

AN INTRODUCTION TO ASTROPHYSICS AND COSMOLOGY BY ANDREW NORTON – 2016

Chapter I	Manipulating numbers and symbols	7
	Introduction	7
1.1	Algebra and physical quantities	7
	1.1.1 Manipulating algebraic expressions	7
	1.1.2 Rearranging algebraic equations	9
	1.1.3 Solving simultaneous equations	10
1.2	Powers, roots and reciprocals	10
	1.2.1 Combining powers	11
	1.2.2 Solving polynomial equations	12
1.3	Imaginary numbers	13
1.4	Unit (dimensional) analysis	14
1.5	Function notation	14
1.6	Powers of ten and scientific notation	15
1.7	Significant figures	16
1.8	Experimental uncertainties	17
	1.8.1 Types of uncertainty	17
	1.8.2 Estimating random uncertainties	19
	1.8.3 Uncertainties when counting randomly occurring events	20
	1.8.4 The uncertainty in a mean value	21
	1.8.5 Combining uncertainties in a single quantity	21
1.9	Logarithms and logarithmic functions	22
1.10	Graphs	24
	1.10.1 Straight-line graphs	25
	1.10.2 Making curved graphs straight	27
1.11	Angular measure	28
1.12	Trigonometry	29
	1.12.1 Trigonometric ratios	29
	1.12.2 The sine rule and cosine rule	30
	1.12.3 Trigonometric functions	31
	1.12.4 Inverse trigonometric functions	33
1.13	Vectors	34
	1.13.1 Vector components	34
	1.13.2 Addition and subtraction of vectors	35

1.13.3	Position and displacement vectors	36
1.13.4	Unit vectors	36
1.13.5	The scalar product	37
1.13.6	The vector product	37
1.14	Coordinates	38
1.15	Scalar and vector fields	40
1.16	Matrices	42
1.16.1	Combining matrices	42
1.16.2	Special types of matrices	44
1.16.3	Transposing matrices	45
1.16.4	The determinant of a matrix	45
1.16.5	Adjoint and reciprocal matrices	46

Chapter 2 Stars and planets 51

	Introduction	51
2.1	Measuring stars and planets	51
2.2	Units in astrophysics	53
2.3	Positions, distances and velocities	54
2.3.1	Observing the positions of stars	54
2.3.2	Measuring the velocities of stars	58
2.4	Spectra and temperatures	59
2.5	Luminosities and fluxes	63
2.6	Astronomical magnitudes	63
2.7	Colours and extinction	66
2.8	The Hertzsprung–Russell diagram	67
2.9	Masses of stars	69
2.10	Life cycles of stars	76
2.11	Stellar end-points	81
2.12	Planetary structure	81
2.12.1	Terrestrial planets	81
2.12.2	Giant planets	83
2.13	Extrasolar planets and how to find them	86
2.14	Astronomical telescopes	88
2.14.1	Telescope characteristics	89
2.14.2	Telescopes in other parts of the electromagnetic spectrum	92

Chapter 3 Galaxies and the Universe 97

Introduction	97
3.1 The Milky Way – our galaxy	97
3.2 Other galaxies	98
3.2.1 Classification of galaxies	98
3.2.2 Origin and evolution of galaxies	101
3.2.3 Measuring galaxy properties	101
3.3 The distances to other galaxies	102
3.4 Active galaxies	105
3.4.1 The spectra of active galaxies	105
3.4.2 Types of active galaxy	110
3.5 The spatial distribution of galaxies	113
3.6 The structure of the Universe	115
3.7 The evolution of the Universe	120
3.8 Observational cosmology	123
3.9 Cosmological questions	125

Chapter 4 Calculus 129

Introduction	129
4.1 Differentiation and curved graphs	129
4.2 Differentiation of known functions	131
4.3 The exponential function	133
4.4 The chain rule	135
4.5 Logarithmic differentiation	138
4.6 Expansions	139
4.7 Partial differentiation	142
4.8 Differentiation and vectors	143
4.9 Differential equations	144
4.10 Integration and curved graphs	146
4.11 Integration of known functions	147
4.12 Integration by substitution	149
4.13 Integration by parts	152
4.14 Multiple integrals	153

Chapter 5	Physics	159
	Introduction	159
5.1	Describing motion	159
	5.1.1 Motion in one dimension	159
	5.1.2 Motion in two or three dimensions	161
	5.1.3 Periodic motion	162
5.2	Newton's laws	163
	5.2.1 Newton's laws of motion	164
	5.2.2 Newton's law of gravitation	165
5.3	Relativistic motion	166
5.4	Predicting motion	168
	5.4.1 Work, energy, power and momentum	168
	5.4.2 Relativistic mechanics	170
5.5	Rotational motion	171
5.6	Properties of gases	174
5.7	Atoms and energy levels	178
	5.7.1 Atomic structure	178
	5.7.2 Photons and energy levels	180
5.8	Quantum physics	184
	5.8.1 Wave mechanics	185
	5.8.2 Quantum mechanics in atoms	189
5.9	Quantum physics of matter	191
	5.9.1 Quantum gases	191
	5.9.2 Nuclear physics	195
	5.9.3 Particle physics	200
5.10	Electromagnetism	202
	5.10.1 Electricity and magnetism	202
	5.10.2 Electromagnetic waves	204
	5.10.3 Spectra	207
	5.10.4 Opacity and optical depth	211
	Solutions	221

Introduction

In order to successfully study one or both of the Open University's Level 3 modules, S382 *Astrophysics* or S383 *The Relativistic Universe*, you should already be familiar with various topics in cosmology, astronomy, planetary science, physics and mathematics. The level of skills, knowledge and understanding that we expect you to have when you embark on either of these modules is equivalent to the end-points of the OU's Level 2 modules: S282 *Astronomy*, S283 *Planetary Science and the Search for Life*, S217 *Physics: from Classical to Quantum* and either MST224 *Mathematical Methods* or MST210 *Mathematical Methods, Models and Modelling*.

To ascertain whether or not you meet the required level before embarking on S382 and/or S383 you should work through the document entitled *Are You Ready For S382 or S383?* which is available from the Faculty website. If, as a result of attempting the questions in that document, you realise that you need to revise your skills, knowledge and understanding in certain areas of mathematics, physics, cosmology, astronomy and planetary science, then you should study the relevant chapters of this document carefully.

There are five main chapters to this document – one each to introduce the astronomy and planetary science, the cosmology and the physics background, plus two chapters of mathematics. It is important to note that, because most of this document revisits concepts and phenomena that are covered in detail in Level 2 Open University modules, the treatment here is much less rigorous than in the modules themselves. For the most part, the subjects covered here are merely *presented* to you rather than developed gradually through detailed argument. This is to enable you to get rapidly 'to the point' and appreciate the key information you need in order to understand what follows, and to allow you to progress quickly to the main substance of the Level 3 modules .

Acknowledgements

The material in this document has been drawn from the physics, maths and astronomy that is taught in various other OU modules, including S282, S283, SXR208/SXPA288, S207/S217, S103/S104 and S151. The authors of the relevant parts of those courses: David Adams, John Bolton, David Broadhurst, Jocelyn Bell Burnell, Derek Capper, Alan Cayless, Andrew Conway, Alan Cooper, Dan Dubin, Alan Durrant, Tony Evans, Stuart Freake, Iain Gilmour, Simon Green, Iain Halliday, Carole Haswell, Keith Higgins, Keith Hodgkinson, Anthony Jones, Barrie Jones, Mark Jones, Sally Jordan, Ulrich Kolb, Robert Lambourne, Ray Mackintosh, Lowry McComb, Joy Manners, David Martin, Pat Murphy, Andrew Norton, Lesley Onuora, John Perring, Michael de Podesta, Shelagh Ross, David Rothery, Sean Ryan, Ian Saunders, Mark Sephton, Richard Skelding, Tony Sudbery, Elizabeth Swinbank, John Zarnecki and Stan Zochowski are gratefully acknowledged, along with the other members of the teams responsible for those modules.

Grateful acknowledgement is also offered to Carolin Crawford for critically reading and Amanda Smith for proof-reading this document, although any remaining errors are the responsibility of the editor.

Grateful acknowledgement is made to the following sources of figures: Figure 1.4 (Photograph of Jupiter): NASA/Science Photo Library; Figure 1.4 (Photograph of the Earth): NASA; Figure 1.4 (Photograph of a galaxy): The Regents, University of Hawaii; Figure 2.2: Till Credner, Allthesky.com; Figure 2.29: Observatoire de Paris; Figure 2.31: S Korzennik, Harvard University Smithsonian Center for Astrophysics; Figure 3.10: Lee, J. C. et al. (2002) 'The shape of the relativistic Iron $K\alpha$ Line from MCG 6-30-15 measured with the Chandra high energy transmission grating spectrometer and the Rossi X-Ray timing explorer', *Astrophysical Journal*, Vol 570. ©The American Astrophysical Society; Figure 3.13: *Le Grand Atlas de l'Astronomie* 1983. Encyclopaedia Universalis; Figure 3.14: W. N. Colley and E. Turner (Princeton University), J. A. Tyson (Bell Labs, Lucent Technologies) and NASA; Figure 3.17: Adapted from Landsberg, P. T. and Evans, D. A. *Mathematical Cosmology*. 1977, Oxford University Press; Figure 3.21: Adapted from Schwarzschild, B. (1998) 'Very Distant Supernovae suggest that the cosmic expansion is speeding up', *Physics Today*, June 1998. American Institute of Physics; Figures 3.22 and 3.23: Bennett, C. L. et al. 'First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations', *Astrophysical Journal Supplement Series*, Volume 148, Issue 1.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked we will be pleased to make the necessary arrangements at the first opportunity.

Chapter 1 Manipulating numbers and symbols

Introduction

In this chapter, we will concentrate on the various rules for manipulating numbers and algebraic symbols, including how to manipulate equations containing fractions, powers, logarithms and trigonometric functions, and how to deal with vectors and matrices.

1.1 Algebra and physical quantities

Physical quantities, such as mass and position, are commonly represented by algebraic symbols such as m or x . Whenever such symbols are used, it should be recalled that they comprise two parts: a numerical value and an appropriate unit of measurement, such as $m = 3.4 \text{ kg}$ or $x = 6.0 \text{ m}$. The units will generally be internationally recognized SI units, although in astrophysics and cosmology, cgs units and other less conventional units are used where convenient.

Quantities may be combined using the standard operations of addition (+), subtraction (−), multiplication (×) or division (/ or ÷). Note that the order of addition or multiplication is not important; i.e. $a + b = b + a$, and $a \times b = b \times a$, but the order of subtraction and division is; i.e. $a - b \neq b - a$, and $a/b \neq b/a$.

Whenever quantities are combined, their units are combined in the same way. For example, if $p = mv$, where mass m is measured in kg and speed v is measured in metres per second (m/s or m s^{-1}), then their product p will have units of kilograms times metres per second, or kg m s^{-1} , pronounced ‘kilogram metres per second’. If two quantities are to be added or subtracted, then they must have the same units. (Unit analysis is discussed in Section 1.4.)

Both lower case and upper case letters are used as algebraic symbols, and in general will represent different quantities with different units. For instance, g is often used to represent the acceleration due to gravity near to the Earth’s surface (9.81 m s^{-2}), whilst G is the universal gravitational constant ($6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$). Note also that (upper and lower case) Greek letters are frequently used as symbols for physical quantities. You will soon become familiar with the letters that are commonly used.

1.1.1 Manipulating algebraic expressions

The most important rule to note when manipulating algebraic expressions is:

Algebraic symbols are manipulated in the same way as pure numbers and algebraic fractions are manipulated in exactly the same way as numerical fractions.

The following examples illustrate some of the rules of manipulating algebraic expressions. To multiply one bracket by another, multiply each term in the right-hand bracket by each term in the left-hand one, taking careful account of the signs, as in the following two cases:

$$(a + b)(c + d) = a(c + d) + b(c + d) = ac + ad + bc + bd$$

$$(a + b)(a - b) = a(a - b) + b(a - b) = a^2 - ab + ba - b^2 = a^2 - b^2$$

Exercise 1.1 Multiply out the following expressions to eliminate the brackets.

(a) $t[2 - (k/t^2)]$ (b) $(a - 2b)^2$

To multiply fractions, multiply the numerators (top lines) together and then multiply the denominators (bottom lines) together. For example, $\frac{2}{3} \times \frac{3}{4} = \frac{6}{12} = \frac{1}{2}$. So in general,

$$\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd} \quad (1.1)$$

To divide by a fraction, multiply by its reciprocal (i.e. by the fraction turned upside down). For example, $\frac{1}{3} \div \frac{1}{6} = \frac{1}{3} \times \frac{6}{1} = 2$. So in general,

$$\frac{a/b}{c/d} = \frac{a}{b} \div \frac{c}{d} = \frac{a}{b} \times \frac{d}{c} = \frac{ad}{bc} \quad (1.2)$$

In order to add or subtract two fractions, it is necessary for them both to have the same denominator. In numerical work, it is usually convenient to pick the smallest possible number for this denominator (the *lowest common denominator*), for example,

$$\frac{1}{3} - \frac{1}{6} = \frac{2}{6} - \frac{1}{6} = \frac{2-1}{6} = \frac{1}{6}$$

If the lowest common denominator is not easy to spot, you can always multiply the top and bottom of the first fraction by the denominator of the second fraction, and the top and bottom of the second fraction by the denominator of the first, for example,

$$\frac{1}{3} - \frac{1}{6} = \frac{1 \times 6}{3 \times 6} - \frac{1 \times 3}{6 \times 3} = \frac{6}{18} - \frac{3}{18} = \frac{3}{18} = \frac{1}{6}$$

This is the method to apply to algebraic fractions:

$$\frac{1}{a} + \frac{1}{b} = \frac{b}{ab} + \frac{a}{ab} = \frac{b+a}{ab} \quad (1.3)$$

$$\frac{1}{a} - \frac{1}{b} = \frac{b}{ab} - \frac{a}{ab} = \frac{b-a}{ab} \quad (1.4)$$

Exercise 1.2 Simplify the following expressions: (a) $\frac{2xy}{z} \div \frac{z}{2}$ (b) $\frac{a^2-b^2}{a+b}$

(c) $\frac{2}{3} + \frac{5}{6}$ (d) $\frac{a}{b} - \frac{c}{d}$

Even if the numerical values of algebraic quantities are known, it is advisable to retain the symbols in any algebraic manipulations until the very last step when the numerical values can be substituted in. This allows you to see the role of each quantity in the final answer, and generally minimizes errors.

1.1.2 Rearranging algebraic equations

Physical laws are often expressed using algebraic equations which may have to be rearranged to obtain expressions for the quantity or quantities of interest. For example, the equation relating the pressure, volume and temperature of a gas is usually written $PV = NkT$, but it may be that we wish to obtain an expression for T in terms of other quantities. The basic rule for manipulating an equation is that the equality must not be disturbed. That is,

Whatever you do on one side of the 'equals' sign, you must also do on the other side, so that the equality of the two sides is maintained.

This is illustrated by the following examples. (Do not be concerned at this stage with the meaning of the various symbols.)

Worked Example 1.1

- Rearrange $PV = NkT$ to obtain an expression for T .
- Rearrange $v = u + at$ to find an expression for t .
- Rearrange $E = \frac{1}{2}mv^2$ to obtain an expression for v .
- Rearrange $\omega = \sqrt{\frac{g}{L}}$ to obtain an expression for L , where ω is the Greek letter *omega*.

Solution

- To isolate T on the right-hand side, divide both sides by Nk . So $PV/Nk = NkT/Nk = T$, i.e. $T = PV/Nk$.
- First subtract u from both sides, to give $v - u = u + at - u = at$. Then divide both sides by a to give $(v - u)/a = at/a = t$. Hence $t = (v - u)/a$.
- The easiest route is first to isolate v^2 on one side of the equation. So first multiply both sides by 2 and divide both sides by m to give $2E/m = 2mv^2/2m = v^2$. Then taking the square root of both sides,

$$v = \pm \sqrt{\frac{2E}{m}}$$

The square root of a number has two values, one positive and one negative. The square root symbol ($\sqrt{\quad}$) denotes only the positive square root, hence the need for the ' \pm ' sign, which is read as 'plus or minus'. This reflects the *mathematics* of the problem. Sometimes the *physics* of the problem allows you to rule out one of these two values. For example, if v represented a speed, then it would have to be greater than or equal to zero, and thus only the positive square root would be retained.

- The first step is to square both sides of the equation: $\omega^2 = g/L$, and the next step is to multiply both sides by L/ω^2 , to give $L = g/\omega^2$.

Essential skill:
Rearranging equations

Having seen some examples, try the following for yourself.

Exercise 1.3 Rearrange each of the following equations to give expressions for the mass m in each case.

(a) $E = -\frac{GmM}{r}$ (b) $E^2 = p^2c^2 + m^2c^4$ (c) $T = 2\pi\sqrt{\frac{m}{k}}$

1.1.3 Solving simultaneous equations

Two different equations containing the same two unknown quantities are called **simultaneous equations** if both equations must be satisfied (hold true) simultaneously. It is possible to solve such equations by using one equation to eliminate one of the unknown quantities from the second equation. An example should make the procedure clear.

Worked Example 1.2

In a certain binary star system it is determined that the sum of the masses of the two stars is $m_1 + m_2 = 3.3$ times the mass of the Sun, whilst the ratio of the two masses is $m_1/m_2 = 1.2$. What are the individual masses of the two stars?

Solution

If we rewrite the first equation to give an expression for m_1 in terms of m_2 , then we can insert this result into the second equation to give an expression for m_2 alone.

Rearrangement of the first equation gives $m_1 = 3.3 - m_2$, then substituting for m_1 in the second equation gives $(3.3 - m_2)/m_2 = 1.2$. Multiplying both sides of the equation by m_2 , we have $(3.3 - m_2) = 1.2m_2$; then adding m_2 to both sides gives $3.3 = 2.2m_2$; from which clearly $m_2 = 3.3/2.2 = 1.5$ times the mass of the Sun. Substitution for m_2 into *either* of the first two equations shows that $m_1 = 1.8$ times the mass of the Sun.

Note the important fact that in order to find two unknowns, two different equations relating them are required. By extension, *it is always necessary to have as many equations as there are unknowns*. You will find yourself constantly applying this principle as you solve numerical problems in astrophysics and cosmology.

Exercise 1.4 Solve the following pairs of equations to find the values of a and b . (a) $a - b = 1$ and $a + b = 5$ (b) $2a - 3b = 7$ and $a + 4b = 9$

1.2 Powers, roots and reciprocals

The **power** to which a number is raised is also called its *index* or *exponent*. So $2^4 = 2 \times 2 \times 2 \times 2 = 16$ can be said as ‘two to the power of four’ or simply ‘two

Essential skill:
Solving simultaneous equations

to the four'. Symbols and units of measurements can also bear indices. For example, the area of a square of side L is $L \times L = L^2$ and could be measured in square metres, written m^2 .

