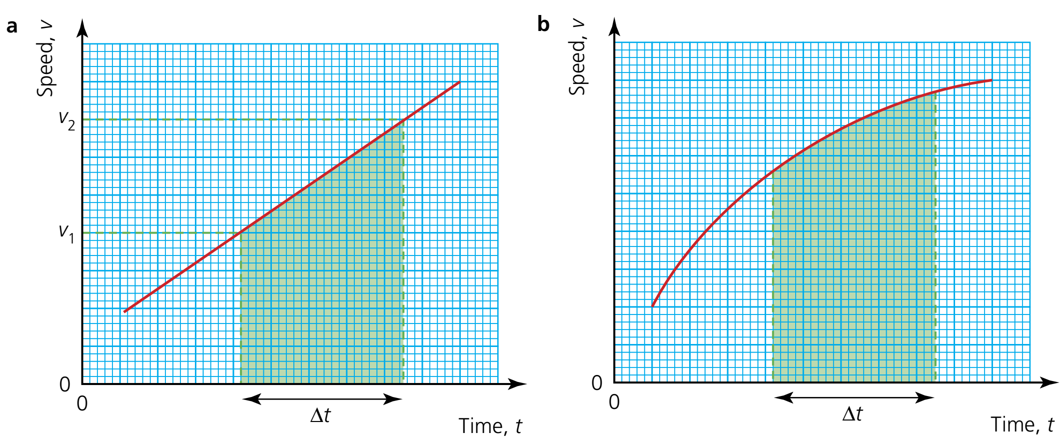


Figure 24.9 Area under a speed–time graph for **a** constant acceleration and **b** changing acceleration

Worked example

3 Figure 24.10 represents the motion of a train that travels at a constant speed for 30 s and then decelerates for 60 s. Calculate the distance travelled in 90 s (equal to the area under the graph).

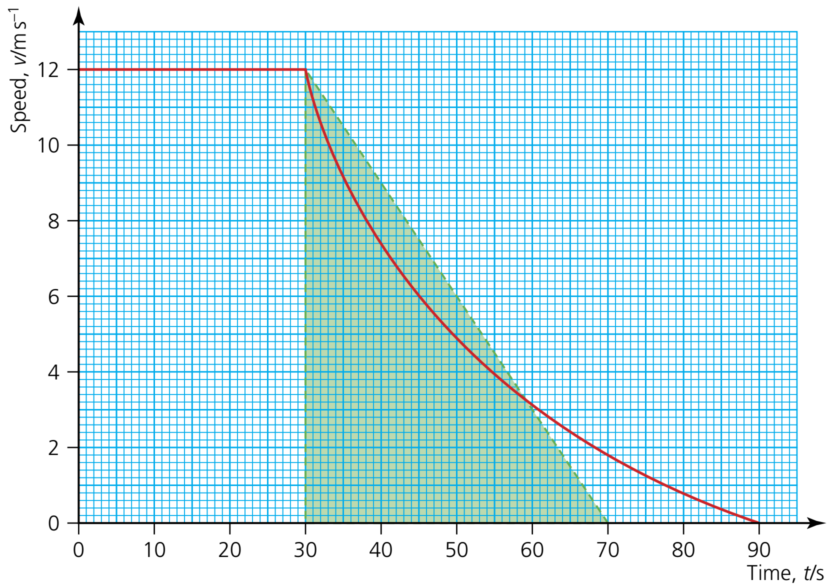


Figure 24.10 Graph of velocity v against time t for a train

The area under the graph up to a time of 30 s = 12 × 30 = 360 m
The area under the graph between 30 s and 90 s can be estimated from the shaded triangle, which has been drawn so that its area appears to be the same as the area under the curved line:

$$\text{area} = \frac{1}{2} \times 12 \times (70 - 30) = 240\,\text{m}$$

total distance (area) = 360 + 240 = 600 m

24.2 The usefulness of straight-line graphs

Straight lines are much easier to understand and analyse than curved lines, but when directly measured experimental data are plotted against each other (x and y, for example), the lines are often curved, rather than linear.

Data that give an x–y curve can be used to draw other graphs to check different possible relationships. For example:

- a graph of y against x^2 could be drawn to see if a straight line through the origin is obtained, which would confirm that y was proportional to x^2
- a graph of y against $1/x$ that passed through the origin would confirm that y was proportional to $1/x$ (in which case x and y are said to be **inversely proportional** to each other)
- a graph of y against $1/x^2$ passing through the origin would represent an **inverse square relationship**.

Figure 24.11 (overleaf) shows graphs of the most common relationships.