1.2.1 Combining powers

It is very important to understand how to manipulate the indices when quantities are multiplied and divided. As an example, consider multiplying 2^3 by 2^2 . This may be written out as $2^3 \times 2^2 = (2 \times 2 \times 2) \times (2 \times 2) = 2^5$. Generalizing from this example, for any quantity y ,

$$y^a \times y^b = y^{a+b} \quad (1.5)$$

From this rule, we can deduce many other properties of indices. For example, Equation 1.5 shows what a *power of zero* is. Since $y^a \times y^0 = y^{(a+0)} = y^a$, multiplying any quantity by y^0 leaves it unchanged. So, for any value of y ,

$$y^0 = 1 \quad (1.6)$$

Equation 1.5 can also be used to demonstrate the meaning of a *negative power*. Since $y^a \times y^{-a} = y^{a-a} = y^0 = 1$, dividing both the left- and right-hand sides of this equation by y^a shows that

$$y^{-a} = 1/y^a \quad (1.7)$$

Negative powers are frequently used with symbols in the units of physical quantities. For instance, speed is measured in metres per second, written in symbols as m/s or m s^{-1} . By use of negative indices, Equation 1.5 can easily be applied to situations in which quantities are divided by one another. For example, $10^5/10^3 = 10^5 \times 10^{-3} = 10^{5-3} = 10^2$, or more generally,

$$y^a/y^b = y^{a-b} \quad (1.8)$$

A *fractional power* denotes the root of a number and this too can be deduced from Equation 1.5. Thus $y^{1/2} \times y^{1/2} = y^{(1/2+1/2)} = y^1 = y$. So $y^{1/2}$ is the quantity that when multiplied by itself gives y . In other words, $y^{1/2}$ is the square root of y , $y^{1/2} = \sqrt{y}$. More generally, the quantity $y^{1/n}$ is the n th root of y ,

$$y^{1/n} = \sqrt[n]{y} \quad (1.9)$$

Consider raising to some power a quantity that already has an index, such as $(2^2)^3$. Writing this out in full shows that $(2^2)^3 = (2^2) \times (2^2) \times (2^2) = 2^{(2+2+2)} = 2^6 = 2^{2 \times 3}$, or in general,

$$(y^a)^b = y^{ab} \quad (1.10)$$

Like Equation 1.5 or Equation 1.8, this rule applies to any powers, whether positive or negative, integer or fractional.

Exercise 1.5 Simplify the following to the greatest possible extent. (You should not need to use a calculator, but you may find that doing so helps you to understand some of the individual steps in the calculation.)

- (a) $10^2 \times 10^3$ (b) $10^2/10^3$ (c) t^2/t^{-2} (d) $1000^{1/3}$ (e) $(10^4)^{1/2}$ (f) $125^{-1/3}$
 (g) $(x^4/4)^{1/2}$ (h) $(2 \text{ kg})^2/(2 \text{ kg})^{-2}$

1.2.2 Solving polynomial equations

A **polynomial** is an expression which is composed of one or more variables and constants, which are combined using only addition, subtraction and multiplication, or raising a variable to a non-negative, integer power. For instance, $x^3 + 7x - 15$, is a polynomial, but $x^3 - 7/x + 15x^{5/2}$ is not, because the second and third terms involve division by a variable (i.e. x to the power -1) and a non-integer power of x respectively.

Polynomials may be used to form **polynomial equations**, which are used to describe a wide range of problems in maths, physics and astronomy. The simplest polynomial equations to deal with are **quadratic equations**, namely equations containing terms no higher than the variable squared. They can be solved relatively easily, and in general will have two solutions. For instance, the quadratic equation $x^2 + 2x - 35 = 0$ has two solutions, $x = -7$ and $x = 5$. Substitution of either value for x into the quadratic equation will leave it balanced.

A quadratic equation can be written as the product of two factors. The example above is simply $x^2 + 2x - 35 = (x + 7)(x - 5)$. Clearly this allows the solution to be found by simply setting either factor equal to 0, i.e. either $(x + 7) = 0$ so $x = -7$, or $(x - 5) = 0$ so $x = 5$.

In general, the solutions to a quadratic equation of the form $ax^2 + bx + c = 0$ are given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (1.11)$$

The original quadratic equation may then be written as the product of two factors, as

$$ax^2 + bx + c = a \left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right) \left(x + \frac{-b - \sqrt{b^2 - 4ac}}{2a} \right)$$

Exercise 1.6 Determine the solutions of the following quadratic equations and write each as the product of two factors. (a) $4x^2 + 10x - 6 = 0$
 (b) $x^2 - 0.9x - 17.86 = 0$ (c) $8x^2 - 50 = 0$

The highest power in a polynomial equation is referred to as the *degree* of the polynomial. So $x^3 + x^2 + x + 1 = 0$ is a polynomial of degree three (or a cubic equation) and $x^4 + x^3 + x^2 + x + 1 = 0$ is a polynomial of degree four (or a quartic equation). The solutions to cubic and quartic equations may also be found using rules, although they are rather more complicated than Equation 1.11. There are no general formulae to solve polynomials of degree five or higher in terms of their coefficients.

Some polynomials, such as $x^2 + 1 = 0$, do not have any solutions that are real numbers. In order to solve such equations we must consider imaginary numbers, which are the topic of the next section.

1.3 Imaginary numbers

One of the main reasons for introducing the concept of imaginary numbers is that not every polynomial equation has solutions that are real numbers. For instance, the equation $x^2 + 1 = 0$ has no real solution, since it implies $x^2 = -1$ and there is no real number which, when multiplied by itself, gives the answer -1 . As there are no *real* numbers that satisfy this equation, we *imagine* there is such a number and give it the symbol i . This is referred to as the **imaginary unit**. It is important to realize, though, that i is a well-defined mathematical construct, despite its name!

i is the number which, when multiplied by itself, gives the answer -1 .

We may therefore write the solutions to $x^2 = -1$ as $x = +i$ and $x = -i$. The original quadratic equation may then be written in terms of factors as $(x + i)(x - i) = 0$.

- What are the result of the following multiplications: $(i \times i)$, $(-i \times -i)$, $(i \times -i)$?
- $(i \times i) = -1$,
 $(-i \times -i) = (-1 \times i \times -1 \times i) = (1 \times i \times i) = -1$,
 $(i \times -i) = (i \times -1 \times i) = (-1 \times -1) = +1$

A **complex number** is one which has both real and imaginary parts, such as $4 + 5i$. Operations that you are familiar with carrying out on real numbers (such as multiplication, division, raising to a power, etc.) can be extended to imaginary and complex numbers. The rule is simply to treat i as an unknown quantity whilst you manipulate the expression, and then use the definition above to replace any occurrence of i^2 with -1 . Higher integer powers of i can also be replaced with one of $\pm i$ or ± 1 .

- What are i^3 , i^4 , i^5 , and i^6 ?
- $i^3 = i^2 \times i = -1 \times i = -i$
 $i^4 = i^2 \times i^2 = -1 \times -1 = +1$
 $i^5 = i^4 \times i = +1 \times i = i$
 $i^6 = i^2 \times i^4 = -1 \times +1 = -1$

Often, i is loosely referred to as the ‘square root of minus one’ but this is not quite correct, and treating it as such can produce the wrong result. For instance if we write $-1 = i \times i = \sqrt{-1} \times \sqrt{-1}$, then we could combine the square root terms to give $-1 = \sqrt{(-1) \times (-1)} = \sqrt{1} = 1$ which is *incorrect*.

The calculation rule $\sqrt{a} \times \sqrt{b} = \sqrt{a \times b}$ is only valid for real, non-negative values of a and b .

To avoid making mistakes like this when you are manipulating complex numbers, the best strategy is simply *never to use a negative number under a square root sign*. For instance, if you are finding the solution of $x^2 + 9 = 0$, instead of writing $x = \sqrt{-9}$, you should write $x = i\sqrt{9}$, and clearly the answer is $x = \pm 3i$.

1.4 Unit (dimensional) analysis

Any equations involving physical quantities must have the same units on both sides; it would clearly be nonsensical to write 2 metres = 6 seconds. This is the basis of *unit analysis* or *dimensional analysis*. Dimensions in the sense used here are analogous to units: they express the nature of a physical quantity in terms of other quantities that are considered more basic. Thus, we can say that the units of area are metre², or equivalently, that area has the dimensions of length².

A good habit to cultivate is that of checking, whenever you write an equation, whether the units on both sides match – in other words whether the equation is ‘dimensionally correct’. For example, suppose you had derived the expression $\sqrt{2Ft^2/a}$ for a change in energy, where F was the magnitude of a force, t a time and a the magnitude of an acceleration. A quick check on the units would be enough to alert you that something was amiss.

The units of energy are $J = N\ m = \text{kg}\ \text{m}^2\ \text{s}^{-2}$.

The units of $\sqrt{Ft^2/a}$ are $\sqrt{\text{N}\ \text{s}^2/\text{m}\ \text{s}^{-2}} = \sqrt{\text{kg}\ \text{m}\ \text{s}^{-2}\ \text{s}^2/\text{m}\ \text{s}^{-2}} = \sqrt{\text{kg}\ \text{s}^2}$.

The units on either side of the proposed ‘equation’ are not the same, and this is sufficient reason to state unequivocally that the equation is wrong. Note however, that the reverse is not necessarily true. An equation may have the same units on either side (i.e. be dimensionally correct), but still be wrong because of a missing numerical factor or some other error.

Unit analysis can also help if you cannot quite remember an equation.

Worked Example 1.3

Is the equation for the speed of a sound wave $v = T/\lambda$ or $v = 1/\lambda T$ or $v = \lambda T$ or $v = \lambda/T$?

Solution

The units of wavelength λ are metres (m), and the units of wave period T are seconds (s). The only way to combine these to get the units of speed, $\text{m}\ \text{s}^{-1}$, is to divide wavelength by wave period, so the correct formula is $v = \lambda/T$.

Essential skill:
Unit analysis

1.5 Function notation

If the value of one variable, x say, is wholly or partly determined by the value of another, t say, then x is said to be a **function** of t . If we know the precise relationship between x and t then we can represent it by an equation. For instance, the position x of a body moving along a straight line at constant speed v , starting at x_0 , is given by the equation $x = x_0 + vt$. However, it is also possible to indicate the existence of a relationship between x and t in a very *general* way, writing $x = f(t)$ to indicate that ‘ x is a function of t ’. In this example we already know the form of f , it is simply $f(t) = x_0 + vt$, but it is possible to imagine cases where the functional form is different, or even unknown. In these cases, the more general expression $x = f(t)$ can be very useful. (There is nothing special about using the letter f to represent the function; we could equally have written $x = g(t)$, where $g(t) = x_0 + vt$, or indeed used any other letter.)

The use of brackets here is very different from in the sections above where they were used to indicate the precedence of algebraic operations. Now the brackets indicate that t is a variable which determines the value of x , and the value of x is given by the function f evaluated at t . The variable t is called the **argument** of the function f . In particular, the notation $f(t)$ does *not* indicate that some variable f is to be multiplied by another variable t . It will be clear from the context when brackets are being used to indicate precedence and when they are being used to indicate a functional relation.

1.6 Powers of ten and scientific notation

Powers of ten, such as $10^6 = 1000\,000$ (a million) or $10^{-3} = 1/1000 = 0.001$ (one thousandth), very often appear in astrophysics or cosmology because they provide a shorthand way of writing down very large or very small quantities. A quantity is said to be in **scientific notation** if its value is written as a decimal number between 1 and 10 multiplied by 10 raised to some power.

For example, the diameter of the Earth is about 12 760 km. In scientific notation this would be written as 1.276×10^4 km or 1.276×10^7 m. Scientific notation is equally useful for very small quantities: for instance, the mass of an electron is conveniently written as 9.1×10^{-31} kg. Scientific notation is particularly valuable in relation to significant figures, described in the next section. Furthermore, scientific notation is a great aid to calculation, since the decimal numbers and the powers of ten can be dealt with separately, as shown by the following example.

Worked Example 1.4

Calculate the value of a light-year (the distance light travels in one year), given that the speed of light is 3.00×10^8 m s⁻¹.

Solution

First we can write

$$1 \text{ year} = 365 \text{ days} \times 24 \frac{\text{hours}}{\text{day}} \times 60 \frac{\text{minutes}}{\text{hour}} \times 60 \frac{\text{seconds}}{\text{minute}} = 3.154 \times 10^7 \text{ s}$$

In that time, light travels a distance given by

$$\begin{aligned} \text{distance} &= \text{elapsed time} \times \text{speed} \\ &= 3.154 \times 10^7 \text{ s} \times 3.00 \times 10^8 \text{ m s}^{-1} \\ &= (3.154 \times 3.00) \times (10^7 \times 10^8) \text{ s m s}^{-1} \\ &= 9.46 \times 10^{15} \text{ m} \end{aligned}$$

Essential skill:
Using scientific notation

Exercise 1.7 The nearest star, Proxima Centauri, is 4.2 light-years away. Write this distance in kilometres expressed in scientific notation. ■

1.7 Significant figures

Suppose you are told that a particular neutron star orbits its companion star, travelling a distance of 4.1×10^6 km in 1.806×10^5 s. What is the speed v of the neutron star in its orbit? Clearly $v = (4.1 \times 10^6 \text{ km}) / (1.806 \times 10^5 \text{ s})$, and if you input this division into your calculator, you will get 22.702 104 0975 km s⁻¹. However, given the accuracy with which the distance and time were quoted, there is no justification for retaining this many digits in the answer.

The number of accurately known digits in the value of a physical quantity, plus one uncertain digit, is called the number of **significant figures**. Experimental results should always be quoted to a number of significant figures consistent with the precision of the measurement. If two or more quantities are combined, for instance by dividing one by the other, then the result is known only to the same number of significant figures as the *least* precisely known quantity.

In the example above, the calculator display could be ‘rounded’ to give

20 (to 1 significant figure)

23 (to 2 significant figures)

22.7 (to 3 significant figures)

22.70 (to 4 significant figures)

22.702 (to 5 significant figures)

If the last significant figure was followed by a digit from 0 to 4 it is unchanged in rounding; if it was followed by a digit from 5 to 9 it is increased by one.

So, how many significant figures should we quote in giving the neutron star’s speed? Clearly, the orbit has been timed to 4 significant figures, but the distance is only known to an accuracy of 2 significant figures (often abbreviated to 2 s.f.). So the final result for v is probably best quoted as 23 km s⁻¹, the same accuracy as the *least* accurate quantity used in the calculation. Note though that if a calculation involves multiple steps, it is best to retain more digits through the calculation, and to round to the correct number of significant figures only at the final stage. In this way you will avoid the introduction of rounding errors.

Scientific notation is also very helpful in dealing with significance in small numbers. Suppose you read that a certain pulsar flashes a pulse of radiation every 69 milliseconds, and you need to convert this period into seconds. Clearly, the value in seconds should be quoted with the same precision, i.e. the same number of significant figures, as the original measurement. In decimal notation, you would express the result as 0.069 s, which has *two* significant figures; *leading zeros do not count as significant figures*. In scientific notation, the measurement would be 6.9×10^{-2} s, having the same number of significant figures.

Very large numbers are often written (e.g. in the popular press) in a misleading fashion. The speed of light, for example, could be stated as 3×10^5 km s⁻¹ or 3.00×10^5 km s⁻¹, but it would be incorrect to write it as 300 000 km s⁻¹ because this implies that all six digits are significant. To such precision, the value would actually be 299 792 km s⁻¹. Some astrophysics texts adopt the convention that, in contrived examples, data can be assumed to have arbitrarily high precision, and that all zeros in provided numbers such as 6000 are significant. However, you

should always think about the context and the physical realities of the situation before making assumptions about the precision to which data are quoted.

Exercise 1.8 What, in km s^{-1} and to an appropriate number of significant figures, are the speeds involved in the following situations?

- An atom in the photosphere of a star travels 6093 km in 500 s.
- A star orbiting the centre of a galaxy travels 2.0×10^{18} km in 8.86×10^{15} s.
- Light travels 3000 km in 0.01 s.



I.8 Experimental uncertainties

Measured values of physical quantities are never exact. There are always uncertainties associated with measurements, and it is important to assess the size of the uncertainties and to quote them alongside the measured values. So if astronomers carried out some observations to determine the apparent magnitude of a particular star (see Section 2.6), then the form in which they would quote their result would be $m_v = 12.3 \pm 0.2$. This means that their best estimate of the value is a V-band magnitude of 12.3, and their confidence in this value is quantified by the uncertainty ± 0.2 , that is, the true value is probably between 12.5 and 12.1.

I.8.1 Types of uncertainty

Uncertainties arise in a variety of ways in astronomy. These include uncertainties caused by: lack of skill, instrumental limitations, extraneous influences, real variations in the quantity that is measured, and random fluctuations. These various uncertainties can be divided into two quite different types, those that are **random** and those that are said to be **systematic**.

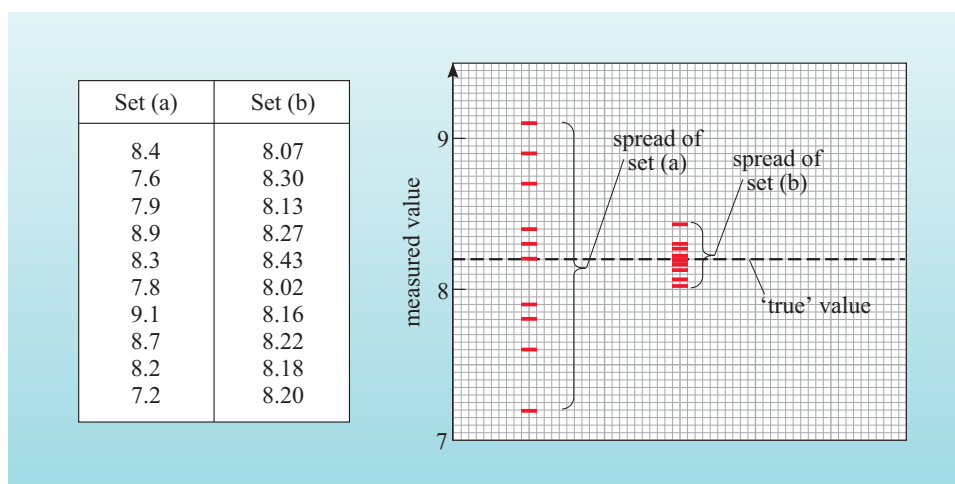


Figure 1.1 Two examples of random uncertainties. The two sets of measurements in the table, (a) and (b), are represented by the vertical positions of the dashes on the graph. The ten measured values for each set are scattered around the same true value. However, the range over which the measurements are scattered is much larger for set (a) than for set (b). This indicates that the random uncertainty is greater for set (a) than for set (b), which means that the precision of the measurements is lower for set (a) than for set (b).

A random uncertainty leads to measured values that are scattered in a random fashion over a limited range, as shown in Figure 1.1. The smaller the random uncertainty in the measurements, the smaller is the range over which they are

scattered. Measurements for which the random uncertainty is small are described as **precise**.

The best estimate that we can make for the value of the measured quantity is the **mean**, or average, of the measured values. As you might expect, if we make more measurements, then the mean value that we calculate is likely to be a better estimate of the quantity that we are measuring. We will make this statement quantitative later.

Systematic uncertainties have a different effect on measurements. A systematic uncertainty leads to measured values that are all displaced in a similar way from the true value, and this is illustrated in Figure 1.2. Such a situation may arise, for instance, if measurements are made using a ruler whose divisions are closer together or further apart than they should be. The two examples shown have the same random uncertainty: in both cases the spread, or scatter, of the values is the same. However, in both cases the measured values are systematically displaced from the true value. The values in set (b) are all larger than the true value, and the values in set (a) are all smaller. The difference between the mean value of a set of measurements and the true value is the systematic uncertainty. Measurements in which the systematic uncertainty is small are described as **accurate**. Therefore, to improve the accuracy of a measurement we need to reduce the systematic uncertainties.

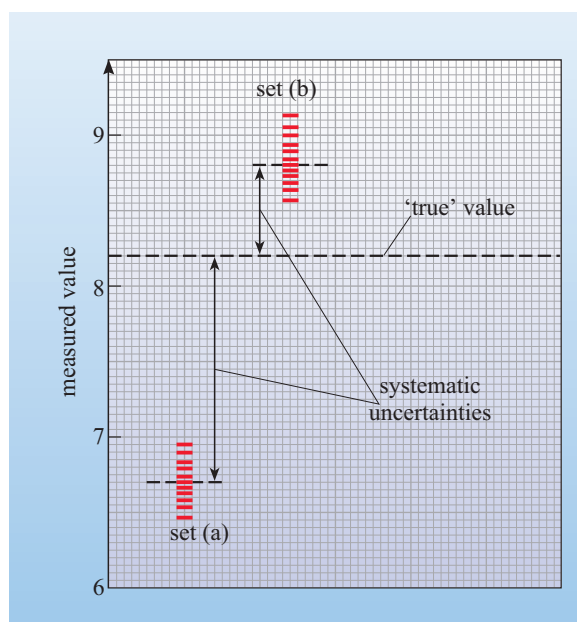


Figure 1.2 The effect of systematic uncertainties. Two sets of measurements, (a) and (b), are represented by the vertical positions of the dashes on the graph. For set (a), the systematic uncertainty causes all of the measured values to be smaller than the true value. For set (b) the systematic uncertainty causes all of the values to be larger than the true value, but the size of the uncertainty is smaller than for set (a). The measurements in set (b) are therefore more accurate than those in set (a).

I.8.2 Estimating random uncertainties

One way to estimate the size of random uncertainties in a measured value is by making a series of repeated measurements of the quantity. Random uncertainties lead to a scatter in measured values, and the uncertainty in the measurements can be deduced from the range over which the values are scattered. As a rough rule of thumb, we generally take the uncertainty in each measurement as about $2/3$ of the spread of the values. Generally, you should not be satisfied with making a single measurement of a quantity, but should repeat the measurement several times. The presence of a random uncertainty in a measurement can be detected – and its size estimated – by repeating the measurement a number of times.

As more measurements are made however, using the overall spread of the measurements, or even $2/3$ of the overall spread, as a measure of the random uncertainty may give a misleading estimate of how far a typical measurement lies from the mean, since the spread is calculated from only the maximum and minimum values. To avoid this problem, a measure of the random uncertainty is used that depends on the values of all of the measurements, not just the two most extreme. This is known as the standard deviation of the measurements, defined as follows:

The standard deviation s_n of a set of n measured values x_i is the square root of the mean of the squares of the deviations d_i of the measured values from their mean value $\langle x \rangle$,

$$s_n = \sqrt{\frac{\sum d_i^2}{n}} \quad (1.12)$$

where the deviation d_i of the measured value x_i from the mean value $\langle x \rangle$ is

$$d_i = x_i - \langle x \rangle \quad (1.13)$$

and the mean value $\langle x \rangle$ of the measurements is

$$\langle x \rangle = \frac{\sum x_i}{n} \quad (1.14)$$

The standard deviation is the most commonly used measure of the scatter of a set of measurements, and is used to quantify the likely random uncertainty in a single measurement. The standard deviation is sometimes known as the root-mean-square (rms) deviation, for obvious reasons.

A useful model to describe how often you will count a certain number of occurrences of an event (like the detection of a photon) in a certain time interval is the **Poisson distribution**. The Poisson distribution can be used to describe a large variety of phenomena that are relevant to astronomy, however it is only applicable when each event counted is independent of all the other events. The Poisson distribution is not symmetric about its mean value, but as the number of events increases, it does become symmetrical, and approaches the shape of a standard mathematical form known as the **normal distribution**, which is also known as the **Gaussian distribution**.

Figure 1.3 The standard deviation s_n characterizes the width of the Gaussian distribution. The shaded area under this Gaussian distribution curve represents the measurements that lie within $\pm s_n$ of the mean. This area is 68% of the total area under the curve, indicating that 68% of measurements are expected to fall within this range, hence the rule of two-thirds used earlier.

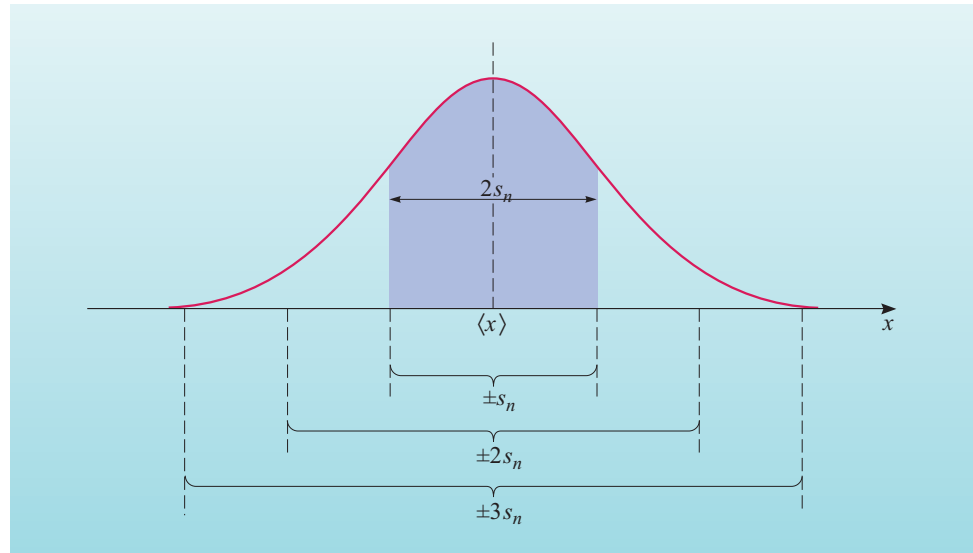


Figure 1.3 shows how the standard deviation of a Gaussian distribution curve is related to the spread of the curve. It is clear that a substantial fraction of measurements deviates from the mean value by more than the standard deviation s_n . For a particular range of the measured variable, the area under the distribution curve represents the fraction of measurements that lie within that range. For a Gaussian distribution, 68% of measurements lie within one standard deviation, i.e. within $\pm s_n$, of the mean value. Therefore, 32% of measurements are expected to differ from the mean by more than the standard deviation s_n . Note that the distribution curve falls off rapidly as the measurements deviate further from the mean. 95% of measurements lie within $\pm 2s_n$ of the mean and 99.7% of measurements lie within $\pm 3s_n$ of the mean.

1.8.3 Uncertainties when counting randomly occurring events

An important type of random uncertainty arises when investigating processes that involve counting events that fluctuate randomly, such as the number of photons from a star arriving on a detector. It turns out that if the number of *randomly fluctuating events* counted in a given period is n , the uncertainty in this number is given by

$$\text{uncertainty} = \sqrt{n} \quad (1.15)$$

This uncertainty is a measure of the likely difference between the value n that would be counted in any *single* measurement and the mean value of *many* measurements of n , namely $\langle n \rangle$, that would be found from a long series of repeated measurements. It is important to note that *increasing* the number of events *reduces* the fractional uncertainty:

$$\text{fractional uncertainty} = \frac{\text{uncertainty}}{\text{measured value}} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}} \quad (1.16)$$

I.8.4 The uncertainty in a mean value

The standard deviation s_n of a set of measurements tells us about how widely scattered the measurements are – it indicates how far the individual measurements are likely to be from the mean value. We usually take the mean value of the measurements as our best estimate of the true value, and so what we really need to know is how far the mean value is likely to be from the true value. In other words, we want to know the uncertainty in the mean value.

Not unreasonably, it turns out that the uncertainty in a mean value *decreases* as the number of measurements used to calculate the mean *increases*. In other words, you can reduce the uncertainties in an experiment by increasing the number of measurements that you make. The uncertainty σ_m in a mean value that is derived from n measurements that have a standard deviation s_n is

$$\sigma_m = \frac{s_n}{\sqrt{n}} \quad (1.17)$$

Note that, whereas the standard deviation tells us about the scatter of individual measurements, the uncertainty in the mean of n measurements tells us about the scatter of the mean values that are each derived from n measurements.

Exercise I.9 Ten measurements were made of the magnitude of a quasar, and the values obtained were:
 $m_v = 22.0, 21.6, 21.8, 22.3, 22.1, 22.0, 21.9, 22.2, 21.9, 22.2$

(a) What is the mean value of the quasar's magnitude? (b) Use the spread of the measurements to estimate the random uncertainty in an individual measurement of the quasar's magnitude. (c) Calculate the standard deviation of the ten measurements, and compare it with the estimate of the random uncertainty obtained in part (b). (d) Calculate the uncertainty in the mean magnitude.



I.8.5 Combining uncertainties in a single quantity

The rules for combining independent uncertainties are given below. If A and B are measured quantities, with uncertainties δA and δB respectively, then the uncertainty δX in the quantity X is as follows:

$$\text{If } X = kA \quad \text{then } \delta X = k \delta A \quad (1.18)$$

$$\text{If } X = kA \pm jB \quad \text{then } \delta X = \sqrt{(k \delta A)^2 + (j \delta B)^2} \quad (1.19)$$

$$\text{If } X = kA \times jB \quad \text{then } \frac{\delta X}{X} = \sqrt{\left(\frac{\delta A}{A}\right)^2 + \left(\frac{\delta B}{B}\right)^2} \quad (1.20)$$

$$\text{if } X = kA^n \quad \text{then } \frac{\delta X}{X} = n \frac{\delta A}{A} \quad (1.21)$$

where j , k and n are constants (i.e. they have no associated uncertainty).

In general, note that for any mathematical function $X = f(A)$ the uncertainty in X , indicated by δX , is related to the uncertainty in the quantity A , indicated by

δA , by the relationship

$$|\delta X| = |f(A \pm \delta A)| - |f(A)| \quad (1.22)$$

In other words, the magnitude of the uncertainty in X is equal to the magnitude of the value of the function evaluated at $(A \pm \delta A)$ minus the magnitude of the value of the function evaluated at A .

1.9 Logarithms and logarithmic functions

As discussed above, numbers such as 100 and 0.01 can be expressed in terms of powers of 10, respectively as 10^2 and 10^{-2} . In fact, by the use of decimal powers, *any* number can be expressed as a power of ten. If you like, you can check the following on your calculator, using the y^x or 10^x button.

$$1 = 10^{0.0}$$

$$2 = 10^{0.301} \text{ (to 3 s.f.)}$$

$$3 = 10^{0.477} \text{ (to 3 s.f.)}$$

$$3.16 = 10^{0.5} = \sqrt{10} \text{ (to 3 s.f.)}$$

$$10 = 10^{1.0}, \text{ etc.}$$

In each case, the power to which 10 is raised is called the **logarithm to base ten** or *common logarithm* (abbreviated \log_{10} or \log) of the resulting number. For example,

$$\log_{10} 100 = 2 \text{ since } 100 = 10^2$$

$$\log_{10} 0.1 = -1 \text{ since } 0.1 = 10^{-1}$$

$$\log_{10} 2 = 0.301 \text{ since } 2 = 10^{0.301} \text{ (to 3 s.f.)}$$

$$\log_{10} 3.16 = 0.5 \text{ since } 3.16 = 10^{0.5} \text{ (to 3 s.f.), etc.}$$

So in general,

$$\text{If } x = 10^a \text{ then } \log_{10} x = a \quad (1.23)$$

That is, the \log_{10} function reverses the operation of the 10^x function.

Exercise 1.10 Without using a calculator, write down the value of
(a) $\log_{10} 1000$ (b) $\log_{10} 0.001$ (c) $\log_{10} \sqrt{10}$

Logarithms are useful for dealing with numbers that range from very large to very small. Figure 1.4 illustrates a range of lengths in the natural world, most conveniently written in scientific notation and plotted using a logarithmic scale.

Base ten is commonly used for logarithms, since ten is the base of our counting system, but of course any number can be raised to a power and hence used as a base for logarithms. Logarithms to base e (where e is a very special number having a value 2.718 to 4 significant figures) are called **natural logarithms** (abbreviated \log_e or \ln), and are used extensively in describing many features of

the natural world. (The significance of e will become clearer in Section 4.4 on the *exponential function*.)

If $y = e^b$ then $\log_e y = b$ (1.24)

(Note: Some texts use 'log' (without a subscript) to refer to \log_e rather than \log_{10} . We recommend writing the subscript 'e' or writing 'ln' to avoid ambiguity.)

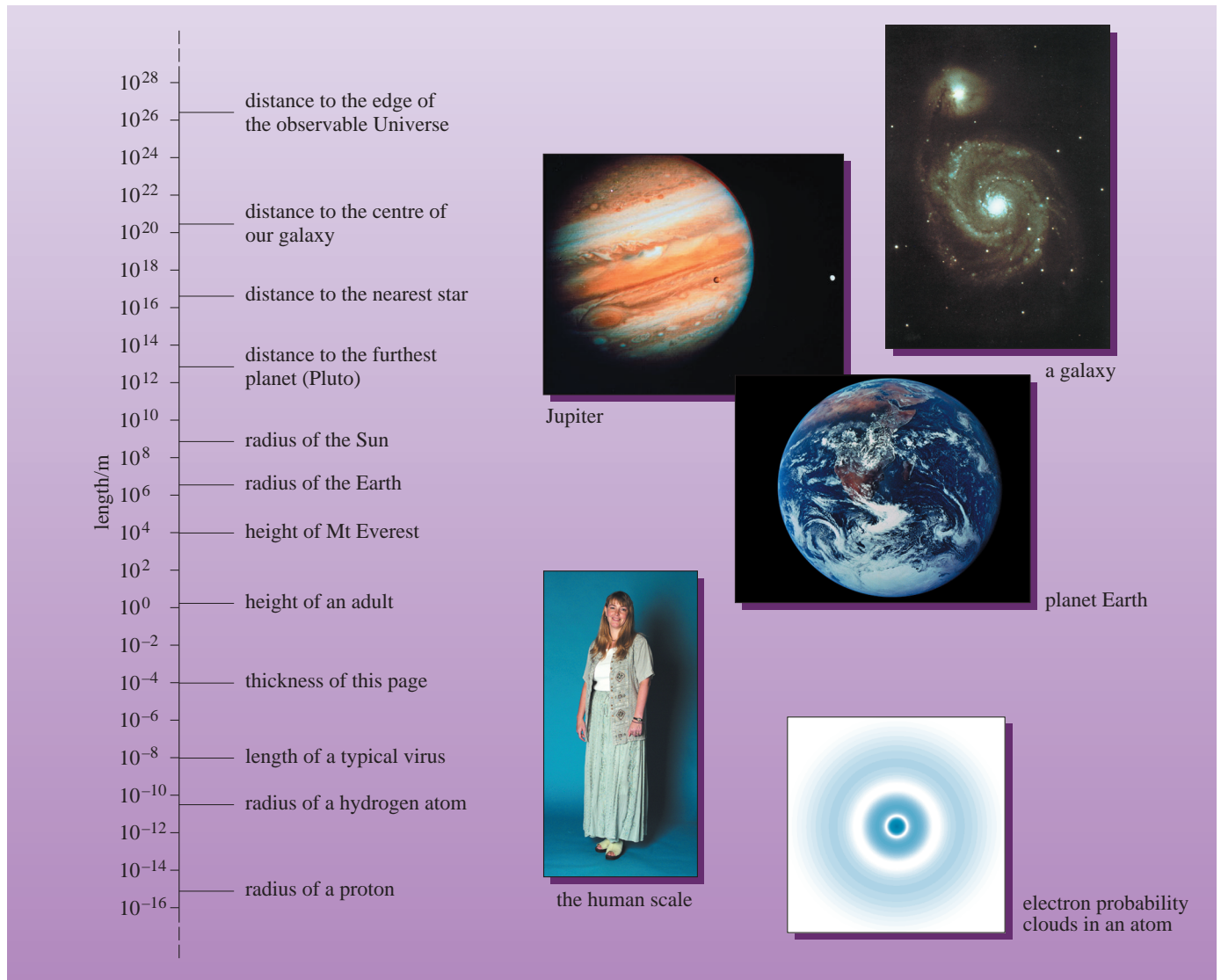


Figure I.4 Some examples of lengths relevant to astrophysics and cosmology. This diagram is plotted using a 'logarithmic' scale, in which each step represents *multiplication* by a factor of 10^2 . Note the difference from the more usual ('linear') type of scale, in which each interval along the axis represents a constant *addition*.

Three important rules for the manipulation of logarithms follow from the definitions above and the rules for combining powers. These rules apply

irrespective of which base (10, e, or some other) is adopted:

$$\log(a \times b) = \log a + \log b \quad (1.25)$$

$$\log(a/b) = \log a - \log b \quad (1.26)$$

$$\log a^b = b \log a \quad (1.27)$$

Taking the logarithms of both sides of an equation can often make the situation easier to deal with. The following example makes use of each of the rules above to illustrate this.

Essential skill:
Manipulating logarithms

Worked Example 1.5

The following equation expresses the angular momentum of a binary star system (see Sections 2.10 and 5.5, although the details of the astrophysics are not important for this example).

$$J = M_1 M_2 \left(\frac{Ga}{M_1 + M_2} \right)^{1/2}$$

Take the natural logarithms of each side of this equation.

Solution

Using the three rules above for the logarithms of products, ratios and powers of numbers, we have

$$\log_e J = \log_e (M_1 M_2) + \frac{1}{2} \log_e \left(\frac{Ga}{M_1 + M_2} \right)$$

so

$$\log_e J = \log_e M_1 + \log_e M_2 + \frac{1}{2} \log_e(Ga) - \frac{1}{2} \log_e(M_1 + M_2)$$

Make sure you can see how the expression above follows from the three rules.

Exercise 1.11 (a) Given that $\log_{10} 2 = 0.301$, without using a calculator, work out the values of:

(i) $\log_{10} 200$ (ii) $\log_{10} 32$ (iii) $\log_{10} 0.25$

(b) Using the basic rules of combining logarithms outlined above, rewrite each of the following expressions as a single logarithm:

(i) $\log_{10} 3 + \log_{10} 8$ (ii) $\log_{10} 4 - \log_{10} 3 - \log_{10} 5$ (iii) $3 \log_{10} 2$

1.10 Graphs

In astrophysics and cosmology, as in most areas of science, you will frequently encounter graphs as a means of depicting how one quantity varies as a function of another.

I.10.1 Straight-line graphs

Two quantities related in such a way that if one is doubled the other also doubles are said to be **directly proportional**. To take a simple example, Hubble's law at low speeds (see Chapter 3) states that the apparent speed of recession, v , of a galaxy is directly proportional to its distance away from us, d . If we find a galaxy with double the value of v , then d doubles. Trebling v trebles d , and when $v = 0$, $d = 0$. This relationship of proportionality is indicated by writing $v \propto d$ (read as ' v is proportional to d ') or $v = kd$, where k is called the constant of proportionality. A graph of v against d is a straight line that passes through the origin (Figure 1.5).

Note how the axes in Figure 1.5 are labelled. The scales show pure numbers, so in each case the quantity involved is divided by the units in which it is measured ($v/\text{km s}^{-1}$ and d/Mpc) to give a dimensionless number. The shorthand way of describing a graph in which y is plotted vertically and x horizontally is a 'graph of y against x ' or ' y versus x '.