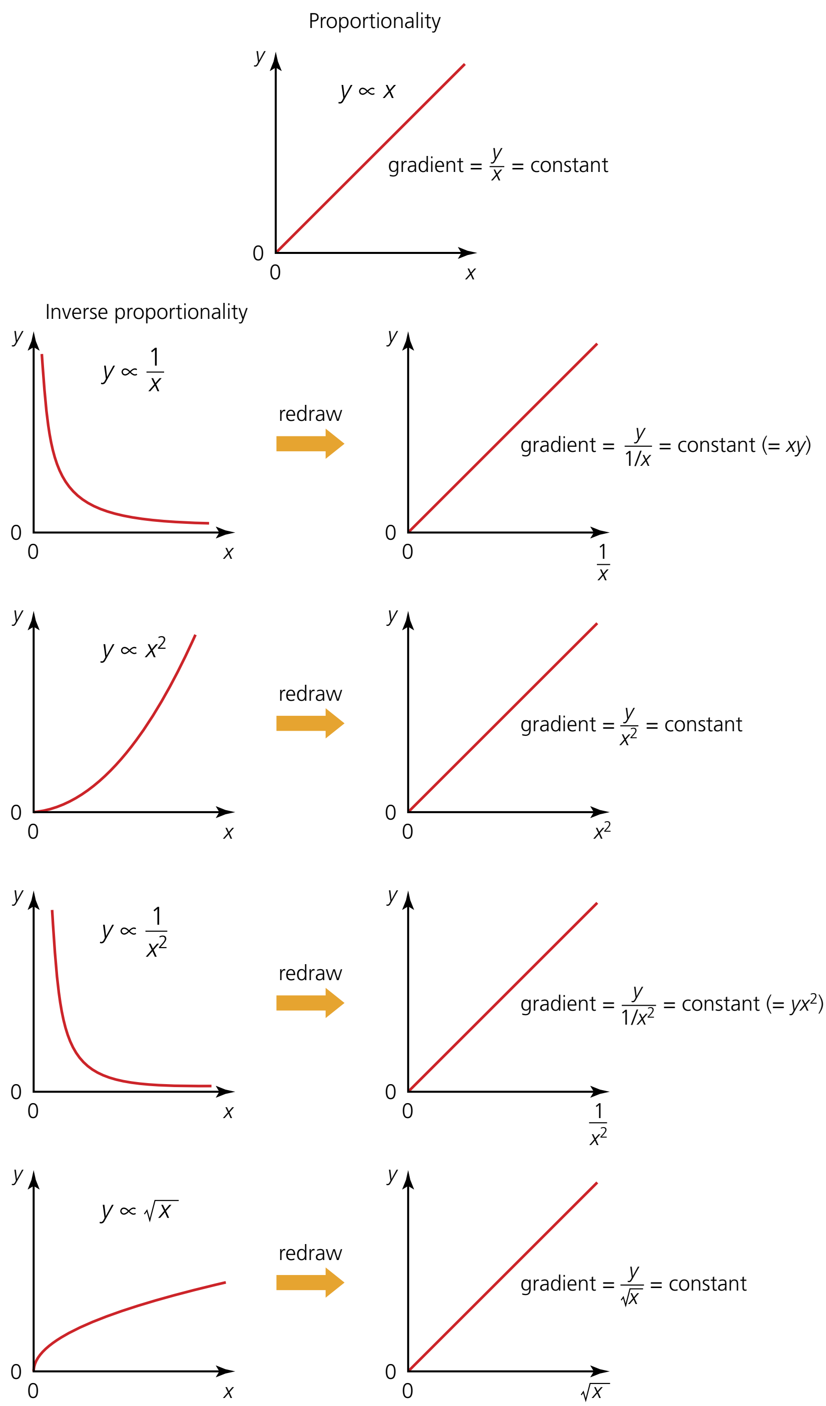


Figure 24.11 Some common graphical relationships, showing how curves can be re-plotted to produce straight lines

Worked example

4 In an internal assessment students were asked to investigate one factor that affected the deflection of a wooden beam fixed only at one end (a cantilever), and which had a mass (load) hanging from somewhere on the part that extended from the bench top (see Figure 24.12).

A student listed the following variables: (i) type of wood, (ii) thickness of wood, (iii) width of wood, (iv) length of wood from where it was fixed, L, (v) position of load, (vi) mass of load. He decided to investigate how the deflection, D, depended on the length L, keeping all the other variables constant. His results are shown in Table 24.1 (for simplicity, uncertainties have not been included).

a A graph of the raw data produces a curved line, so the student thought that maybe the deflection was proportional to the length squared or the length cubed. Perform numerical checks on the data to see if either of these possibilities is correct.
b Plot a suitable graph to confirm the correct relationship.

Table 24.1

L/cm	*D*/cm
30.0	0.8
40.0	2.0
50.0	3.7
60.0	6.5
70.0	10.8
80.0	15.8
90.0	22.2
100.0	31.2

a If the deflection, D, is proportional to the length, L squared, L^2/D (or D/L^2) will be constant, within the limits of experimental uncertainties. Calculations produce the following results (all $\times 10^2$ cm): 11, 8.0, 6.8, 5.5, 4.5, 4.1, 3.6, 3.2.These values are getting smaller for longer lengths, and are clearly *not* constant.

If the deflection, D, is proportional to the length, L cubed, L^3/D (or D/L^3) will be constant, within the limits of experimental uncertainties. Calculations produce the following results (all $\times 10^4$ cm^2): 3.4, 3.2, 3.4, 3.3, 3.2, 3.2, 3.3, 3.2. These values are all very similar (within 3% of their average), confirming that the deflection is proportional to the length cubed.

b See Figure 24.13. A graph of D against L^3 produces a straight line through the origin. Note that it would have been better if the student had used lengths such that the points were spread out evenly along the line.

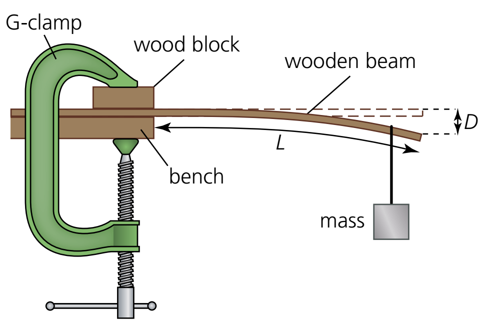


Figure 24.12

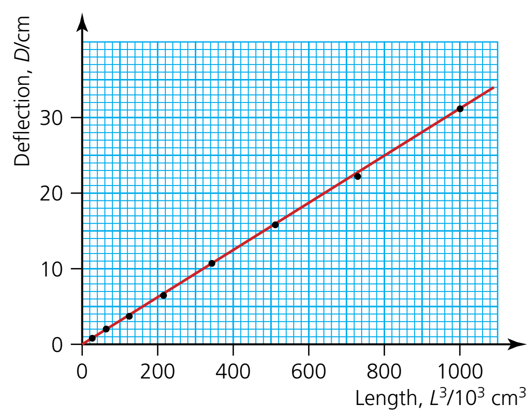


Figure 24.13

Equation of a straight line

All linear graphs can be represented by an equation of the form:

$$y = mx + c$$

where m is the gradient and c is the value of y when $x = 0$, known as the ***y*-intercept** (Figure 24.14).

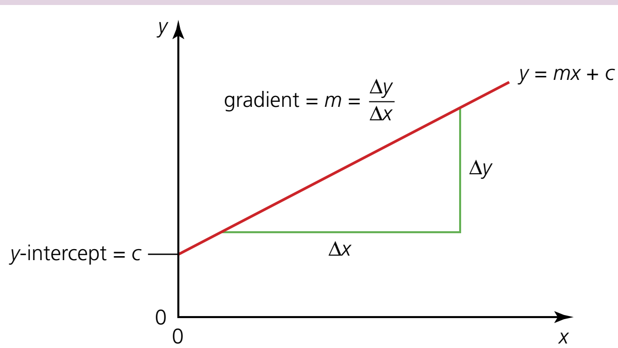


Figure 24.14 Graph of $y = mx + c$

Once a linear graph has been drawn, values for the gradient and the y-intercept can be determined and the results used to produce a mathematical equation to describe the relationship.

Worked example

5 Experimental data connecting two variables, x and y, is represented by the graph in Figure 24.15 (for simplicity units have been ignored). Take measurements off the graph to enable you to write an equation to represent the relationship.

The gradient of the line, m, is $\frac{(0.2-1.2)}{(30-5)} = -0.04$ and the y-intercept, c, is 1.4.

Substituting into $y = mx + c$ we get:

$y = -0.04x + 1.4$

(which could be rewritten as $25y = 35 - x$)

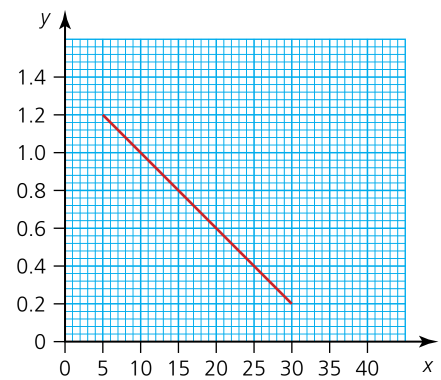


Figure 24.15

Power laws and logarithmic graphs

Sometimes there is no 'simple' relationship between two variables, or we may have no idea what the relationship may be. So, in general, we may suggest that the variables x and y are connected by a relationship of the form:

$$y = kx^p$$

where k and p are constants. That is, y is proportional to x to the power p.