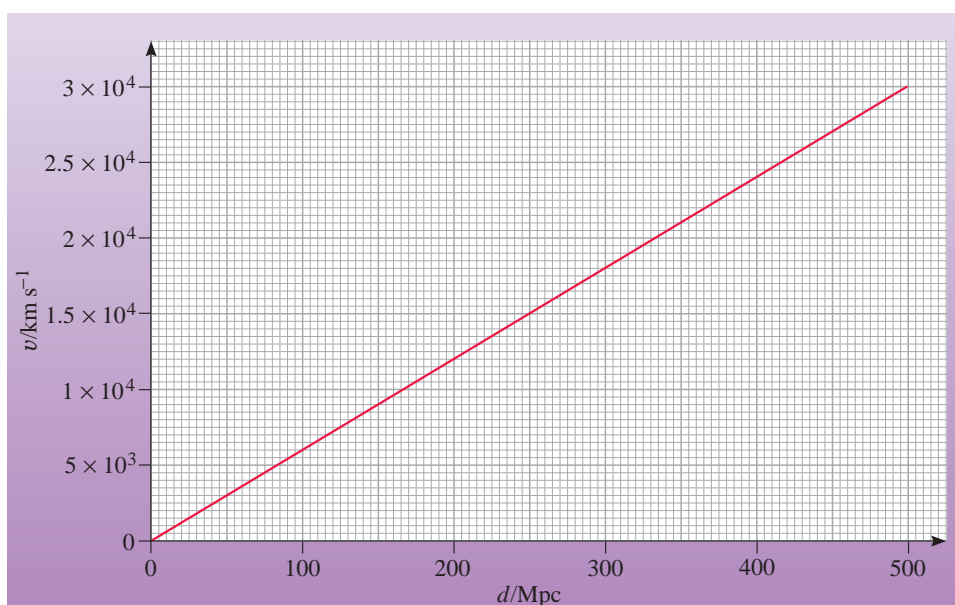


Figure 1.5 A graph showing how the apparent speed of recession v of distant galaxies varies with their distance d from Earth. Notice that the unit of the distance axis is given as 'Mpc' – this is the megaparsec – a commonly used unit of astronomical distance (see Section 2.2).

Quantities related in such a way that if one halves, the other doubles, are said to be **inversely proportional**. The pressure P and volume V of a fixed amount of gas at constant temperature (such as in an interstellar cloud) are inversely proportional: $P \propto 1/V$. A graph of P against V is thus a curve (of a shape called a hyperbola, Figure 1.6a), but a graph of P against $1/V$ or $1/P$ against V is a straight line through the origin (Figures 1.6b and c).

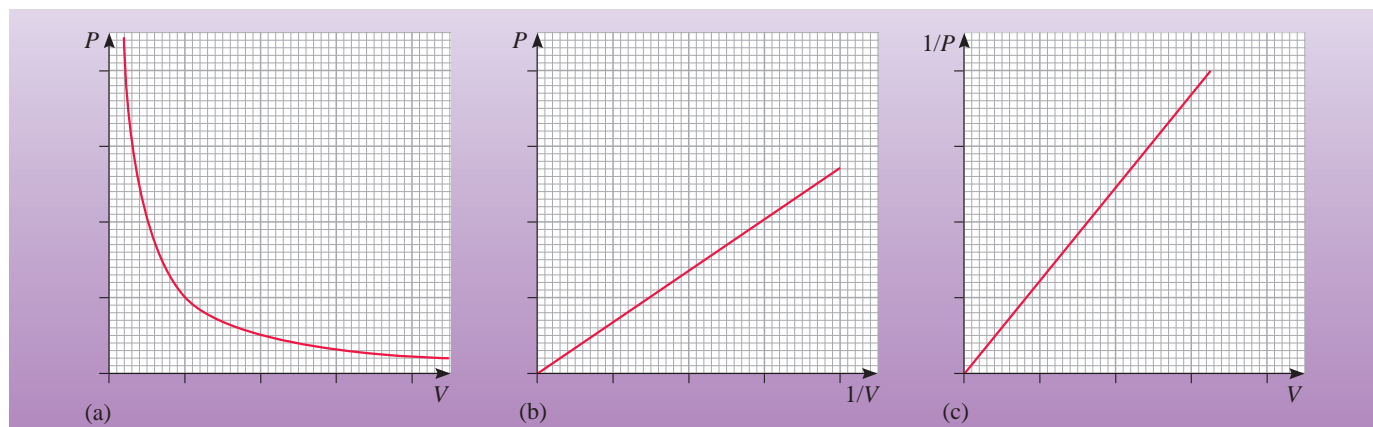


Figure 1.6 Graphs showing the way in which the pressure P of a fixed amount of gas at constant temperature depends on its volume V .

Any linear (straight-line) graph has a constant slope or **gradient**. We can use Figure 1.5 to calculate the gradient for the Hubble's law example. First, choose two convenient but well separated points A and B on the graph, and read off the corresponding values of d , to be called d_A and d_B . The difference between them is the change in d . Changes are usually denoted by the upper case Greek letter *delta*, symbol Δ . Thus, Δd (read as 'delta dee') means the change in d : $\Delta d = d_B - d_A$. The corresponding change in v , written Δv , can also be read off the graph: $\Delta v = v_B - v_A$. The gradient of the line is then defined as

$$\text{gradient} = \frac{\Delta v}{\Delta d} = \frac{v_B - v_A}{d_B - d_A}$$

More generally, for a graph of y versus x ,

$$\text{gradient} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

The gradient of a graph represents the rate of increase of one quantity as another quantity increases.

Essential skill:

Measuring the gradient of a graph

Worked Example 1.6

Compute the gradient for the Hubble's law graph in Figure 1.5.

Solution

In this case

$$\Delta v = (3.0 \times 10^4 \text{ km s}^{-1}) - (1.5 \times 10^4 \text{ km s}^{-1}) = 1.5 \times 10^4 \text{ km s}^{-1}$$

and

$$\Delta d = (500 \text{ Mpc}) - (250 \text{ Mpc}) = 250 \text{ Mpc}$$

So

$$\Delta v / \Delta d = (1.5 \times 10^4 \text{ km s}^{-1}) / (250 \text{ Mpc}) = 60 \text{ km s}^{-1} \text{ Mpc}^{-1}$$

The gradient $\Delta v / \Delta d$ represents the rate of increase of apparent speed with distance, which in this case is 60 kilometres per second per megaparsec. (Note: This value is in fact the Hubble constant, although the currently accepted value is somewhat different from this, see Chapter 3.)

In Example 1.6, the graph slopes upwards to the right, because v increases with d , and so the gradient $\Delta v/\Delta d$ is *positive*. A graph sloping downwards from left to right tells us that one quantity *decreases* as the other increases, and it will have *negative* gradient.

If we plot one quantity against another and get a straight line that crosses the axes at points away from the origin, then the quantities are not proportional to each other. However, since the graph is a straight line, we can still say that there is a *linear relation* between them. The points at which the line crosses the axes are called the **intercepts**. The general equation for a straight-line graph of y versus x is

$$y = mx + c \quad (1.28)$$

where c is the intercept on the y -axis, i.e. the value of y when $x = 0$, and m is the gradient. (Symbols other than m and c are sometimes used, but the meaning is the same.)

1.10.2 Making curved graphs straight

When a graph of one quantity plotted against another is a curve, and you think you know the relationship between them, it may be possible to re-plot the graph to produce a straight line.

If you suspect that $y = a_0 + a_1x^2$, where a_0 and a_1 are unknown constants, then you could plot y versus x^2 (rather than y versus x), to give a straight line of slope a_1 and y -intercept a_0 . Of course, if in reality $y = a_0 + a_1x^3$ then the graph will *not* be straightened by this approach, and another attempt, perhaps plotting y versus x^3 could be tried.

If you suspect that $y = ax^k$, where a and k are unknown constants, you can use logarithms to straighten the graph. (Such an equation is referred to as a *power law* for obvious reasons.) Taking the logarithm of both sides gives

$$\begin{aligned} \log y &= \log(ax^k) \\ &= \log a + \log x^k \\ &= \log a + k \log x \end{aligned}$$

Since $\log y = \log a + k \log x$, by plotting $\log y$ versus $\log x$ you would get a straight line whose slope is k and whose y -intercept is $\log a$. (This technique works irrespective of which base – 10, e , or some other – is used for the logarithms, so subscripts have been ignored.) Both the slope and intercept can be measured, so plotting a logarithmic graph allows a power law to be tested and the relevant power to be found.

Worked Example 1.7

The volumes V of stars are related to their radii r by the equation $V = \frac{4}{3}\pi r^3$. A plot of V against r will therefore be curved (Figure 1.7a). What straight-line graph involving V and r could be plotted instead?

Essential skill:
Making curved graphs straight

Solution

Since $\log_{10} V = \log_{10}(4\pi/3) + 3 \log_{10} r$, a plot of $\log_{10} V$ against $\log_{10} r$ is a straight line of gradient 3 and intercept $\log_{10}(4\pi/3)$ as shown in Figure 1.7b.

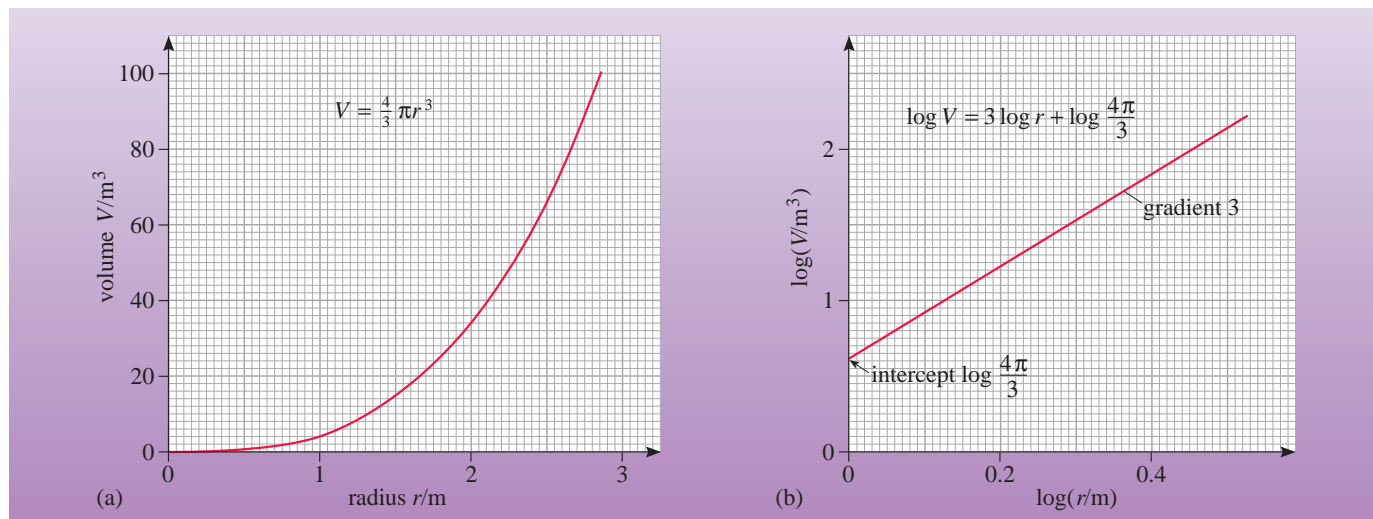


Figure 1.7 A log–log plot can identify a power law. (a) A plot of V against r is a curve, (b) but a plot of $\log V$ against $\log r$ is a straight line.

Exercise 1.12 Suppose that you have taken measurements of two quantities U and x , for various chosen values of x . You believe the two quantities to be related according to the equation $U = (kx^2/2) + c$, where k and c are constants. What graph would you draw to find out whether your data supported this belief?

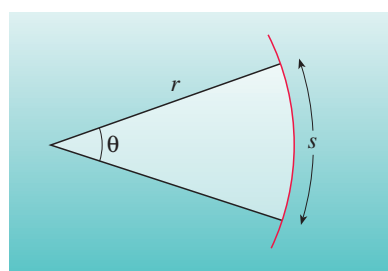


Figure 1.8 An arc of length s subtending an angle θ at the centre of a circle of radius r .

1.11 Angular measure

Plane angles may be measured in **degrees**, where 360° equals one complete turn. Subdivisions of a degree are the **arc minute** (often abbreviated to arcmin), where 1° equals 60 arcmin ($1^\circ = 60'$) and the **arc second** (abbreviated to arcsec), where 1 arcmin equals 60 arcsec ($1' = 60''$).

- How many arc seconds are there in one degree?
- 60 arcmin per degree \times 60 arcsec per arcmin = 3600 arcsec per degree.

In astrophysics and cosmology, angles are sometimes measured in the SI unit of radians, denoted by the abbreviation rad. A **radian** is defined as the angle subtended at the centre of a circle by an arc (i.e. part of the circumference of a circle) whose length is equal to that of the radius of the circle. So in general, if an arc of length s subtends an angle θ at the centre of a circle of radius r , as shown in Figure 1.8, then

$$\theta \text{ (in radians)} = s/r \quad (1.29)$$

Note that arbitrary plane angles are often denoted by the Greek lower case letter θ (*theta*) and that angles expressed in radians are the ratio of two lengths (s/r) and so are dimensionless (though not unitless). An arc of length $2\pi r$, i.e. the whole circumference, subtends an angle (in radians) of $2\pi r/r = 2\pi$. Therefore, 2π radians = 360° and so 1 radian = $360^\circ/2\pi = 57.3^\circ$ (to 3 s.f.).

Exercise I.13 (a) Convert the following into radians, expressing your answer as a fraction/multiple of π : 90° ; 30° ; 180° .

(b) Convert the following into degrees: $\pi/8$ radians; $3\pi/2$ radians.

The three-dimensional analogue of plane angles measured in radians is the concept of the solid angle measured in the SI unit of steradians, denoted by the abbreviation sr. A **steradian** is defined as the solid angle subtended at the centre of a sphere by a part of the surface of the sphere whose area is equal to the radius of the sphere squared. So in general, for a sphere of radius r , if part of the surface with area a subtends a solid angle Ω at the centre of the sphere, as shown in Figure 1.9, then

$$\Omega \text{ (in steradians)} = a/r^2 \quad (1.30)$$

Note that arbitrary solid angles are often denoted by the Greek upper case letter Ω (*omega*) and that solid angles expressed in steradians are the ratio of two areas (a/r^2) and so, like radians, are dimensionless (though not unitless). Since the surface area of a sphere is given by $4\pi r^2$, clearly there are 4π steradians in a complete sphere, 2π steradians in a hemisphere, and so on.

Exercise I.14 The power per unit solid angle emitted over a particular frequency range by a star is $1.4 \times 10^6 \text{ W sr}^{-1}$. What is the power per unit area received by a detector sensitive to this frequency range and placed $1.0 \times 10^{17} \text{ m}$ away from the star? (*Hint*: Calculate the solid angle subtended by a detector of unit area situated at this distance from the star.)

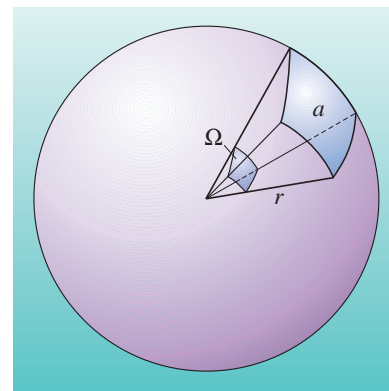


Figure 1.9 A sphere of radius r showing part of its surface of area a subtending a solid angle Ω at the centre.

I.12 Trigonometry

Trigonometric ratios and trigonometric functions are widely used in astrophysics and cosmology when solving geometric problems and when dealing with periodically varying phenomena.

I.12.1 Trigonometric ratios

Given the lengths of two sides of a right-angled triangle, the length of the third side is uniquely determined, being given by **Pythagoras's theorem**, which in terms of the symbols given in Figure 1.10 may be written as

$$h^2 = o^2 + a^2 \quad (1.31)$$

Furthermore, knowledge of the value of one of the acute angles of a right-angled triangle automatically gives the other angle too, since the angles of a triangle

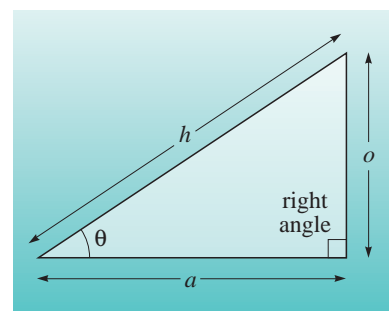


Figure 1.10 A right-angled triangle. The side opposite the right angle, called the hypotenuse, is of length h . The side opposite the angle θ has length o , the side adjacent to it has length a .

always add up to exactly π radians (180°). Although the triangle might be any size, it can only have one shape appropriate to those angles, and the ratios of any two sides of the right-angled triangle are defined. These ratios are given the following names:

$$\frac{\text{side opposite an angle}}{\text{hypotenuse}} = \text{sine of the angle; } \sin \theta = o/h \quad (1.32)$$

$$\frac{\text{side adjacent to an angle}}{\text{hypotenuse}} = \text{cosine of the angle; } \cos \theta = a/h \quad (1.33)$$

$$\frac{\text{side opposite an angle}}{\text{side adjacent to an angle}} = \text{tangent of the angle; } \tan \theta = o/a \quad (1.34)$$

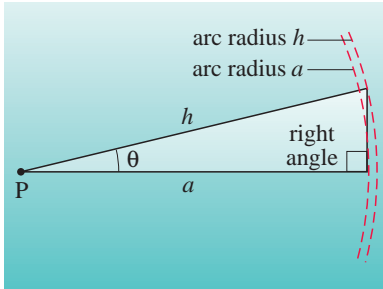


Figure 1.11 A right-angled triangle with a small angle θ .

Most calculators can calculate the sine, cosine or tangent of an angle. Many calculators can be set up to receive the angle in either degrees or radians, but *you must take care to set up your calculator correctly* according to your data. Usually this involves selecting ‘radian mode’ or ‘degree mode’; see your calculator handbook for details.

As the angle θ becomes smaller and smaller, o decreases and h becomes more and more nearly equal to a , as shown in Figure 1.11. So, from the definition of cosine,

$$\cos \theta \approx 1 \text{ when } \theta \text{ is small} \quad (1.35)$$

Figure 1.11 also shows that, for small θ , the length o approximates the length of an arc of a circle with centre P, and radius a (or h). So for an arc at radius a , the arc length is $o = a\theta$ and so $\theta = o/a$. Similarly, for an arc at radius h , the arc length is $o = h\theta$ and so $\theta = o/h$. Therefore, from the definitions of sine and tangent,

$$\tan \theta \approx \sin \theta \approx \theta \text{ when } \theta \text{ is small and in radians} \quad (1.36)$$

The so-called **small-angle approximations** in Equations 1.35 and 1.36 hold within 1% accuracy for angles less than about 0.2 radians ($\approx 11^\circ$).

Additional trigonometric relationships can be derived based on the properties introduced above. Taking Pythagoras’s theorem (Equation 1.31) and dividing both sides by h^2 , we see that $(o/h)^2 + (a/h)^2 = 1$, so

$$\sin^2 \theta + \cos^2 \theta = 1 \quad (1.37)$$

Also, $\tan \theta = o/a = (o/h)/(a/h)$, so

$$\tan \theta = \sin \theta / \cos \theta \quad (1.38)$$

Exercise 1.15 The disc of the Sun subtends an angle of 31.9 arcmin when viewed from the Earth which is 1.50×10^{11} m away. What is the diameter of the Sun?



1.12.2 The sine rule and cosine rule

Whilst the trigonometric ratios are useful when determining the properties of right-angled triangles, they can also be extended to triangles in general, via two formulae, known as the sine rule and the cosine rule.