Taking logarithms of this equation we get:

$$\log y = p \log x + \log k$$

Compare this to the equation for a straight line, $y = mx + c$.

If a graph is drawn of $\log y$ against $\log x$, it will have a gradient p and an intercept of $\log k$. Using this information, a mathematical equation can be written to describe the relationship. Note that logarithms to the base 10 (log or lg) have been used in the above equation, but **natural logarithms (ln)** could be used instead (Higher level only).

Worked examples

6 The relationship between two variables, x and y, is shown in Figure 24.16. Take measurements off the graph so that you can write an equation to represent the relationship.

The gradient of the line, p, is 2.9 and the intercept on the $\log y$ axis, $\log k$, is –1.2. Substituting into $\log y = p \log x + \log k$ we get:

$\log y = 2.9 \times \log x - 1.2$

So, $y = 0.063x^{2.9}$

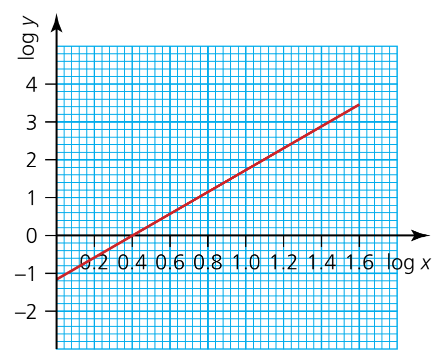


Figure 24.16

7 Refer back to Worked example **4**.
Use the data to draw a log graph which verifies that the relationship is described by the equation $D = kL^m$. Determine values for m and k from the graph.

The graph is a straight line (Figure 24.17), confirming the form of the relationship.
Taking logs of the equation, we get

$$\log D = m \log L + \log k$$

Comparing this to $y = mx + c$, we know that a graph of $\log D$ against $\log L$ will have a gradient of m, and k can be determined from an intercept.

$m = 3.1$ and $k = 0.27$.

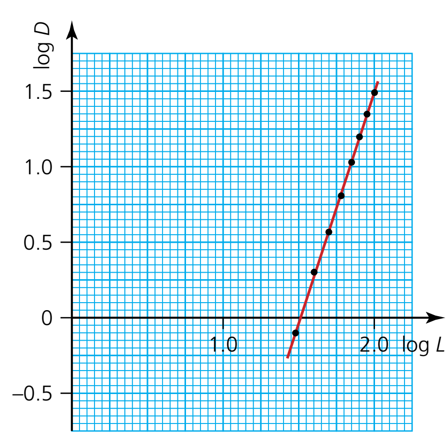
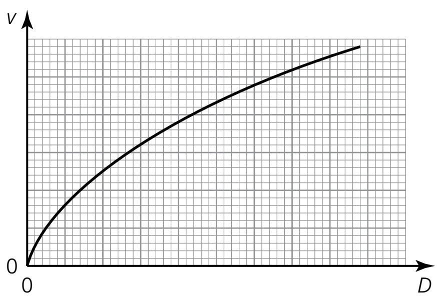


Figure 24.17

Examination questions – a selection

1 As part of a road-safety campaign, the braking distances of a car were measured. A driver in a particular car was instructed to travel along a straight road at a constant speed v. A signal was given to the driver to stop and he applied the brakes to bring the car to rest in as short a distance as possible. The total distance D travelled by the car after the signal was given was measured for corresponding values of v. A sketch graph of the results is shown here.



a State why the sketch graph suggests that D and v are not related by an expression of the form

$D = mv + c$,

where m and c are constants. [1]

b It is suggested that D and v may be related by an expression of the form

$D = av + bv^2$,

where a and b are constants.

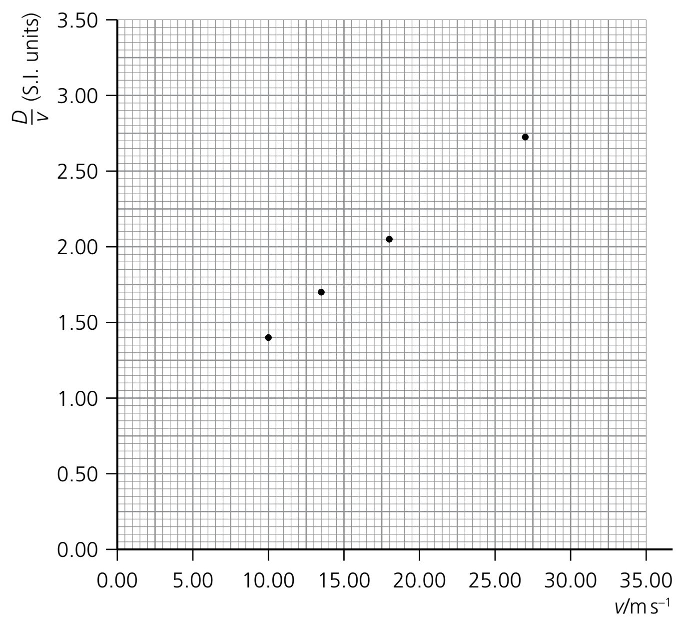
In order to test this suggestion, the data shown below are used. The uncertainties in the measurements of D and v are not shown.

v/m s^{-1}	D/m	$\frac{D}{v}$
10.0	14.0	1.40
13.5	22.7	1.68
18.0	36.9	2.05
22.5	52.9	
27.0	74.0	2.74
31.5	97.7	3.10

i State the unit of $\frac{D}{v}$. [1]

ii Calculate the magnitude of $\frac{D}{v}$, to an appropriate number of significant digits, for v = 22.5 m s^{-1}. [1]

c Data from the table are used to plot a graph for $\frac{D}{v}$ (y-axis) against v (x-axis). Some of the data points are plotted on a graph as shown.



i On a copy of the graph, plot the data points for speeds corresponding to 22.5 m s^{-1} and to 31.5 m s^{-1}. [2]

ii Draw the best-fit line for all the data points. [1]

d Use your graph from **c** to determine

i the total stopping distance D for a speed of 35 m s^{-1}. [2]

ii the intercept on the $\frac{D}{v}$ axis [1]

iii the gradient of the best-fit line. [2]

e Using your answers to **d ii** and **d iii**, deduce the equation for D in terms of v. [1]

f The uncertainty in the measurement of the distance D is ±0.3 m and the uncertainty in the measurement of the speed v is ±0.5 m s^{-1}.

i For the data point corresponding to v = 27.0 m s^{-1}, calculate the absolute uncertainty in the value of $\frac{D}{v}$. [2]

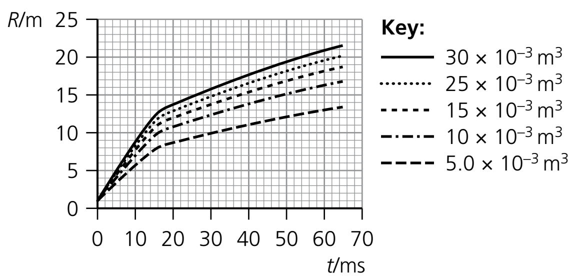
ii Each of the data points in **b** was obtained by taking the average of several values of D for each value of v. Suggest what effect, if any, the taking of averages will have on the uncertainties in the data points. [2]

Standard Level and **Higher** Level Paper 2, Nov 07, Q1

2 The question is about investigating a fireball caused by an explosion.

When a fire burns within a confined space, the fire can sometimes spread very rapidly in the form of a circular fireball. Knowing the speed with which these fireballs can spread is of great importance to fire-fighters. In order to be able to predict this speed, a series of controlled experiments was carried out in which a known amount of petroleum (petrol) stored in a can was ignited.