The sine rule states that in any triangle with sides of length a , b and c , and internal angles opposite to each of these sides of A , B and C respectively, then

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} \quad (1.39)$$

Similarly, the cosine rule may be stated as

$$c^2 = a^2 + b^2 - 2ab \cos C \quad (1.40)$$

That these equations reduce to the simple trigonometric ratios and pythagoras's theorem may be seen simply by assuming that the triangle in question is right-angled with $C = 90^\circ$ and c then becomes the hypotenuse.

I.12.3 Trigonometric functions

So far we have discussed the trigonometric ratios (\sin , \cos , \tan) in the context of acute angles ($0 < \theta < 90^\circ$) within right-angled triangles, but angles in nature are not so constrained and may take any value $-\infty < \theta < +\infty$. (For example, if you turn around twice, you have turned through an angle of $720^\circ = 4\pi$ radians.)

Trigonometric functions can be defined over the full range of angles through the concept of a unit circle (of radius 1 unit) as shown in Figure 1.12.

As the radius arm of a unit circle sweeps out an angle from $\theta = 0$ on the x -axis to $\theta = 90^\circ$ on the y -axis, the x -coordinate of the tip of the radius arm traces $\cos \theta$ and the y -coordinate traces $\sin \theta$. As θ increases beyond 90° , the x - and y -coordinates continue to trace out the cosine and sine functions, the cosine function becoming negative as x goes negative. The y -coordinate, and hence $\sin \theta$, becomes negative once θ exceeds 180° , at which time the radius arm lies just below the negative x -axis. As the radius arm continues to sweep out angles from $\theta = 0$ to 360° , the full range of the trigonometric functions is revealed. For angles greater than 360° , the functions repeat themselves. That is, the trigonometric functions are periodic with period 360° or 2π , as shown in Figure 1.13.

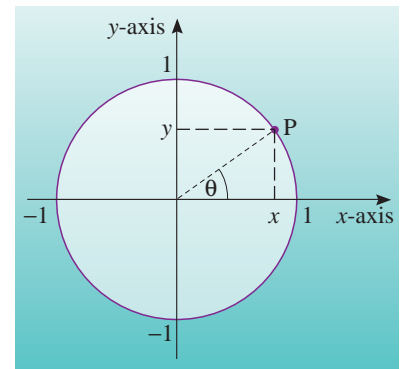


Figure 1.12 A unit circle.

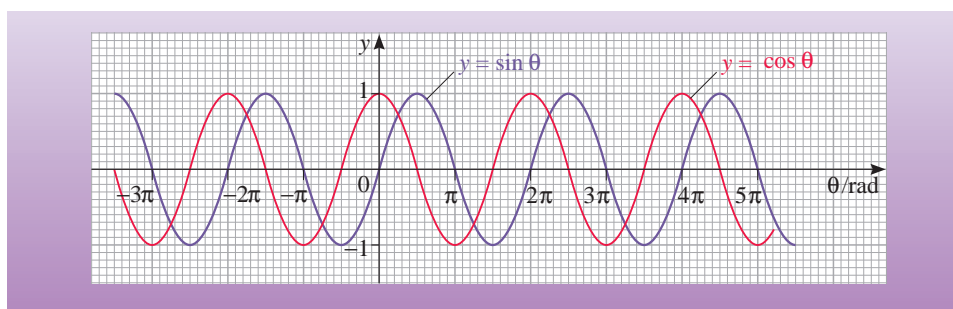


Figure 1.13 Graphs of $\sin \theta$ and $\cos \theta$ against θ .

If we plot a graph of the way one quantity varies with another and get either a **sine curve** or a **cosine curve**, we say in either case that the quantity plotted on the vertical axis varies *sinusoidally*. The *argument* is the expression or number whose sine is being computed. So the argument of $\sin x$ is x , the argument of $\sin 0.4$ is 0.4 , and so on. The argument to the trigonometric functions is often specified in radians rather than degrees. Furthermore, recall that a radian is really a

dimensionless value, originally introduced as the dimensionless ratio between the length of an arc of a circle and the radius, $\theta = s/r$. The trigonometric functions can also be thought of in this form, where the argument of the functions is not an angle in degrees but a dimensionless number possibly unrelated to triangles.

An example which illustrates sinusoidal functions is that of the voltage V of mains electricity which varies sinusoidally with time t according to the equation $V = V_{\max} \sin(2\pi ft)$, as illustrated in Figure 1.14. V_{\max} is known as the **amplitude**; this is the quantity that scales the sine curve. The argument $2\pi ft$ of the sine function includes the **frequency** f (in the UK, the frequency of the mains is 50 Hz). The greater the frequency, the shorter is the period of time before the curve repeats itself. Figure 1.13 shows that a sinusoidally varying quantity repeats exactly the same pattern of variation every time the argument of the sine function increases by 2π . In Figure 1.14, the argument $2\pi ft$ increases by 2π when ft increments by 1, i.e. when t increments by $1/f$. This time $t = 1/f$ is called the **period** of oscillation, T . Thus in Figure 1.14 the curve crosses the axis going positive at $t = 0$ (since $\sin 0 = 0$), and again, one full cycle later, at $t = 1/f$, where again $\sin(2\pi f \times 1/f) = \sin(2\pi) = 0$. Alternative expressions for a sinusoidal oscillation in time are thus:

$$y = A \sin(2\pi ft) \quad (1.41)$$

or

$$y = A \sin(2\pi t/T) \quad (1.42)$$

It is also useful to define an **angular frequency** $\omega = 2\pi f$, from which it follows that

$$y = A \sin(\omega t) \quad (1.43)$$

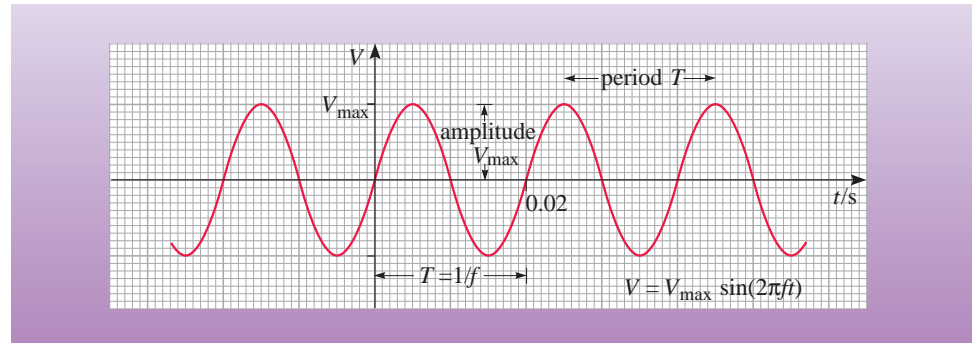


Figure 1.14 Mains voltage varies sinusoidally with time.

If an oscillation doesn't start with $y = 0$ at $t = 0$, the graphical representation of Figure 1.13 must be modified by displacing the sine curve (Figure 1.15), and the algebraic representation supplemented by adding a constant term, known as the *initial phase* or **phase constant**, ϕ (the Greek letter *phi*, pronounced 'fie'), to the argument, i.e.

$$y = A \sin(2\pi ft + \phi) \quad (1.44)$$

or

$$y = A \sin(2\pi t/T + \phi) \quad (1.45)$$

or

$$y = A \sin(\omega t + \phi) \quad (1.46)$$

Note that the phase difference between a sine function and a cosine function is $\pi/2$, i.e. $\sin \theta = \cos(\theta + \pi/2)$.

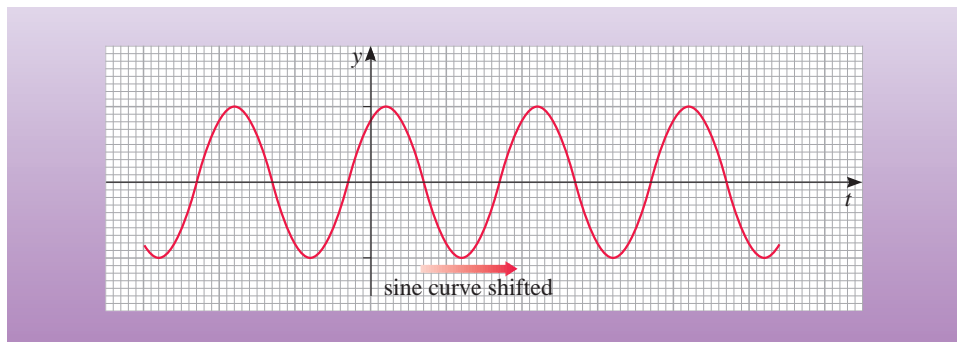


Figure 1.15 When the oscillation doesn't start with $y = 0$ at $t = 0$, the sine curve is shifted.

Exercise 1.16 The sinusoidal curve shown in Figure 1.16 represents the apparent radial speed of a star in a binary system, as derived from Doppler shift measurements (i.e. the speed relative to our line of sight). What are (a) the amplitude of the star's radial velocity, (b) the period of the motion, (c) the frequency of the motion, and (d) the angular frequency of the motion?

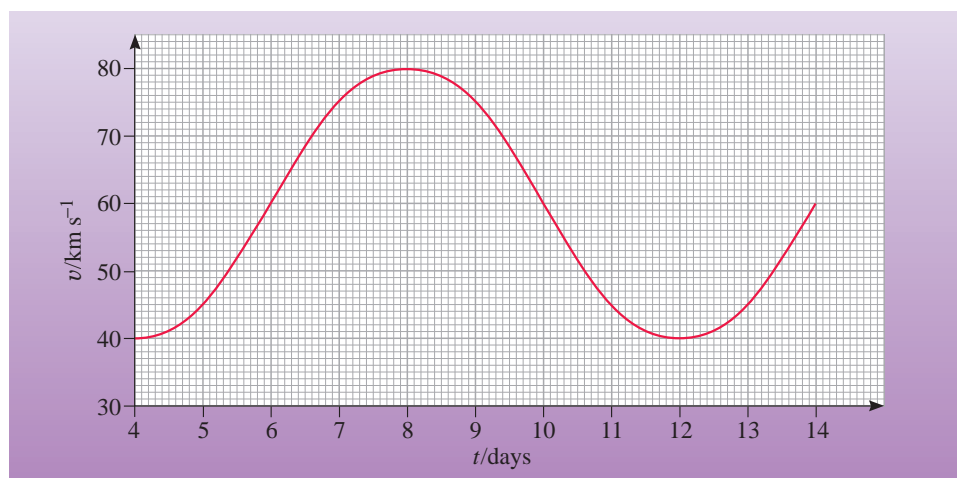


Figure 1.16 The radial velocity curve of a star referred to in Exercise 1.16.

I.12.4 Inverse trigonometric functions

Consider the general expression $\sin \theta = x$. If x is known, what value of θ satisfies the expression? We write this value as $\theta = \sin^{-1} x$ or $\arcsin x$, called the *inverse sine* function or *arcsine* function. $\sin^{-1} x$ is therefore the value whose sine is x .

This applies to the trigonometric functions, not just trigonometric ratios. The argument x takes negative as well as positive values, but clearly must be in the range $-1 \leq x \leq 1$.

Consider $\sin^{-1}(\sqrt{3}/2)$. Not only does $\sin(\pi/3) = \sqrt{3}/2$, but so does $\sin(2\pi/3)$. What then is $\sin^{-1}(\sqrt{3}/2)$? Is it $\pi/3$, $2\pi/3$, or $7\pi/3$ even? This debate is settled by the convention that the \sin^{-1} function is defined only on the range $-\pi/2 \leq \sin^{-1} x \leq \pi/2$. The *inverse tangent* or *arctangent* function $\tan^{-1} x$ or $\arctan x$ is defined on the range $-\pi/2 < \tan^{-1} x < \pi/2$, the same range as for $\sin^{-1} x$, but with any real number valid as the argument, x . The *inverse cosine* or *arccosine* function $\cos^{-1} x$ or $\arccos x$ is defined over a different range, $0 \leq \cos^{-1} x \leq \pi$, with $-1 \leq x \leq 1$.

Exercise 1.17 The sine of a particular angle may be expressed as the fraction $2/\sqrt{5}$. (a) What is the size of the angle in degrees? (b) If this is one of the angles of a right-angled triangle, what is the ratio of the lengths of the three sides to each other? (c) What is the tangent of the smallest angle in the triangle? ■

1.13 Vectors

Vectors occur in many areas of astrophysics and cosmology when describing the physical properties of systems. A **vector** is a quantity that has both magnitude and direction, such as the velocity of a star. In contrast, a **scalar** has magnitude only. For example, the temperature of a star is a scalar variable; it may vary from point to point within a star, but there is no direction associated with each measurement. A vector may be represented diagrammatically by an arrow, the length of which specifies the vector's magnitude and the direction of which is the same as the vector's direction. By convention, vectors are printed as bold, italic symbols, e.g. \mathbf{r} , while the magnitude is written as a normal italic symbol, r . *Handwritten* vector symbols are written with a wavy underline, e.g. \tilde{r} (which in the printing trade means 'make bold').

To specify a vector fully, both its magnitude (which is always positive) and its direction must be stated, e.g. ' \mathbf{F} is a force of 10 N acting vertically downwards'. The **magnitude** of \mathbf{F} may be written as $F = |\mathbf{F}| = 10 \text{ N}$, where the pair of vertical lines ($|$ $|$) surrounding the vector indicates that we take the magnitude (i.e. the positive numerical value) of the quantity.

1.13.1 Vector components

Any vector \mathbf{a} in three-dimensional space can be resolved into three mutually perpendicular **components** a_x , a_y and a_z given by

$$a_x = a \cos \theta_x \quad (1.47)$$

$$a_y = a \cos \theta_y \quad (1.48)$$

$$a_z = a \cos \theta_z \quad (1.49)$$

where θ_x , θ_y and θ_z are respectively the angles between the direction of vector \mathbf{a} and the x -, y - and z -axes. The components are written as an ordered set in brackets, e.g.

$$\mathbf{a} = (a_x, a_y, a_z) \quad (1.50)$$

and the magnitude of \mathbf{a} is given by

$$a = \sqrt{a_x^2 + a_y^2 + a_z^2} \quad (1.51)$$

In many astrophysics or cosmology problems, only two dimensions are required, so only two components are considered, as in Figure 1.17 where the vector $\mathbf{f} = (f_x, f_y)$ is illustrated.

Exercise 1.18 A vector representing the gravitational force acting on a planet orbiting a star has a magnitude of $F = 3.50 \times 10^{22}$ N and acts at an angle of 30° with respect to the x -axis of a particular coordinate system. What are the x - and y -components of the force?

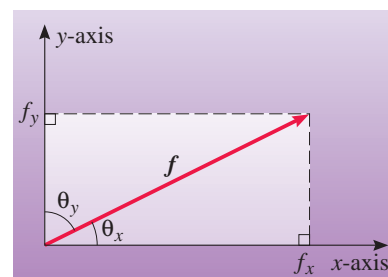


Figure 1.17 Any two-dimensional vector \mathbf{f} is characterized by two components, f_x and f_y , found by projecting perpendiculars to the x - and y -axes.

1.13.2 Addition and subtraction of vectors

For the rare case of two vectors \mathbf{a} and \mathbf{b} having the *same* direction, addition is easy: the resultant vector $\mathbf{c} = \mathbf{a} + \mathbf{b}$ is also in the same direction and of magnitude $c = a + b$. However, for the more common case of vectors with *different* directions, this simple rule does not apply, and addition must be carried out using graphical methods, such as the **triangle rule** or the **parallelogram rule**, shown in Figure 1.18 for two-dimensional vectors, or in terms of components.

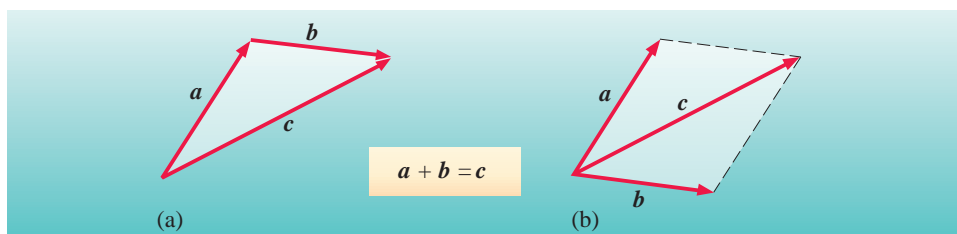


Figure 1.18 Equivalent methods of adding 2 two-dimensional vectors graphically: (a) the triangle rule for addition, (b) the parallelogram rule for addition. To sum more than two vectors, repeat the application of either rule.

Knowledge of the components greatly simplifies the addition, since if $\mathbf{c} = \mathbf{a} + \mathbf{b}$, then $c_x = a_x + b_x$, $c_y = a_y + b_y$ and $c_z = a_z + b_z$. So unless \mathbf{a} and \mathbf{b} have the same direction, $|\mathbf{c}| \neq |\mathbf{a}| + |\mathbf{b}|$. Similarly, if $\mathbf{c} = \mathbf{a} - \mathbf{b}$, then $c_x = a_x - b_x$, $c_y = a_y - b_y$ and $c_z = a_z - b_z$.

Note that a vector can always be resolved into ‘component vectors’ along arbitrary directions at right angles to each other. However, vectors are normally specified in terms of scalar components tied to the x -, y -, z -coordinate axes.

1.13.3 Position and displacement vectors

Vectors are frequently used to specify the positions of points of interest. In three dimensions, the position of a point can be specified by giving its position coordinates (x, y, z) , as shown in Figure 1.19a, but alternatively we can specify the vector \mathbf{r} from the origin to the point (x, y, z) . Since the components of \mathbf{r} are (x, y, z) , the shorthand for this is $\mathbf{r} = (x, y, z)$. The vector \mathbf{r} is known as the **position vector** of the point (x, y, z) , and its magnitude is equal to the distance of the point from the origin: $r = \sqrt{x^2 + y^2 + z^2}$.

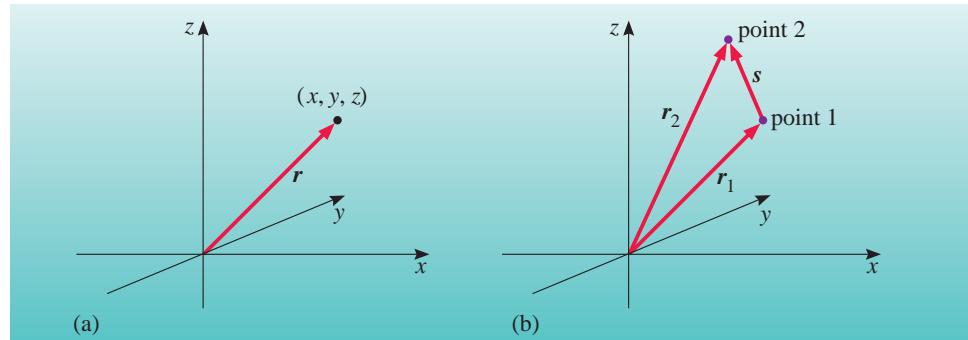


Figure 1.19 (a) The position vector \mathbf{r} defines the position of a point relative to the origin. (b) The displacement vector \mathbf{s} defines the difference in position of two points with position vectors \mathbf{r}_1 and \mathbf{r}_2 , where $\mathbf{s} = \mathbf{r}_2 - \mathbf{r}_1$.

Exercise 1.19 The position vector of the Sun from the Earth has components $a_x = 0.90 \times 10^{11}$ m and $a_y = 1.20 \times 10^{11}$ m in a particular coordinate system. What is the magnitude of this vector and what does the magnitude represent?