The radius R of the resulting fireball produced by the explosion of some petrol in a can was measured as a function of time t. The results of the experiment for five different volumes of petroleum are shown on the graph. (Uncertainties in the data are not shown.)



a The original hypothesis was that, for a given volume of petrol, the radius R of the fireball would be directly proportional to the time t after the explosion. State two reasons why the plotted data do not support this hypothesis. [2]

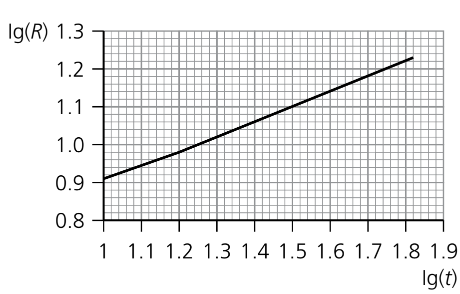
b The uncertainty in the radius is ±0.5 m. The addition of error bars to the data points would show that there is in fact a systematic error in the plotted data. Suggest one reason for this systematic error. [2]

c (HL only) Since the data do not support direct proportionality between the radius R of the fireball and time t, a relation of the form

$R = kt^n$

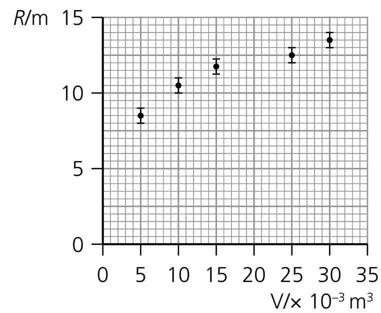
is proposed, where k and n are constants.

In order to find the value of k and n, lg(R) is plotted against lg(t). The resulting graph, for a particular volume of petrol, is shown. (Uncertainties in the data are not shown.)



Use this graph to deduce that the radius R is proportional to $t^{0.4}$. Explain your reasoning. [4]

d It is known that the energy released in the explosion is proportional to the initial volume of petrol. A hypothesis made by the experimenters is that, at a given time, the radius of the fireball is proportional to the energy E released by the explosion. In order to test this hypothesis, the radius R of the fireball 20 ms after the explosion was plotted against the initial volume V of petrol causing the fireball. The resulting graph is shown.



The uncertainties in R have been included. The uncertainty in the volume of petrol is negligible.

i Describe how the data for the above graph are obtained from the graph in **a**. [1]

ii Make a copy of the graph and draw the line of best-fit for the data points. [2]

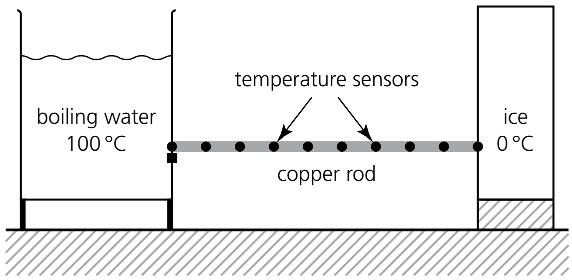
iii Explain whether the plotted data together with the error bars support the hypothesis that R is proportional to V. [2]

e Analysis shows that the relation between the radius R, energy E released and time t is in fact given by $R^5 = Et^2$.
Use data from the graph in **d** to deduce that the energy liberated by the combustion of 1.0×10^{-3} m³ of petrol is about 30 MJ. [4]

Standard Level and **Higher** Level Paper 2, May 07 TZ2, Q1

3 This question is about thermal energy transfer through a rod.

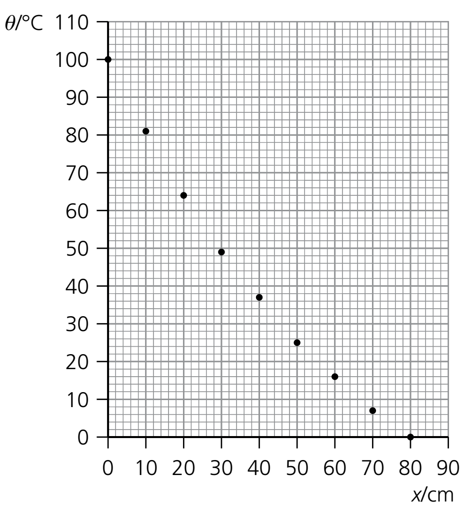
A student designed an experiment to investigate the variation of temperature along a copper rod when each end is kept at a different temperature. In the experiment, one end of the rod is placed in a container of boiling water at 100 °C and the other end is placed in contact with a block of ice at 0.0 °C as shown in the diagram (on the next page).



not to scale

Temperature sensors are placed at 10 cm intervals along the rod. The final steady state temperature θ of each sensor is recorded, together with the corresponding distance x of each sensor from the hot end of the rod.

The data points are shown plotted on the axes below.



The uncertainty in the measurement of θ is ±2 °C. The uncertainty in the measurement of x is negligible.

a Make a copy of the graph and draw the uncertainty in the data points for x = 10 cm, x = 40 cm and x = 70 cm. [2]

b Draw the line of best-fit for the data. [1]

c Explain, by reference to the uncertainties you have indicated, the shape of the line you have drawn. [2]

d **i** Use your graph to estimate the temperature of the rod at x = 55 cm. [1]

ii Determine the magnitude of the gradient of the line (the temperature gradient) at x = 50 cm. [3]

e The rate of transfer of thermal energy R through the cross-sectional area of the rod is proportional to the temperature gradient $\frac{\Delta\theta}{\Delta x}$ along the rod. At x = 10 cm, R = 43 W and the magnitude of the temperature gradient is $\frac{\Delta\theta}{\Delta x}$ = 1.81 °C cm^{-1}.

At x = 50 cm the value of R is 25 W.

Use these data and your answer to **d ii** to suggest whether the rate R of thermal energy transfer is in fact proportional to the temperature gradient. [3]

f (HL only) It is suggested that the variation with x of the temperature θ is of the form

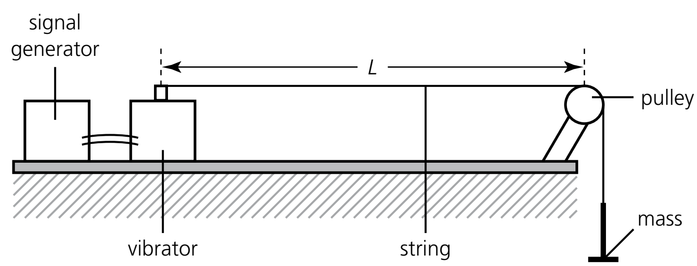
$\theta = \theta_0 e^{-kx}$

where θ_0 and k are constants.

State how the value of k may be determined from a suitable graph. [2]

Standard Level and **Higher** Level Paper 2, May 07 TZ1, Q1

4 This is a data analysis question. The frequency f of the fundamental vibration of a standing wave of fixed length is measured for different values of the tension T in the string, using the apparatus shown.

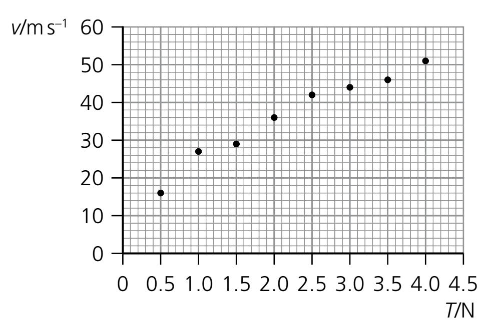


In order to find the relationship between the speed v of the wave and the tension T in the string, the speed v is calculated from the relationship

$v = 2fL$

where L is the length of the string.

The data points are shown plotted on the axes below. The uncertainty in v is ±5 m s^{-1} and the uncertainty in T is negligible.



a Make a copy of the graph and draw error bars on the first and last data points to show the uncertainty in speed v. [1]

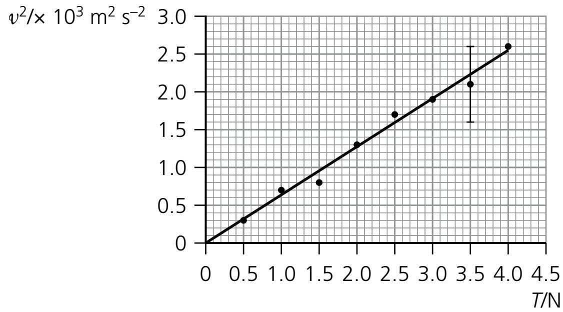
b The original hypothesis is that the speed is directly proportional to the tension T. Explain why the data do **not** support this hypothesis. [2]

c It is suggested that the relationship between speed and tension is of the form

$v = k\sqrt{T}$

where k is a constant.

To test whether the data support this relationship, a graph of v^2 against T is plotted as shown below.



The best-fit line shown takes into account the uncertainties for each data point. The uncertainty in v^2 for T = 3.5 N is shown as an error bar on the graph.

i State the value of the uncertainty in v^2 for T = 3.5 N. [1]

ii At T = 1.0 N the speed v = 27±5 m s^{-1}. Calculate the uncertainty in v^2. [3]

d Use the graph in **c** to determine k without its uncertainty. [4]

Standard Level and **Higher** Level Paper 2, Nov 09, Q1

25 Preparing for the IB Diploma physics examination

25.1 Revision techniques

Students have many different ways of revising for examinations, but some general words of advice apply to everyone:

- Use the IB syllabus and the chapter summaries in this book to identify your strengths and weaknesses at an early stage of your revision. Effective revision will concentrate on the parts of the syllabus with which you are less confident, rather than repeating topics you already know well. Monitor your progress, and you will be motivated by the fact that the list of topics left to revise becomes shorter and shorter!
- It is important to realize that many examination questions are *very closely linked* to statements in the IB Diploma physics syllabus. For example, wherever the syllabus uses *define*, *derive* or *state*, this is a clear indication of what may be asked in examination questions. Effective revision will take this into account. Such statements are highlighted throughout this book in blue boxes.
- Revision should usually be *active* rather than passive. Discussing physics or answering questions is generally much more useful than reading or watching a physics video. However, it is sensible to *start* your revision of a topic by re-reading the appropriate section of this book.
- Answering questions from past examination papers is very important and many students and teachers believe that it is the best way to revise. You should do as many questions as you can, and check your answers against the IB mark schemes (see page 10), or have them checked by someone else. Completing 'mock exams', in which you answer all the required questions on complete examination papers in the regulation times, will also help you judge whether you are working too quickly or slowly. You may also have taken a series of tests and examinations during your course; these are valuable resources for revision – we should all learn from our mistakes!
- Very few students enjoy revision so it is a good idea to use a variety of different revision techniques to stimulate interest. Some students also find that working in different surroundings can be a way to freshen up their revision. It is not a good idea to force yourself to revise when you are tired, nor to work for too long at one time. (Between 40 and 60 minutes may be the ideal length of time for revision without a break.)
- Most students find that planning a revision schedule helps to organize and structure their work, but it can be a waste of time if you don't stick to your schedule, so that you need to keep re-writing it!
- You must be familiar with the structure of the examination papers and style of different exam questions (see Table 25.1).

25.2 Examination paper details

Table 25.1

	Standard Level	Higher Level
Paper 1 (Core)	45 minutes 30 multiple-choice questions 20% of total examination mark	1 hour 40 multiple-choice questions 20% of total examination mark
Paper 2 (Core)	1 hour 15 minutes **Section A** has a data analysis question and about two or three other short questions (all compulsory). In **Section B** students must choose *one* of the three questions. Each question is usually in two parts on different sections of the syllabus. All questions to be answered in the boxes provided on the examination paper. 32% of total examination mark	2 hour 15 minutes **Section A** has a data analysis question and about four or five other short questions (all compulsory). In **Section B** students must choose *two* of the four questions. Each question is usually in two parts on different sections of the syllabus. All questions to be answered in the boxes provided on the examination paper. 36% of total examination mark

	Standard Level	**Higher Level**
Paper 3 (Options)	1 hour The paper contains questions on the Options A–G. Students should answer *all* of the questions on any *two* of these options in the boxes provided on the examination paper. 24% of total examination mark	1 hour 15 minutes The paper contains questions on the Options E–J. Students should answer *all* of the questions on any *two* of these options in the boxes provided on the examination paper. 20% of total examination mark
Internal Assessment	24% of total examination mark	24% of total examination mark

25.3 Taking the examination

Make sure you take into the examination room all of the equipment that you may need: 30 cm ruler, protractor, compass, pens and pencils and a calculator of an approved type (but not allowed in Paper 1).

There are very few students who (with the same knowledge of physics) could not improve their marks simply by improving their examination technique! Here are some tips.