It is sometimes more convenient to specify the position of a point relative to another point not necessarily at the origin. In Figure 1.19b, the position of point 2 relative to point 1 is described by the vector \mathbf{s} , which is known as a **displacement vector**. Since point 1 has position vector \mathbf{r}_1 and point 2 has position vector \mathbf{r}_2 the triangle rule for addition of vectors tells us that $\mathbf{r}_1 + \mathbf{s} = \mathbf{r}_2$ or $\mathbf{s} = \mathbf{r}_2 - \mathbf{r}_1$. In general, a displacement vector is the *difference* between two position vectors.

1.13.4 Unit vectors

It is often useful to divide a vector \mathbf{r} by its own magnitude, to produce a **unit vector** $\hat{\mathbf{r}}$ defined as

$$\hat{\mathbf{r}} = \mathbf{r}/r \quad (1.52)$$

$\hat{\mathbf{r}}$ points in the same direction as \mathbf{r} , but has unit magnitude, i.e. $|\hat{\mathbf{r}}| = 1$. Note that $|\hat{\mathbf{r}}|$ is *dimensionless*, not 1 m or 1 unit for example.

- What is the effect of multiplying a scalar by a unit vector?
- The result is a vector whose *magnitude* is the same as that of the original scalar, but whose *direction* is that of the unit vector.

Now, the three components of a vector are each scalars. Consequently, unit vectors can be used to express a single vector as a sum of three mutually

perpendicular vectors. For instance, the vector \mathbf{a} has components (a_x, a_y, a_z) but it can be written as the sum of three vectors:

$$\mathbf{a} = i a_x + j a_y + k a_z \quad (1.53)$$

where i , j and k are unit vectors in the x -, y - and z -directions, respectively.

I.13.5 The scalar product

There are two completely different ways of multiplying two vectors: one produces a scalar, the other a vector. The **scalar product** (also called the ‘dot product’) of two vectors \mathbf{a} and \mathbf{b} is a scalar equal to the product of their magnitudes multiplied by the cosine of the angle between their directions:

$$\mathbf{a} \cdot \mathbf{b} = ab \cos \theta \quad (1.54)$$

An alternative expression, useful if the components of \mathbf{a} and \mathbf{b} are known, is

$$\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y + a_z b_z \quad (1.55)$$

Combining these equations, we can also express the angle between the vectors as

$$\cos \theta = \frac{a_x b_x + a_y b_y + a_z b_z}{ab} \quad (1.56)$$

where the value of θ is found from the inverse cosine (arccosine) function.

I.13.6 The vector product

The **vector product** (also called the ‘cross product’) of \mathbf{a} and \mathbf{b} is a vector with magnitude equal to the product of the magnitudes of \mathbf{a} and \mathbf{b} multiplied by the sine of the angle between them. The direction of the vector product is given by the right-hand rule, as illustrated in Figure 1.20.

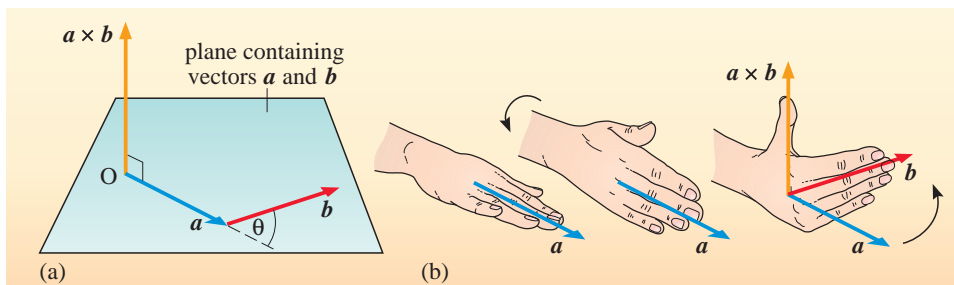


Figure 1.20 (a) The vector product. (b) The right-hand rule defines the direction of the vector product of two vectors. The palm and outstretched fingers and thumb of the right hand are aligned with the first vector \mathbf{a} , until the fingers can be bent to the direction of the second vector \mathbf{b} . The outstretched thumb then points in the direction of the vector product $\mathbf{a} \times \mathbf{b}$.

So $\mathbf{a} \times \mathbf{b}$ is a vector with magnitude $ab \sin \theta$ and direction perpendicular to both

\mathbf{a} and \mathbf{b} as given by the right-hand rule.

$$|\mathbf{a} \times \mathbf{b}| = ab \sin \theta \quad (1.57)$$

In terms of the components of the vectors,

$$\mathbf{a} \times \mathbf{b} = (a_y b_z - a_z b_y, a_z b_x - a_x b_z, a_x b_y - a_y b_x) \quad (1.58)$$

Note that the order of the vectors is unimportant in forming the scalar (dot) product, because $\cos(-\theta) = \cos \theta$, so $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$, but that the order of the vectors is crucial in forming the vector (cross) product, because $\sin(-\theta) = -\sin \theta$, so $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$. Note also that some textbooks and articles use the symbol \wedge rather than \times to indicate a vector product, but the meaning is identical. Finally, you should be aware that there is no mathematical operation defined as division by a vector, and so expressions such as \mathbf{a}/\mathbf{b} or \mathbf{a}/b are meaningless and should never be written.

Exercise 1.20 By considering the definition of the scalar and vector products, evaluate: (a) $\mathbf{a} \cdot \mathbf{a}$ (b) $\mathbf{b} \times \mathbf{b}$



1.14 Coordinates

The position of a point in space may be described by reference to a set of perpendicular x -, y -, z -axes, as shown in Figure 1.21. The (x, y, z) values of a point are called its **Cartesian coordinates**. These are also the components of the point's position vector $\mathbf{r} = (x, y, z)$.

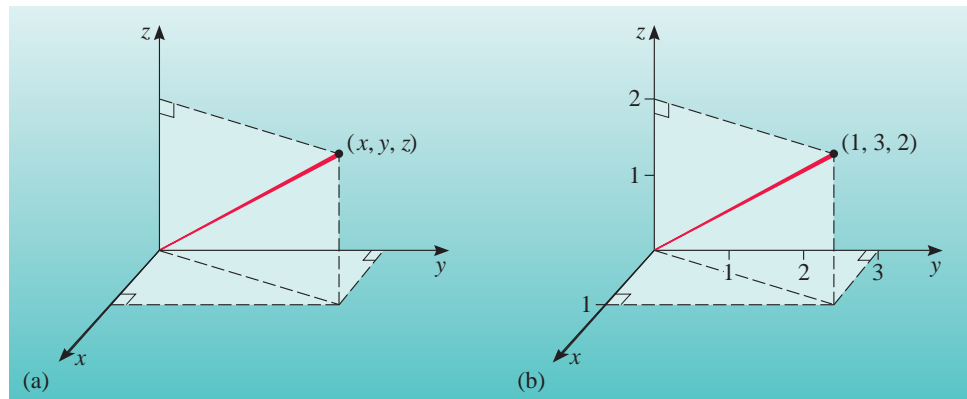


Figure 1.21 (a) The position of a point is specified by its x -, y - and z -coordinates. (b) The point for which $x = 1$, $y = 3$ and $z = 2$ has coordinates $(1, 3, 2)$.

In two dimensions, an alternative to Cartesian coordinates $\mathbf{r} = (x, y)$ is **plane polar coordinates** $\mathbf{r} = [r, \theta]$ (see Figure 1.22). Two numbers are still required to locate a point, but now the distance r and direction θ are specified rather than the x and y values. (We have also chosen to use square brackets around the pair of polar coordinates to help distinguish them from the pair of Cartesian coordinates, though this practice is neither essential nor universally adopted.) The units

attached to the two coordinate pairs also differ; Cartesian coordinates have units of length, whereas plane polar coordinates have units of length and angle.

Coordinate transformations allow conversions to be made between the coordinates expressed in one system and those in another system, albeit for the same point in space. To convert two-dimensional Cartesian coordinates into plane polar coordinates, the transformations are:

$$r = \sqrt{x^2 + y^2} \quad (1.59)$$

$$\text{and} \quad \tan \theta = y/x \quad (1.60)$$

where θ is the angle between the x -axis ($\theta = 0$) and r . Conversely, to convert plane polar coordinates into two-dimensional Cartesian coordinates,

$$x = r \cos \theta \quad (1.61)$$

$$\text{and} \quad y = r \sin \theta \quad (1.62)$$

Many other coordinate systems can be devised; Cartesian axes do not provide the only possible description of a location in space. For example, locations near the Earth's surface are often described in terms of latitude, longitude, and height above mean sea-level. In astrophysics and cosmology, Cartesian coordinates, whilst valid, often are not the most *convenient* ones to use, especially where rotational symmetry exists.

When describing disc-like structures for instance, **cylindrical coordinates** are often much more useful. In this case, the position of a point in space is described by reference to two distance coordinates r and z and an angular coordinate ϕ , as shown in Figure 1.23. From the figure, the three-dimensional Cartesian to cylindrical coordinate transformations are:

$$r = \sqrt{x^2 + y^2} \quad (1.63)$$

$$\cos \phi = x/r \quad (1.64)$$

$$\text{and} \quad z = z \quad (1.65)$$

whilst the cylindrical to three-dimensional Cartesian coordinate transformations are:

$$x = r \cos \phi \quad (1.66)$$

$$y = r \sin \phi \quad (1.67)$$

$$\text{and} \quad z = z \quad (1.68)$$

In the case of spherical symmetry, such as when dealing with stars, **spherical coordinates** are even more useful. In spherical coordinates, the position of a point in space is described by reference to a range coordinate r and two angular coordinates, the zenith angle θ and the azimuthal angle ϕ , as shown in Figure 1.24. From the figure, the three-dimensional Cartesian to spherical coordinate transformations are:

$$r = \sqrt{x^2 + y^2 + z^2} \quad (1.69)$$

$$\cos \theta = z/r \quad (1.70)$$

$$\text{and} \quad \sin \phi = y/(r \sin \theta) \quad (1.71)$$

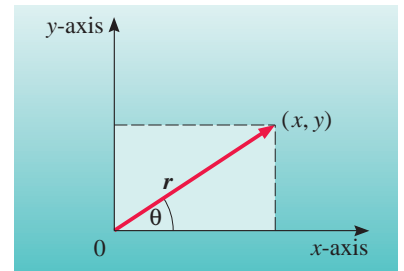


Figure 1.22 Cartesian (x, y) and plane polar $[r, \theta]$ coordinates.

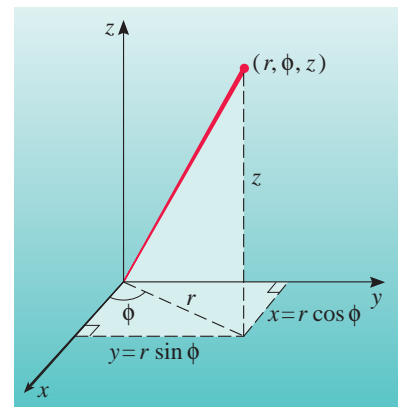


Figure 1.23 In the cylindrical coordinate system, the position of a point is specified by its r -, ϕ - and z -coordinates.

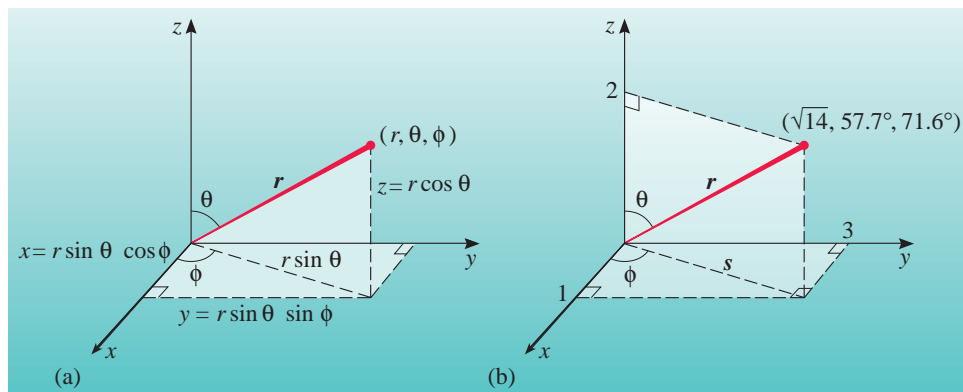
whilst the spherical to three-dimensional Cartesian coordinate transformations are:

$$x = r \sin \theta \cos \phi \quad (1.72)$$

$$y = r \sin \theta \sin \phi \quad (1.73)$$

and
$$z = r \cos \theta \quad (1.74)$$

Figure 1.24 (a) In the spherical system, the position of a point is specified by its r -, θ - and ϕ -coordinates. (b) The point for which $x = 1$, $y = 3$, $z = 2$ has spherical coordinates $(\sqrt{14}, 57.7^\circ, 71.6^\circ)$.



Exercise 1.21 In a certain binary star system at a particular instant, the Cartesian coordinates of one of the stars with respect to a set of axes centred on the other star are $(1.2 \times 10^{10} \text{ m}, 1.6 \times 10^{10} \text{ m}, 0.0 \text{ m})$. What are the spherical coordinates of this star?

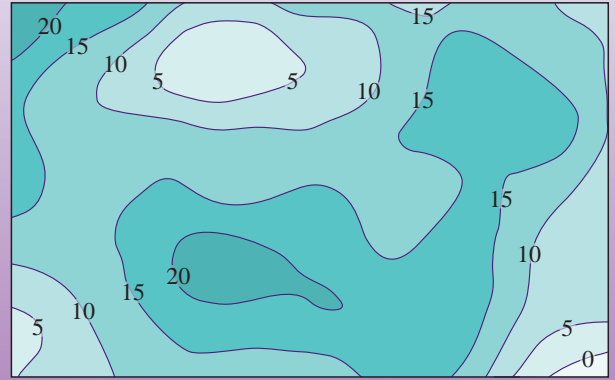


1.15 Scalar and vector fields

A field is a physical quantity that has a value at each point within a region of space. So for instance, the altitude of the landscape would constitute a two-dimensional field (Figure 1.25) and the density of gas within an interstellar cloud would constitute a three-dimensional field. In each case a *single* value (of altitude or density, etc.) can be assigned to each point in two- or three-dimensional space to represent the physical quantity. Notice that in each of these cases, the fields are **scalar fields** – the quantity at each point in space has a *magnitude* only. So a three-dimensional scalar field represented by $\rho(x, y, z)$ has a single value ρ at each point in space (where ρ is the Greek letter *rho* pronounced ‘roe’).

20	17	15	10	5	5	7	8	10	13	15	14	12	10
15	12	7	5	3	3	5	7	10	13	17	16	15	12
15	12	11	10	10	10	10	10	12	15	17	18	17	14
15	13	13	15	14	14	14	14	14	14	15	15	15	11
15	13	15	17	18	18	16	16	15	14	15	15	10	9
10	11	15	20	21	22	20	18	16	16	16	15	10	8
5	10	13	15	20	20	19	20	16	16	16	15	8	6
5	9	11	14	15	15	15	15	16	16	15	10	5	0

(a)



(b)

Figure 1.25 The altitude of the landscape is an example of a scalar field. (a) It can be represented by a value at each point on an imaginary grid, or (b) as a series of contour lines connecting locations of equal altitude.

By contrast, a **vector field** has a magnitude *and* a direction at each point in space. For instance, the wind velocity is often represented on weather maps by arrows (Figure 1.26). The size of each arrow represents the magnitude of the wind velocity (i.e. the wind speed) at each point and the direction of each arrow represents the wind direction. The usual weather map shows a two-dimensional vector field – representing the wind velocity close to the surface of the Earth. However, we could just as easily have a three-dimensional vector field representing wind velocity throughout the atmosphere.

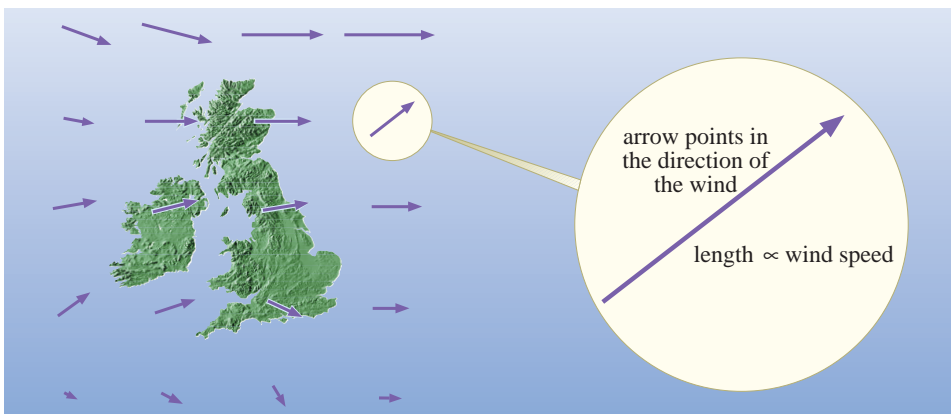


Figure 1.26 The wind velocity across the UK is an example of a vector field. The direction of each arrow represents the direction of the wind, whilst the length of each arrow represents its speed at that point.

The value of a vector field at a particular point is represented by a certain vector, but remember that a two-dimensional vector may be represented by two components (two scalars) and a three-dimensional vector may be represented by three components (three scalars). So a three-dimensional vector field represented by $\mathbf{U}(x, y, z)$ could be written in terms of its components as a vector (U_x, U_y, U_z) at *each point* in space. Pay close attention to the notation here: $\mathbf{U}(x, y, z)$ indicates that the vector field \mathbf{U} is a function of three coordinates x , y and z ; and at each point in space it can be represented by a vector whose components are (U_x, U_y, U_z) .

So, just as a three-dimensional vector can be expressed as three independent scalars – each representing one component of the corresponding vector – a

three-dimensional vector field can be expressed as a set of three independent scalar fields, such as $U_x(x, y, z)$, $U_y(x, y, z)$ and $U_z(x, y, z)$. Each scalar field gives one component of the corresponding vector field at each point in space.

Exercise 1.22 The temperature, pressure, gravitational field and magnetic field can be defined at each point within a star. Which of these are vector fields and which are scalar fields?



1.16 Matrices

A final topic in the area of manipulating numbers and symbols is that of **matrices**. These have uses in many areas of physics from quantum mechanics to general relativity, and it is in this latter context that you will meet them in cosmology.

A matrix is a set of numbers laid out in a rectangular array of rows and columns, and represented by a letter. A matrix of n rows and m columns is said to be of **order** ($n \times m$), and you should note that the elements are *always* listed as '(rows, columns)' *not* as '(columns, rows)'. So, for example, a matrix P of order (3×4) has 3 rows and 4 columns and may be written

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{pmatrix}$$

A matrix is referred to as a **square matrix** if $n = m$.

The individual quantities p_{ij} are referred to as **elements** of the matrix. Two matrices P and Q can only be equal if they are of the same order ($n \times m$) and all their corresponding elements are equal, i.e. $p_{ij} = q_{ij}$ for all values $i = 1$ to n and $j = 1$ to m .

1.16.1 Combining matrices

In order to add together two matrices, or to subtract one from another, they must be of the same order. Given two matrices P and Q of the same order, with elements p_{ij} and q_{ij} respectively, then the sum $P + Q$ is a matrix R whose elements are given by $r_{ij} = p_{ij} + q_{ij}$. Similarly, the difference $P - Q$ is a matrix S whose elements are given by $s_{ij} = p_{ij} - q_{ij}$. Both R and S are clearly of the same order as P and Q .