Paper 1

- The questions are generally arranged in approximate syllabus order.
- Although multiple-choice questions are often considered to be easier than many of the questions in the other two papers, it is common for students to make careless mistakes. If you have time, double-check your answers.
- Never select any answer until you have read *all* of the alternative possibilities.
- If you are unsure of an answer, do not spend too much time on it. Write comments next to the question, cross out any answers that seem obviously wrong (there are usually one or two!) and move on to the next question. If you work quickly enough you should have time to come back to unfinished questions later. The question may seem easier second time around.
- Be aware that sometimes a possible answer is a correct statement in itself, but not the correct answer to the question.
- Look out for the inclusion of two answers which contradict each other. It is likely that one of them is correct (and the other wrong).
- Remember that there is no penalty for wrong answers. Never leave a question without an answer, even if it is only a guess.

Papers 2 and 3

- Read through the *whole* question before you begin to answer any part. This is especially important when you have to decide which questions to choose in Part B of Paper 2.
- Judge the amount of detail you need to supply in your answers from the size of the space allowed and the number of marks allocated to the question. In general you need to make a separate point for each mark.
- Read the questions very carefully and note or underline key words and phrases. If a question asks you to 'use' certain information, graphs, or laws etc., make sure that your answer does exactly that. If you answer the question in some other way, even correctly, you will not get the marks. If you are asked to use a law or definition, begin your answer by quoting it.
- The presentation of your answers is important. Although neatness and good spelling are not directly assessed, they create a favourable impression. Use a ruler to draw straight lines.
- Always show all your working in calculations.
- Give your answers in decimals not fractions. Use scientific notation wherever appropriate.
- Use appropriate significant figures in your answers and do not forget units.

- In general, it is better to express physical quantities in words, rather than symbols, although standard symbols can be used in the working of calculations.
- If you need to change one of yours answers, neatly cross out the work you want to delete. If there is not enough space for your new work, use extra pages and attach them to your answer booklet at the end of the examination.

25.4 Examples of different styles of exam question (Papers 2 and 3)

Command terms

All Paper 2 and 3 questions in IB Diploma physics examinations contain one of a limited number of clear instructions, such as *define*, *outline* and *calculate*. These one-word instructions are known as *command terms* and they indicate the way in which you should answer the question. Some of these words are also used throughout the IB Diploma physics syllabus. The command terms can be divided into three groups:

- demonstrate understanding
- apply and use
- construct, analyse and evaluate.

Demonstrate understanding

These command terms will be used to test your memory of factual knowledge of the syllabus.

Define

You are required to give the *exact* meaning of a word, phrase or quantity. All the definitions that you may need are clearly indicated in the IB Diploma physics syllabus and in this textbook (including in the glossary). Students are strongly advised to make sure that they know all these definitions before taking the examination. A surprisingly high proportion of students fail to achieve these 'easy' marks!

A definition should usually be written in words, rather than as an equation, although an equation is acceptable if the meanings of all the symbols are explained.

Question: *Define* resistance.

Answer: Resistance is the ratio of the potential difference across a conductor to the current passing through it. $\left(R = \frac{V}{I}\right.$ is only acceptable if the symbols are explained.$\left.\right)$

Question: *Draw* a circuit to show how a cell, a thermistor and a variable resistor can be connected to provide a potential dividing input to another circuit.

Answer:

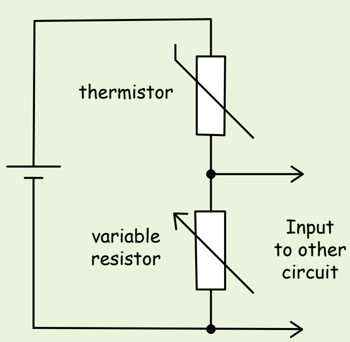


Figure 25.1

Draw

This widely used command term usually requires you to add something to an existing diagram or graph, but sometimes you may be asked to *draw* a completely new diagram. Use a sharp pencil and a ruler for drawing. *Drawing* a line of best-fit on a graph is a common question (see Chapter 24).

Note that, in this example, it is not essential to label the components if standard electrical circuit symbols are used (see the *IB Physics data booklet*). The question is partly aimed at testing whether the student knows these symbols. If you are unsure of the correct symbol, draw a box and write beside it what it is meant to represent.

Label

This common instruction is often combined with an instruction to draw something. It may refer to an existing diagram, or an addition to a diagram that you are asked to make, or occasionally to a new drawing. The labels should normally be in words rather than symbols. It is important to do this neatly and clearly, so that the examiner can be sure exactly what you are labelling.

Question: Draw an arrow on the diagram (Figure 25.1) to show the direction of the current and *label* it with the letter I.

This could be testing if the student knows that current flows conventionally from positive to negative around a circuit. The letter I should be placed close to the arrow. Without the labelling you may not be awarded the mark, even if the direction is correct.

Question: *Label* the drawing to show the meaning of amplitude and period.

Answer:

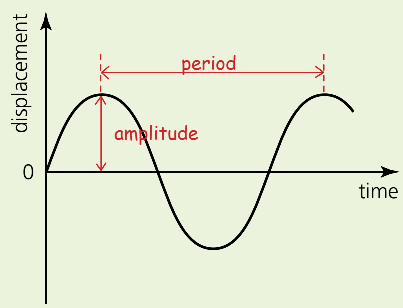


Figure 25.2

Question: A diagram shows a car moving at constant velocity. Draw fully *labelled* vectors to indicate the forces acting on the car.

Answer: At least four force vectors should be labelled as: weight, normal contact force, air resistance and push of road on tyres. The pairs of vectors should have equal lengths.

List

This means that you should provide a series of items *without* explanation. It should be clear from the question whether the terms need to be in any order.

Question: *List* the main energy sources which are used around the world to generate electrical power.

Answer: coal, nuclear, oil, natural gas, hydro-electric. (In this answer the order of the list is not important, nor is the list complete. A complete list would not be required by the examiner because the question is not intended to be definitive.)

Measure

This command term requires that the student measures the value of a physical quantity from a diagram or graph on the examination paper. It could be a length, an area, an angle, or may require interpretation from a scale. Clearly students should be prepared by taking a 30 cm ruler and a protractor into the examination room. Obviously, measurement has to be accurate in order to get the marks available.

Question: A diagram shows a ray of light being refracted as it enters glass. Take *measurements* in order to calculate the refractive index of the medium. (This combines measurements with a calculation.)

Answer: It will be necessary to measure the angles of incidence and refraction.

State

State is similar to *define*, in that a short, precise answer is required, without the need for any further explanation. This is one of the most widely used command terms in IB Diploma physics examinations. The syllabus contains some important laws and terms that may need to be 'stated' in an examination, and these should be memorized in the same way as definitions (they are also highlighted in the glossary).

Question: *State* what is meant by damping.

Answer: Damping is the dissipation of the energy of an oscillation when it is acted on by a resistive force.

Apply and use

These command terms will be used to test your ability to *use* the concepts and principles of physics that you have learned during the course.

Annotate and apply

These are unusual command terms in IB Diploma physics examinations.