- What are the sum and difference of the following two matrices?

$$A = \begin{pmatrix} 7 & 11 \\ 3 & 2 \\ 4 & 8 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 4 \\ 13 & 9 \\ 7 & 2 \end{pmatrix}$$

- The sum of these two matrices is

$$A + B = \begin{pmatrix} 12 & 15 \\ 16 & 11 \\ 11 & 10 \end{pmatrix}$$

The difference of these two matrices is

$$A - B = \begin{pmatrix} 2 & 7 \\ -10 & -7 \\ -3 & 6 \end{pmatrix}$$

Multiplying, or dividing, a matrix by a number (often referred to as a scalar) simply entails multiplying, or dividing, each element of the matrix by that number. So given a matrix P with elements p_{ij} and a scalar k , the matrix kP has elements with values kp_{ij} , and the matrix P/k has elements with values p_{ij}/k . Multiplication of a matrix by a scalar is clearly commutative, that is, $kP = Pk$.

Multiplying one matrix by another is also possible. However, a matrix P can only be multiplied by a matrix Q if the number of columns of P is equal to the number of rows of Q . Matrices for which this is possible are called **conformable**. Given a matrix P of order $(a \times b)$ with elements p_{ik} and a matrix Q of order $(b \times c)$ with elements q_{kj} , then their product PQ is a matrix R of order $(a \times c)$ with elements r_{ij} that are given by the summation over all k as follows

$$r_{ij} = \sum_{k=1}^{k=b} p_{ik}q_{kj}$$

An example should make the process clearer.

Worked Example 1.8

Suppose P is a matrix of order (2×3) and Q is a matrix of order (3×2) , shown below, what is the product PQ ?

$$P = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \quad Q = \begin{pmatrix} 10 & 11 \\ 14 & 15 \\ 18 & 19 \end{pmatrix}$$

Solution

Since the number of columns of P is equal to the number of rows of Q (i.e. both are 3), the two matrices are conformable, and matrix multiplication is possible. The product PQ is therefore a matrix of order (2×2) with elements as follows:

$$PQ = \begin{pmatrix} (1 \times 10) + (2 \times 14) + (3 \times 18) & (1 \times 11) + (2 \times 15) + (3 \times 19) \\ (4 \times 10) + (5 \times 14) + (6 \times 18) & (4 \times 11) + (5 \times 15) + (6 \times 19) \end{pmatrix}$$

$$PQ = \begin{pmatrix} 92 & 98 \\ 218 & 233 \end{pmatrix}$$

Essential skill:
Multiplying matrices

Note that matrix multiplication is not commutative. That is to say given two matrices P and Q of order $(a \times b)$ and $(b \times a)$ respectively, whilst PQ and QP are both conformable, $PQ \neq QP$. If P and Q are not square matrices, this is obvious, since PQ will be a matrix of order $(a \times a)$ and QP will be a matrix of order $(b \times b)$, so they cannot possibly be equal. However, the result is still not commutative even if P and Q are square matrices as the following exercise demonstrates.

Exercise 1.23 Given two square matrices P and Q as defined below, what are the matrices defined by PQ and QP ?

$$P = \begin{pmatrix} 1 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \quad Q = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$



1.16.2 Special types of matrices

From the previous section on vectors, it may be apparent that a single row or column of a matrix has a lot in common with a vector. Both comprise a set of elements and there exist shorthand ways of expressing all the elements of a vector or a matrix. In fact, a set of m elements arranged in a single row (i.e. a matrix of order $(1 \times m)$) is called a row matrix or a **row vector** and may be written as

$$[P] = (p_1, p_2, \dots, p_m)$$

Similarly, a set of n elements arranged in a single column (i.e. a matrix of order $(n \times 1)$) is called a column matrix or a **column vector** and may be written as

$$\{Q\} = \begin{pmatrix} q_1 \\ q_2 \\ \cdot \\ \cdot \\ \cdot \\ q_n \end{pmatrix}$$

- What is the result of multiplying a row vector $[P]$ of order $(1 \times m)$ by a column vector $\{Q\}$ of order $(m \times 1)$?
- The result is a vector of order (1×1) (i.e. a scalar) whose value is given by $(p_1q_1 + p_2q_2 + p_3q_3 + \dots + p_mq_m)$.
- What is the order of the vector obtained when multiplying a column vector $\{Q\}$ of order $(n \times 1)$ by a row vector $[P]$ of order $(1 \times n)$?
- The result is a square matrix of order $(n \times n)$.

Three other special names for matrices are the null, diagonal and unit matrices. A **null matrix** has all its elements equal to zero. A **diagonal matrix** is a square matrix, all of whose elements are zero *except* those in the leading diagonal. So a matrix P is diagonal if all its elements $p_{ij} = 0$ as long as $i \neq j$. For example the following is a diagonal matrix of order (3×3) :

$$P = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 15 \end{pmatrix}$$

An important point about diagonal matrices is that two diagonal matrices of the same order commute when multiplied together, i.e. $PQ = QP$ if both P and Q are diagonal matrices of the same order.

Finally, a **unit matrix** or **identity matrix** is a diagonal matrix whose elements on the leading diagonal are all equal to 1. Such a matrix is often represented by the symbol I . Multiplying any square matrix by a unit matrix leaves it unchanged, i.e. $IP = PI = P$, where P is a square matrix of the same order as the unit matrix I .

I.16.3 Transposing matrices

A **transposed matrix** is the result of interchanging the rows and columns of a matrix. For example, if P is a matrix of order (2×3) given by

$$P = \begin{pmatrix} 5 & 10 & 15 \\ 20 & 25 & 30 \end{pmatrix}$$

then its transpose, denoted by P' , is a matrix of order (3×2) given by

$$P' = \begin{pmatrix} 5 & 20 \\ 10 & 25 \\ 15 & 30 \end{pmatrix}$$

- What are the transpose of (i) a column vector, and (ii) a row vector?
- The transpose of a column vector is a row vector, and the transpose of a row vector is a column vector, i.e. $\{P\}' = [P]$ and $[P]' = \{P\}$.

A final result about transposing matrices to state is that if P and Q are two matrices that are conformable such that $PQ = R$, then the transpose of R , i.e. $R' = (PQ)'$ is equal to $Q'P'$. In other words, if we take the transpose of the product of two matrices, this is equal to the product of the two transposed matrices with the order reversed: $(PQ)' = Q'P'$.

I.16.4 The determinant of a matrix

The **determinant** is a number that may be calculated for any square matrix. For a square matrix of order 2×2 as follows:

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

the determinant, indicated by $|P|$, is simply $(p_{11} \times p_{22}) - (p_{21} \times p_{12})$. For a square matrix of order 3×3 :

$$Q = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{pmatrix}$$

the determinant is $|Q| = q_{11}(q_{22} \times q_{33} - q_{32} \times q_{23}) - q_{21}(q_{12} \times q_{33} - q_{32} \times q_{13}) + q_{31}(q_{12} \times q_{23} - q_{22} \times q_{13})$. The procedure may be extended to square matrices of higher orders, but becomes tedious to calculate by hand.

When the determinant of a square matrix is zero, it is referred to as a **singular** matrix. In fact, any matrix for which any two rows, or any two columns, are identical will have a determinant of zero, and so is singular.

Exercise 1.24 What are the determinants of the following matrices?

$$(a) \quad A = \begin{pmatrix} 2 & 3 \\ 4 & 5 \end{pmatrix} \quad (b) \quad B = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 3 & 4 & 3 \end{pmatrix}$$

Note also that the value of a determinant is unchanged if the elements of all the rows are interchanged with the elements of all the columns. This is equivalent to saying that the determinant of a matrix is the same as the determinant of the transpose of that matrix, i.e. $|P| = |P'|$. However, the sign of a determinant is reversed if any two of its columns, or any two of its rows, are interchanged.

1.16.5 Adjoint and reciprocal matrices

It is often useful to obtain the reciprocal of a matrix, i.e. P^{-1} such that

$$PP^{-1} = P^{-1}P = I \quad (1.75)$$

where I is the unit matrix. The reciprocal of a matrix is defined as the adjoint of a matrix divided by its determinant, i.e.

$$P^{-1} = \text{adj } P / |P| \quad (1.76)$$

where the **adjoint** of a matrix is the transpose of the matrix of its cofactors.

The **cofactor** of an element p_{ij} of a matrix P is found by (i) crossing out the entries that lie in the corresponding row i and column j ; (ii) rewriting the matrix without the marked entries and (iii) finding the determinant of this new matrix. Then if $i + j$ is an even number, the cofactor of p_{ij} is equal to the value found, but if $i + j$ is an odd number, the cofactor of p_{ij} is equal to minus the value found. An example should make the process clear.

Worked Example 1.9

(a) What is the reciprocal P^{-1} of the matrix given by

$$P = \begin{pmatrix} 1 & 6 & 5 \\ 2 & 5 & 1 \\ 3 & 4 & 3 \end{pmatrix}$$

(b) Verify that $PP^{-1} = I$.

Solution

(a) First we find the cofactors of each element. We start with the element $p_{11} = 1$. The cofactor of this is the determinant of the 2×2 matrix remaining after we cross out row 1 and column 1. So the cofactor of this element is $(5 \times 3) - (4 \times 1) = 11$. Since $i + j = 1 + 1 = 2$ is even, the sign remains as calculated.

If we next consider the element $p_{12} = 6$, the cofactor is the determinant of the 2×2 matrix remaining after we cross out row 1 and column 2. So the

Essential skill:

Finding the reciprocal of a matrix

cofactor of this element is $(2 \times 3) - (3 \times 1) = 3$. Since $i + j = 1 + 2 = 3$ is odd, the sign is reversed, and so the cofactor is actually -3 .

Similarly, we can calculate all the other cofactors, and end up with the matrix of cofactors:

$$\begin{pmatrix} 11 & -3 & -7 \\ 2 & -12 & 14 \\ -19 & 9 & -7 \end{pmatrix}$$

Next we take the transpose of this matrix, and the result is the adjoint of the original matrix, i.e.

$$\text{adj } P = \begin{pmatrix} 11 & 2 & -19 \\ -3 & -12 & 9 \\ -7 & 14 & -7 \end{pmatrix}$$

The determinant of the original matrix is

$$|P| = 1(5 \times 3 - 4 \times 1) - 2(6 \times 3 - 4 \times 5) + 3(6 \times 1 - 5 \times 5) = 11 + 4 - 57 = -42.$$

So the reciprocal of the original matrix is

$$P^{-1} = \frac{\text{adj } P}{|P|} = \begin{pmatrix} -11/42 & -2/42 & 19/42 \\ 3/42 & 12/42 & -9/42 \\ 7/42 & -14/42 & 7/42 \end{pmatrix}$$

(b) Multiplying the original matrix by its inverse, we get

$$PP^{-1} = \begin{pmatrix} 1 & 6 & 5 \\ 2 & 5 & 1 \\ 3 & 4 & 3 \end{pmatrix} \times \begin{pmatrix} -11/42 & -2/42 & 19/42 \\ 3/42 & 12/42 & -9/42 \\ 7/42 & -14/42 & 7/42 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{(1 \times -11) + (6 \times 3) + (5 \times 7)}{42} & \frac{(1 \times -2) + (6 \times 12) + (5 \times -14)}{42} & \frac{(1 \times 19) + (6 \times -9) + (5 \times 7)}{42} \\ \frac{(2 \times -11) + (5 \times 3) + (1 \times 7)}{42} & \frac{(2 \times -2) + (5 \times 12) + (1 \times -14)}{42} & \frac{(2 \times 19) + (5 \times -9) + (1 \times 7)}{42} \\ \frac{(3 \times -11) + (4 \times 3) + (3 \times 7)}{42} & \frac{(3 \times -2) + (4 \times 12) + (3 \times -14)}{42} & \frac{(3 \times 19) + (4 \times -9) + (3 \times 7)}{42} \end{pmatrix}$$

$$PP^{-1} = \begin{pmatrix} 42/42 & 0/42 & 0/42 \\ 0/42 & 42/42 & 0/42 \\ 0/42 & 0/42 & 42/42 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

which is the unit matrix, as required.

Summary of Chapter 1

1. Physical quantities are commonly represented by algebraic symbols. The symbol comprises two parts: a numerical value and an appropriate unit. Any equations involving physical quantities must have the same units on both sides.
2. Dimensional analysis may be used to verify that the units on both sides of an

equation are the same.

3. The rearrangement of numerical or algebraic equations is accomplished by following the rule that whatever operation is carried out on one side of the equals sign must also be carried out on the other. Algebraic fractions are manipulated in exactly the same way as numerical fractions.
4. In order to solve simultaneous equations it is always necessary to have as many equations as there are unknowns.
5. The following rules illustrate how powers of numbers are manipulated:

$$y^a \times y^b = y^{a+b}$$

$$y^{-a} = 1/y^a$$

$$y^a / y^b = y^{a-b}$$

$$y^{1/n} = \sqrt[n]{y}$$

$$(y^a)^b = y^{ab}$$

6. A quadratic equation of the form $ax^2 + bx + c = 0$ generally has two solutions given by $x = (-b \pm \sqrt{b^2 - 4ac})/2a$.
7. The imaginary unit i is defined such that $i \times i = -1$.
8. If a variable x is a function of another variable t , we may in general write $x = f(t)$, where t is the argument of the function.
9. Any number may be written in scientific notation as a decimal number between 1 and 10 multiplied by 10 raised to some integer power.
10. The number of accurately known digits in the value of a physical quantity, plus one uncertain digit, is called the number of significant figures. Leading zeros do not count as significant figures.
11. If two or more quantities are combined, the result is known only to the same number of significant figures as the least precisely known quantity.
12. Random uncertainties affect the precision of a measurement; systematic uncertainties affect the accuracy of a measurement. Random uncertainties may be estimated by repeating measurements. The best estimate of the measurement is the mean value: $\langle x \rangle = \sum x_i/n$ and the size of the random uncertainty in any individual measurement is about 2/3 of the spread of the measurements.
13. The standard deviation s_n of a set of measured values x_i is the square root of the mean of the squares of the deviations of the measured values from their mean value:

$$s_n = \sqrt{\frac{\sum (x_i - \langle x \rangle)^2}{n}}$$

In the limit of many measurements, the typical distribution of a set of measurements will follow a Gaussian (normal) distribution. 68% of the measurements will lie within ± 1 standard deviation of the mean value.

14. When counting randomly fluctuating events, the uncertainty in the number of events is given by the square root of the number of events.

15. The uncertainty in the mean value of a set of n measurements that have a standard deviation of s_n is $\sigma_m = s_n/\sqrt{n}$.
16. Any number can be expressed as a power of ten. The power to which 10 is raised is called the logarithm to the base ten of the resulting number, i.e. if $x = 10^a$ then $\log_{10} x = a$. Logarithms to the base e ($= 2.718\dots$) are called natural logarithms, i.e. if $y = e^b$ then $\log_e y = b$ (also written as $\ln y$).
17. Logarithms (to any base) may be combined by applying the following rules:

$$\log(a \times b) = \log a + \log b$$

$$\log(a/b) = \log a - \log b$$

$$\log a^b = b \log a$$

18. If two quantities are directly proportional to one another, then a graph of one quantity plotted against the other will yield a straight line passing through the origin. In general the equation of a straight line graph has the form

$$y = mx + c$$

where m is the gradient of the graph and c is its intercept with the y -axis.

19. If a relationship obeys a power law, such as $y = ax^b$, then a graph of $\log y$ against $\log x$ will be a straight line whose gradient is b and whose intercept is $\log a$.
20. Plane angles may be measured in degrees or radians. An arc whose length is equal to the circumference of a circle subtends an angle of 360° or 2π radians. Solid angles are measured in steradians. The surface of a sphere subtends a solid angle of 4π steradians.
21. The trigonometric ratios may be defined using a right-angled triangle (such as that shown in Figure 1.27) as follows:

$$\sin \theta = o/h, \quad \cos \theta = a/h, \quad \tan \theta = o/a$$

22. The small-angle approximation is such that when θ is small and in radians, $\sin \theta \approx \tan \theta \approx \theta$ and $\cos \theta \approx 1$.
23. Trigonometric functions are periodic with period 360° or 2π radians. If we plot a graph of the way one quantity varies with another and get either a sine curve or a cosine curve, we say in either case that the quantity plotted on the vertical axis varies sinusoidally. The argument of the functions is *not* an angle in degrees but a dimensionless number possibly unrelated to triangles.
24. A general expression for a sinusoidal function of time is $y = A \sin(\omega t + \phi)$, where A is the amplitude of the function, the angular frequency is $\omega = 2\pi f = 2\pi/T$ and ϕ is the initial phase of the function at $t = 0$. The frequency and period of this function are related by $f = 1/T$.
25. A vector has both a magnitude and a direction and may be represented by three mutually perpendicular (Cartesian) components, e.g. $\mathbf{a} = (a_x, a_y, a_z)$. The magnitude of such a vector is given by $a = \sqrt{a_x^2 + a_y^2 + a_z^2}$.
26. Addition or subtraction of vectors may be accomplished by adding or subtracting the individual components (scalars) or graphically using the triangle or parallelogram rule.

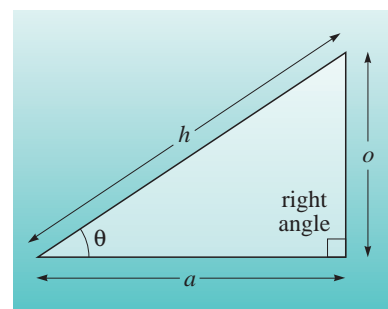


Figure 1.27 A right-angled triangle.

27. A unit vector is the result of dividing a vector by its own magnitude, i.e. $\hat{\mathbf{r}} = \mathbf{r}/r$. If a scalar is multiplied by a unit vector, the result is a vector whose magnitude is that of the original scalar and whose direction is that of the unit vector.
28. The scalar product of two vectors is given by $\mathbf{a} \cdot \mathbf{b} = ab \cos \theta = a_x b_x + a_y b_y + a_z b_z$, where θ is the angle between \mathbf{a} and \mathbf{b} . The result is a scalar. The vector product of two vectors has a magnitude given by $|\mathbf{a} \times \mathbf{b}| = ab \sin \theta$, where θ is the angle between \mathbf{a} and \mathbf{b} . The result is a vector which points in the direction given by the right-hand rule. In terms of components, $\mathbf{a} \times \mathbf{b} = (a_y b_z - a_z b_y, a_z b_x - a_x b_z, a_x b_y - a_y b_x)$.
29. Apart from Cartesian coordinates, other types of coordinate systems include plane polar, spherical and cylindrical coordinates. Trigonometry may be used to establish transformations between one system and another.
30. A scalar field is a physical quantity that has a definite value at each point within a particular region of space. A vector field describes a physical quantity that possesses both a magnitude *and* a direction at each point in space.
31. A matrix is a rectangular array of numbers arranged in rows and columns. Matrices can be added or subtracted if they are of the same order, and may be multiplied or divided by a scalar, to produce another matrix of the same order.
32. One matrix may be multiplied by another matrix if they are conformable, that is, if the number of columns of the first matrix is equal to the number of rows of the second matrix.
33. A row vector $[P]$ consists of a single row of elements, and a column vector $\{Q\}$ consists of a single column of elements.
34. A matrix is transposed by switching its row and columns. The transpose of the product of two conformable matrices is equal to the product of the two transposed matrices with the order reversed: $(PQ)' = Q'P'$. The determinant of a matrix and its transpose are identical.
35. The reciprocal P^{-1} of a matrix P is such that $PP^{-1} = I$, the unit (or identity) matrix. The reciprocal of a matrix is defined as its adjoint divided by its determinant, i.e. $P^{-1} = \text{adj } P / |P|$, where the adjoint of a matrix is the transpose of the matrix of its cofactors.