- To *annotate* is similar to *label*, but requires *brief notes* to be added to a diagram or graph.
- To *apply* means to use knowledge in a new situation. For example, you could be asked to *apply* your understanding of Newton's laws of motion to a fairground ride.

Calculate

In this very common type of question you are required to use data given in the question and/or in the *IB Physics data booklet* in order to determine a numerical answer.

- You must show clearly *how* you obtained your answer. Marks are usually awarded for correct working, even if your final answer is wrong.
- Your answer must have a suitable number of significant figures (see Chapter 1).
- Your answer must have a unit (unless it is a ratio).

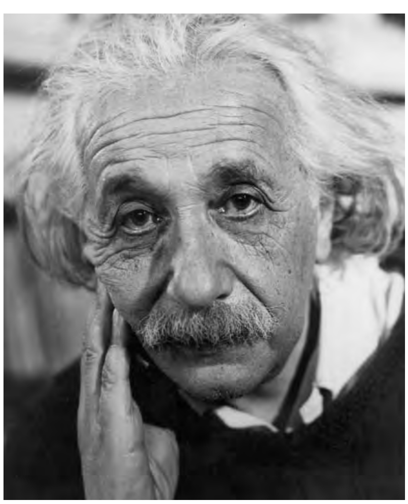


Figure 25.3 Einstein's dislike of examinations has been widely reported

Question: A stone is thrown vertically upwards with a speed of $8.10\,\mathrm{m\,s^{-1}}$. Ignoring air resistance, *calculate* the maximum height reached by the stone.

Answer: $v^2 = u^2 + 2as$ (Quote the equation you are using.)
$0 = 8.10^2 + (2 \times -9.81 \times s)$ (Substitute in data. The value for a was taken from the IB *Physics data booklet*.)
$s = 3.34\,\mathrm{m}$

Describe

Use knowledge from the course and/or information given in the question to give a straightforward account. The amount of detail needed will vary from question to question, and is best judged from the number of marks available for the answer.

The command term *describe* appears frequently in the IB syllabus.

Question: Describe what is meant by the term resonance. (3 marks)

Answer: Resonance is the *increase in the amplitude* of an oscillation when it is *disturbed by an external force* which has the *same frequency as the natural frequency* of the oscillator.

Note that there are three separate ideas included in this answer – in response to the three marks allocated to the question.

Distinguish

This command term means you should explain the essential difference(s) between two things (or more). You may also *briefly* indicate what they have in common. (See also *compare*, page 10.)

Question: *Distinguish* between speed and velocity.

Answer: Speed is calculated from distance/time. Velocity has the same magnitude as speed, but a direction of motion must also be given. (Examples might help, but are not essential, unless asked for in the question.)

Estimate

This is similar to *calculate*, but an accurate answer will not be possible. For example, the question may involve you making a calculation based on your reasonable estimates of unknown quantities. Estimated answers should be given with an appropriate number of significant figures (which may be only one), or just an order of magnitude.

Making estimates is demanding for many students, but marks will be awarded for *any* reasonable estimates, rather than an expected answer. Your assumptions should be stated clearly.

Question: *Estimate* the amount of coal that would be burned in a 100 MW power station in one hour.

Answer: Mass of coal needed every second = output power/(efficiency × energy density)

Assuming that the power station has an efficiency of 35% and the energy density of the coal used is 30 MJ kg^{-1}:

Mass of coal needed per second = $10^8/(0.35 \times 3 \times 10^7) = 9.5$ kg

Mass of coal needed per hour = $9.5 \times 3600 \approx 3 \times 10^4$ kg

Identify

This requires that you give a name for something, or select a correct answer from a number of different possibilities (which may, or may not, be provided in the question). Only a brief answer is required.

Question: The following equation represents the production of carbon-14 in the Earth's atmosphere by neutron bombardment of nitrogen.

$$^{14}_{7}N + ^{1}_{0}n \rightarrow ^{14}_{6}C + X$$

Identify the particle denoted by X.

Answer: a proton ($^{1}_{1}p$ is an acceptable answer because it is a standard symbol).

Outline

This is similar to *describe*, except that full details are *not* needed. This command signifies to the reader to give only a *brief* answer, and there will usually only be one or two marks for such a question.

Question: *Outline* how the energy carried by a water wave can be converted to electrical energy in an ocean-wave energy converter.

Answer: As the crests and troughs of the water pass through the converter, air is forced backwards and forwards past a turbine. The turbine causes coils of wire to rotate in a magnetic field and generate a voltage.

Construct, analyse and evaluate

This group of command terms may involve less familiar situations or a greater depth of understanding. They can test more complex skills, such as critical thinking and imagination.

Comment

This command term requires your judgment or opinion of a calculated numerical answer or a statement provided in a question. Usually only one comment is required.

Question: The value for the specific latent heat of fusion of ice determined by experiment using an immersion heater was lower than the accepted value. *Comment* on this difference.

Answer: This was probably because thermal energy was transferred to the ice from the hotter surroundings, so that less energy was needed from the heater in order to melt it.

Question: There is an enormous amount of energy in the waves on the world's oceans. *Comment* on the fact that very little of this energy is transferred to forms which are useful to us.

Answer: The construction and maintenance costs of ocean-wave energy converters are currently much more expensive than for most other energy sources.

Deduce

This is a widely used command term. To *deduce* means to reach a conclusion (stated in the question) from the information provided. Most commonly, this will require you to show all the steps in a calculation and, in this respect, *deduce* has a similar meaning to *show*; however, *deducing* is not quite as straightforward and may involve more steps (and marks). As with *show*, it is important to show every step of the calculation.

Question: A laser has an output power of 4.0 mW and forms a parallel beam of area 0.46 cm^2 which strikes a surface perpendicularly. If the wavelength of the light is 630 nm, *deduce* that photons are striking the surface at a rate of about $3.0 \times 10^{16}\,cm^{-2}s^{-1}$. (3 marks)

Answer: Number of photons per second in the beam = power/energy of each photon $\left(\frac{hc}{\lambda}\right)$

$= (4.0 \times 10^{-3})\,/(6.63 \times 10^{-34} \times 3.00 \times 10^{8}/6.30 \times 10^{-9})$

$= 1.3 \times 10^{16}\,s^{-1}$

Rate per $cm^2 = 1.3 \times 10^{16}/0.46 = 2.8 \times 10^{16}\,cm^{-2}s^{-1}$

Note that you should show the answer that is produced from the data provided, *not* just the approximate answer provided in the question.

Derive

To *derive* means to show all the physics principles and mathematical reasoning that leads to a particular equation (usually a well-used relationship, the derivation of which is clearly referred to in the IB Diploma physics syllabus).

Question: *Derive* an expression for the gravitational field strength on the surface of a planet in terms of its mass M, its radius R and the universal gravitational constant G. (2 marks)

Answer: gravitational field strength, g, is defined as gravitational force/mass

$$g = \frac{F}{m}$$

From Newton's universal law of gravitation we know that

$$F = G\frac{Mm}{R^2}$$

(In this case there is no need to explain the symbols, because they are explained in the question).

Combining these two equation gives

$$g = \frac{GM\cancel{m}/R^2}{\cancel{m}} = GM/R^2$$

Determine

This command term usually relates to questions requiring numerical answers. It has a meaning very similar to *calculate*, although the term itself relates to finding a definite answer. The context of the questions may be more difficult than straightforward calculations. It is important that every step in the manipulation of equations is shown.

Discuss

This command term requires that the student presents (and compares) alternative explanations and opinions, or the advantages and disadvantages of various choices.

Question: *Discuss*, in terms of transportability and reproducibility, the advantages and disadvantages of storing text in analogue form or digital form. (4 marks)

Answer: Storing text in analogue form involves usually writing or printing words on paper (or making copies). Storing text in digital form involves converting data to binary form and recording onto optical disc or hard drive. Analogue text needs much more space for storage and transportation. This also involves time, whereas digital text can be transferred around the world at the touch of a button. Reproducing text that has been stored in analogue form will result in some deterioration in quality, but digital text can be reproduced accurately as many times as is required. Making multiple copies is quick and easy for digital text and this means that losing the text is less likely. It is difficult to see any advantage for analogue text in the terms of this question.

Answers to this kind of open-ended question can easily become too lengthy. Four marks are allocated to the answer, but the mark scheme may award marks to *any* four relevant comments. Note that the question requires answers *only* related to *transportability* and *reproducibility* and if you discuss other features, those comments will be ignored.

Explain

This command term is very widely used in examination papers, usually requiring the student to make something understandable by giving details, or the reasons why something may, or may not, happen. The detail required in an answer can be assessed from the number of marks allocated to the question.

Question: A constant forward force is used to accelerate a car. *Explain* why the magnitude of the acceleration produced by a constant forward force decreases as the car moves faster. (4 marks)

Answer: Acceleration is proportional to the resultant force acting on the car. The resultant force equals the forward force less the resistive forces opposing motion. As the car moves faster, the resistive forces (mainly air resistance) increase. So that the resultant force and acceleration decrease. (Four marks require that four different points of explanation are made.)

Show

This is similar to *calculate* and *determine*, but the main intention here is for students to *show in detail* how an answer (given in the question) was obtained, rather than just to perform the calculation. This kind of question may be asked in the first part of a series of calculations, so that you then have the correct data for further calculations. It is important that every step in the manipulation of equations is shown.

Question: An electron moved between charged parallel plates with a p.d. of 250 V across them. Show that the electric potential energy of the electron changed by 4.0×10^{-17} J

Answer: $\text{potential difference} = \frac{\text{energy transferred}}{\text{charge flowing}}$

$$V = \frac{E}{q}$$

Which can be rearranged to give $E = Vq$
(Include details of *every step* of the calculation.)

So that, $E = 250 \times 1.6 \times 10^{-19} = 4.0 \times 10^{-17}$ J

Sketch

This command term requires that you draw a graph, but without any numerical data. The word *sketch* does not imply 'untidy'. Your drawing *does* need to neat, so draw with a ruler! The axes should be labelled with the quantities that they represent, as should any important features of the graph (for example gradients or intercepts). You may need to take information from another part of question and add it to the graph.

Question: *Sketch* a graph to show how the force between two point charges varies with their separation.

Answer: The y-axis should be labelled as *force*, and the x-axis labelled as *separation*. The origin should be labeled (0,0). An appropriate, neatly drawn and labelled curve should indicate an inverse square relationship. The graph does not need to be plotted accurately, but it should be clear that the curved line will not touch the axes.

Suggest

This command term is usually used when a single correct answer is not possible, perhaps because not enough information is available; because a definite answer requires knowledge beyond that covered in the syllabus; or simply because there are many possible answers. Generally, there are several acceptable answers to this kind of question.

Question: *Suggest* a reason why the melting point of ice was measured to be –1.5 °C and not 0.0 °C (referring to an experiment in which the temperature of ice and water was measured over a period of time).

Answer: The thermometer used was wrongly calibrated. (Only *one* suggestion is required here. There is no way of knowing if this suggestion is actually correct. For example, an alternative answer could be that 'the ice was not pure'.)

Other terms

The following command terms are listed in the IB Diploma physics syllabus, but they are rarely used in examinations.

- **Analyse**. This means to use data or information provided in a question in order to reach some kind of conclusion.
- **Compare**. This is similar to *distinguish*, but requires a more detailed analysis of both the similarities and differences between two (or more) items.
- **Construct**. This command term is similar to sketch, but usually involves more detail, and may involve a step-by-step process.
- **Design**. This command term usually means that you are required to make some kind of plan.
- **Evaluate**. This requires you to consider the advantages and disadvantages of a process.
- **Predict**. This means that you are required to give the expected result of a course of action or calculation.
- **Solve**. This command term usually requires that an answer is determined by using algebraic methods.

25.5 Understanding mark schemes

How marks are allocated

After you have taken your IB Diploma physics examination, your answers will be sent to the IB office in the UK. The papers are then scanned and made available on a secure IB website for examiners (based all around the world), together with an agreed *mark scheme*. Examiners must use this detailed mark scheme when carefully assessing students' work. The marking of all examiners is checked carefully to ensure that the work of all students is treated fairly and equally. Examiners know nothing about the students – except their examination number.

As already mentioned, past examination questions and their mark schemes should be an important part of your revision. You should be aware of the following points when using the mark schemes.

- Marking is meant to be *positive*. Answers are given credit for the understanding that they demonstrate. The examiner will *not* usually look for the exact words shown in the mark scheme, but will award marks if the same ideas are shown clearly in some other way. (Anticipated alternative answers or wording are indicated in the mark scheme by a '/'.)
- **OWTTE** means 'or words to that effect'. This is used on the mark scheme when it is expected that different students will write different acceptable answers to a certain question.
- Each marking point starts on a new line and ends with a semicolon (;).
- Occasionally for some answers, a certain word is considered to be *essential*, indicated by <u>underlining</u> that word in the mark scheme.
- Words in brackets (…) are used to clarify an issue for the examiners. They are not required to gain the mark.
- The separate points in a mark scheme do not need to be in any particular order.
- You will *not* be penalized for poor grammar or spelling, as long as your meaning is clear.
- Sometimes there are more relevant points (which can be made in response to a certain question) than there are marks allocated to that question. For example there may be six or more marking points for a question which only has four marks. In this case, *any four* of those points will result in the maximum mark of four being awarded.
- **ECF** means 'error carried forward'. This is used by examiners to explain why they have given marks to an incorrect answer; for example, in a calculation, the student has only got the wrong answer because they used their incorrect answer to a previous part of the same question.
- In your answers to calculations, don't forget to give a *unit* (unless it is a ratio) and use the correct number of *significant figures*.
- If you have to take a measurement off a graph on the examination paper, there will be a range of acceptable answers, but accurate measurement is required, so be careful.