Chapter 2 Stars and planets

Introduction

This chapter will allow you to revise and consolidate your knowledge of the astrophysics of stars and the planets that orbit them. If you have recently completed the OU's Level 2 astronomy and planetary science courses (S282, S283 and SXR208), then a large part of this chapter will be familiar to you, but perhaps not all of it.

2.1 Measuring stars and planets

Since stars and planets are the building blocks of the Universe, we begin this brief review by examining just what stars and planets are and how astronomers go about measuring their physical characteristics.

A **star** is a luminous body, composed chiefly of hydrogen and helium, that generates energy through thermonuclear fusion processes occurring in its interior. Our Sun is a typical example of a star. The three main parameters determining the behaviour of stars are their mass, age and composition, and it is various combinations of these parameters which give rise to a star's other measurable characteristics, such as its temperature, luminosity and radius.

- What are the things that astronomers can measure which come from stars and planets around other stars?
- Virtually the only thing astronomers can measure is the light, or other electromagnetic radiation, which stars emit or which planets reflect. Some particles (such as neutrinos or cosmic rays) may be emitted by certain objects and in the future it may become possible to measure gravitational radiation too.

Of course astronomers can investigate light and other electromagnetic radiation in a number of ways – looking at its intensity, its spectrum, how it varies with time, and so on. But at the root of all measurements of stars is the measurement of light.

A **planet** is a gaseous, rocky or icy body which orbits a star. Our own Solar System consists of 4 major rocky planets and 4 gas giant planets, most of which are themselves orbited by a number of rocky or icy moons; an inner asteroid belt of small rocky bodies; and an outer Edgeworth–Kuiper belt of small icy bodies; plus a distant Oort cloud of comets. Several hundred planets around other stars are now known, but these have all been discovered by looking for the effect they have on the light emitted by their parent star.

Star names

In antiquity, many bright stars were given individual names and some of the names we still use today are derived from ancient Arabic names. Examples include *Betelgeuse* in the constellation of Orion and *Algol* in the constellation of Perseus.

A more convenient way of naming stars was introduced in 1603 by Johann Bayer. He suggested naming stars in a constellation in order of brightness using letters of the Greek alphabet and the genitive of the Latin name of the constellation. So the brightest star in Orion is *Alpha Orionis* (α Ori), the second brightest is *Beta Orionis* (β Ori), and so on. When the 24 Greek letters have been used up, the 26 lower case Roman letters a ... z are used, then the upper case letters A ... Q.

Fainter stars, usually invisible to the naked eye, are usually just referred to by their number in a particular catalogue. The first such catalogue was that compiled by John Flamsteed in 1725, but this has been superseded by more extensive catalogues as yet fainter and fainter stars are recorded. Thus, the seventy-seven thousand five hundred and eighty-first star in the Henry Draper Catalog (which happens to be a seventh magnitude star in the constellation of Vela) is referred to as HD77581. Other catalogues commonly encountered are the Bright Star Catalog (BS numbers) the Bonner Durchmusterung (BD numbers) and the Córdoba Durchmusterung (CD numbers). One of the largest star catalogues so far compiled is the Hubble Space Telescope Guide Star Catalog (GSC numbers), which contains over 15 million stars brighter than sixteenth magnitude.

A slightly different system is adopted for many variable stars. The letters R ... Z (unused in the naming of normal stars) are used for the first nine variable stars in a constellation. Then the lettering follows the sequence RR ... RZ, SS ... SZ, TT ... TZ up to ZZ. If yet more variable star names are needed, the sequence reverts to AA ... AZ, BB ... BZ, up to QZ which would be the 334th variable star in a constellation (leaving out those with the letter J to avoid confusion with I). After that, variable stars are numbered V335, V336, etc. which is fortunately a limitless sequence! So for example, V709 Cas is a fourteenth-magnitude variable star in the constellation of Cassiopeia, and the 709th such variable star to be identified in that constellation.

Another system of nomenclature is often used when dealing with designations in catalogues constructed in different parts of the electromagnetic spectrum, such as in X-ray astronomy. Here the convention is often to name objects according to their position in the sky (see Section 2.3) in terms of right ascension and declination. Thus the X-ray source corresponding to the variable star V709 Cas mentioned above is also known as RX J0028.8+5917. This designation informs us that the star appears in the catalogue of X-ray sources discovered by *ROSAT* (*Röntgenstrahlung Satellit*) and is located at a right ascension of 00 h 28.8 m and a declination of $+59^{\circ}17'$. The 'J' indicates that the coordinates are assigned in a system known as J2000.0 (rather than the previously used B1950.0).

As you will appreciate, most stars can have a variety of different names in different catalogues. As an example, the following list shows names currently associated with the interacting binary star Vela X-1 (the first X-ray source discovered in the constellation of Vela).

GP Vel	Vela X-1	3A 0900–403
CPC 0 6891	CPD-40 3072	GC 12502
GEN# +1.00077581	GX 263+03	H 0900–403
Hbg 267	HD 77581	HIC 44368
JP11 1751	LS 1227	1M 0900–403
PPM 313886	SAO 220767	SBC 366
TD1 13466	2U 0900–40	3U 0900–40
UBV 8734	UBV M 15041	1XRS 09002–403
GSC 07681–02303	uvby98 100077581	CD-40 4838
GCRV 25807	1H 0859–403	HIP 44368
MCW 1168	SKY# 17441	4U 0900–40
TYC 7681 2303 1		

2.2 Units in astrophysics

Although astrophysicists and cosmologists do use conventional SI (Système International) units based on the metre, kilogram and second, they also sometimes use units based on the cgs (centimetre gram second) system instead. So in the books or scientific papers that you read you may come across, for instance, the radius of a star expressed in centimetres and the rate of transfer of mass from one star to another expressed in grams per second. Even more unfamiliar may be the use of the erg as the cgs unit of energy. 1 erg may be defined as the kinetic energy possessed by an object of mass 2 g moving at a speed of 1 cm s^{-1} , from which it may be deduced that $1 \text{ erg} = 10^{-7} \text{ joules}$ (see Section 5.4). Unfortunately this is just the way things are, and although it would be very nice if all scientists always used SI units, in the real world things are not always that well organized!

Astrophysicists and cosmologists sometimes find it convenient to use other units in certain situations. For instance, when talking about the energies involved in atomic transitions, the electronvolt (eV) is a convenient unit; when discussing wavelengths of light the ångström (Å) is often used; and when talking about astronomical distances the astronomical unit (AU) or parsec (pc) are useful. A summary of conversion factors between some of these alternative units is given in the Appendices to this document.

Another feature you will find is that astrophysical quantities are often expressed in terms of units relative to the Sun. The mass, radius and luminosity of the Sun are usually denoted by the symbols M_{\odot} , R_{\odot} and L_{\odot} respectively, where $M_{\odot} = 1.99 \times 10^{30} \text{ kg}$, $R_{\odot} = 6.96 \times 10^8 \text{ m}$ and $L_{\odot} = 3.83 \times 10^{26} \text{ W}$. So a star which has a mass 12 times that of the Sun would have a mass $M = 12 M_{\odot}$, whilst another star with a luminosity of fifty-thousand times that of the Sun would have a luminosity $L = 5 \times 10^4 L_{\odot}$.

- What is the radius in metres of a star whose radius is given as $R = 0.3 R_{\odot}$?
- The radius of the star is $R = 0.3 \times 6.96 \times 10^8 \text{ m} = 2.1 \times 10^8 \text{ m}$.

Finally, another notation that is sometimes encountered in astrophysics or cosmology is to express variables in terms of a power of ten and a particular unit. For instance, the radius of a white dwarf star may be written as simply

$R_9 = 7.5$ where (in this case) R_9 is defined as the radius of the star (R) divided by 10^9 cm; that is $R_9 = R/10^9$ cm. You can think of ‘ R_9 ’ as ‘the radius of the star in units of 10^9 cm’. In this case the radius of the star is therefore $R = R_9 \times 10^9$ cm = 7.5×10^9 cm. Notice that the number represented by R_9 is a dimensionless number, it has no units, and notice also that R_9 is defined in this case using a cgs basis, you will need to be careful to check whether quantities are defined in terms of cgs or SI units in each case you come across.

- The variable M_1 is defined as the mass of a star divided by the mass of the Sun, that is, $M_1 = M/M_\odot = M/(1.99 \times 10^{30}$ kg). If a particular star has a mass given by $M_1 = 3.2$, what is its mass in kg?
- $M = M_1 \times M_\odot = 3.2 \times 1.99 \times 10^{30}$ kg = 6.4×10^{30} kg.

2.3 Positions, distances and velocities

In this section we consider how the positions of the stars may be quantified and how precise measurements of their positions may be used to infer their distances and velocities, in some cases.

2.3.1 Observing the positions of stars

The stars appear to us as (almost) fixed points of light scattered on the background sky. To quantify the positions of celestial objects, a system of coordinates is adopted which is reminiscent of the system of latitude and longitude on the surface of the Earth. The celestial equivalents are called right ascension (RA) and declination (Dec), as shown in Figure 2.1.

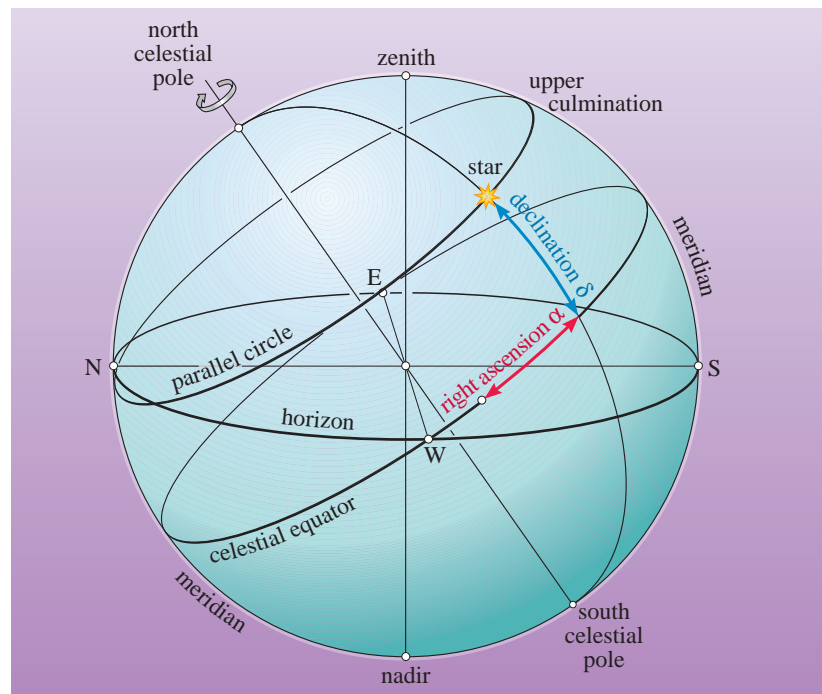


Figure 2.1 Celestial coordinates are defined in terms of right ascension and declination.

Declination, represented by the Greek lower case letter *delta* δ , is comparable to terrestrial latitude. The celestial equator is defined by a projection of the Earth's equator out into space. Positions north of the celestial equator are assigned positive values and are measured in degrees from 0° at the equator to $+90^\circ$ at the north celestial pole. The Pole Star (Polaris or Alpha Ursa Minoris) is close to the north celestial pole. Positions south of the celestial equator are assigned negative values, such that the south celestial pole is at -90° . Subdivisions in declination are measured in arc minutes and arc seconds.

Right ascension, represented by the Greek letter *alpha* α , is comparable to terrestrial longitude. The equivalent of the Greenwich Meridian, the zero point for celestial right ascension, is a circle perpendicular to the celestial equator which passes through the point where the Sun crosses the celestial equator at the spring equinox. Right ascension increases to the east of this line and is measured in hours. Consequently, RA increases from 0 h (running through constellations such as Cassiopeia, Andromeda and Pisces) through to 12 h (running through constellations such as Ursa Major, Virgo and Centaurus) before coming full circle back to 24 h which coincides with the 0 h line. Subdivisions in right ascension are measured in minutes and seconds, where 1 hour (1 h) = 60 minutes, and 1 minute (1 min) = 60 seconds (60 s). An angular distance of 1 h on the celestial equator corresponds to 15° (since $360^\circ/24 \text{ h} = 15 \text{ degrees per hour}$).



about 20°

Figure 2.2 Distances on the sky between objects are quantified in terms of their angular separation. The constellation of Orion spans about 20° from Orion's shoulder to his knee. (© Till Credner, AlltheSky.com)

Distances between stars and other objects, as they appear on the sky, are quantified in terms of their angular separation (Figure 2.2), measured in degrees or subdivisions of degrees. A constellation may extend over many tens of degrees; the full Moon or the Sun each subtend an angle of about half a degree, or 30 arc minutes ($30'$) as seen from the Earth; whilst the smallest angular size discernible to the naked eye is about $1'$; and the best angular resolution obtained with Earth-bound optical telescopes is less than 1 arc second ($1''$). Whilst the separation of objects on the sky are often described in terms of the angle between them, areas of the sky are often described in terms of square degrees, square arc minutes or square arc seconds.

Figure 2.3 shows a map of part of the sky with a grid of celestial coordinates superimposed on it. This is near to the north celestial pole, so you should notice how lines of constant right ascension converge towards the north. Also notice that with north at the top (increasing declination upwards), east is to the left (increasing right ascension leftwards). This is the opposite of terrestrial maps, where east would be to the right and longitude increases rightwards. The fact that east and west are reversed is a consequence of the fact that celestial maps are looking 'outwards' (away from the Earth) whereas terrestrial maps are looking 'inwards' (towards the Earth).

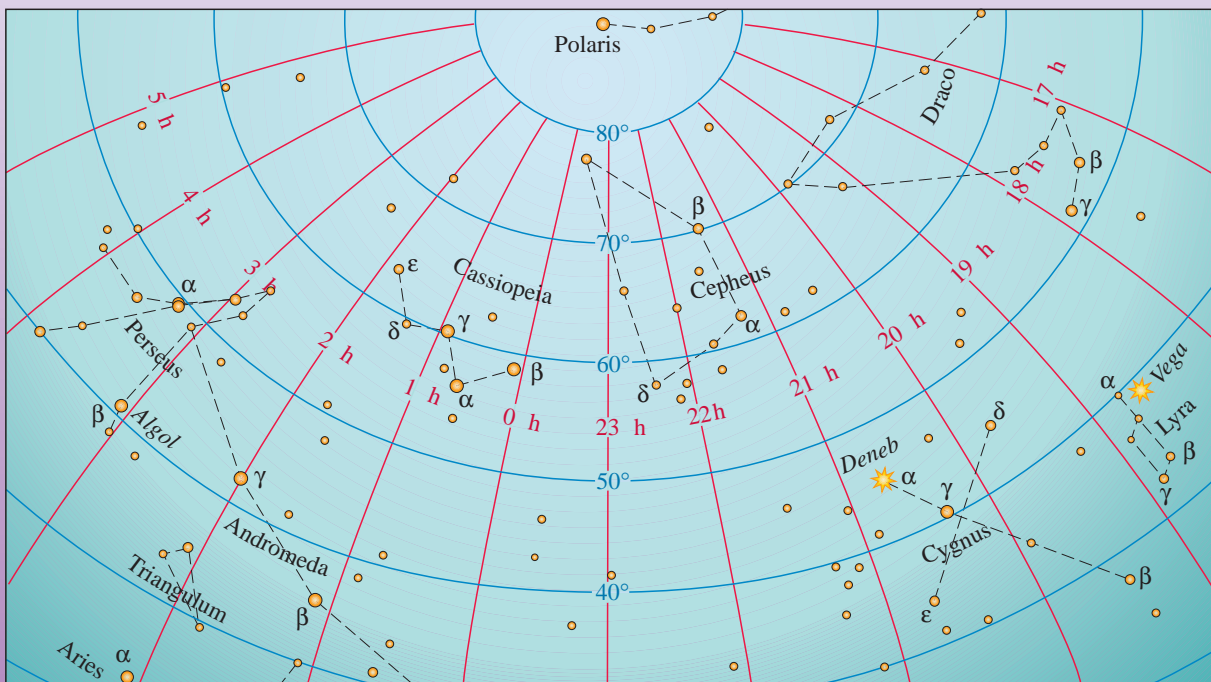


Figure 2.3 A map of part of the sky with celestial coordinates superimposed.

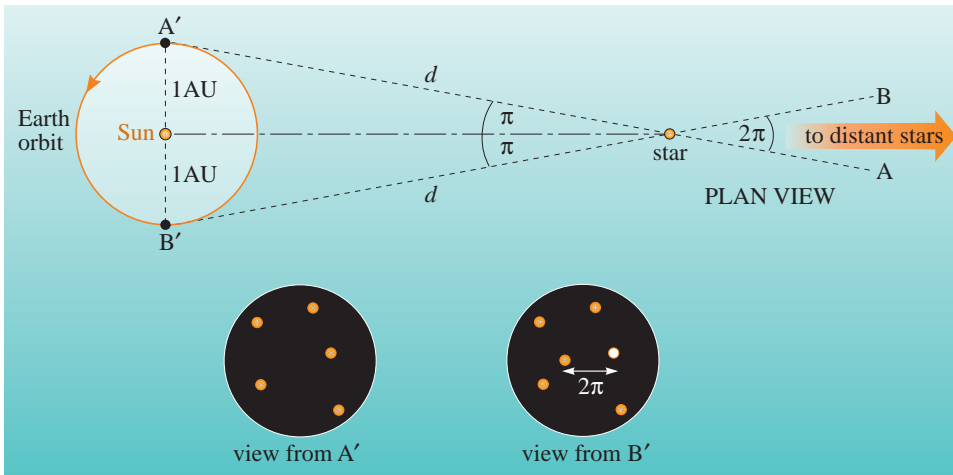


Figure 2.4 An illustration of parallax. The position of a hypothetical nearby star, situated a distance d away, exhibits a tiny angular shift when viewed from the Earth at two intervals. The nearby star appears at location A in January and location B in July. The parallax angle π is defined as half the angular shift between two such positions, six months apart.

Distances to stars may be measured using the method of **trigonometric parallax**, illustrated in Figure 2.4. The angle labelled π is the **parallax angle** of the star. (*Note: In this case π is not being used to represent the numerical constant 3.141 59...)* Using simple trigonometry,

$$\sin \pi = 1 \text{ AU}/d \quad (2.1)$$

where 1 AU (astronomical unit = 1.5×10^{11} m) is the average distance from the Earth to the Sun. By definition, a star with a parallax angle of 1 arcsec ($1^\circ/3600$) lies at a distance from the Earth of 1 parsec = $(1.5 \times 10^{11} \text{ m})/\sin(1^\circ/3600) = 3.1 \times 10^{16}$ m. Since stars are so far away, the parallax angles of all stars are extremely small, so we can always use the small-angle approximation. Therefore the distance d is given by

$$d/\text{pc} = 1/(\pi/\text{arcsec}) \quad (2.2)$$

Parallax angles as small as 10^{-3} arcsec can now be measured using satellite techniques, so stars as far away as 1000 pc (1 kpc) can have their distances measured accurately by this method.

- Barnard's star has a parallax angle of 0.55 arcsec. What is its distance from the Earth in parsecs and in metres?
- $d/\text{pc} = 1/0.55$, so $d = 1.8$ parsec = $(1.8 \times 3.1 \times 10^{16} \text{ m}) = 5.6 \times 10^{16}$ m.

The parsec and its multiples the kiloparsec, megaparsec and gigaparsec (1 Gpc = 10^3 Mpc = 10^6 kpc = 10^9 pc) are the most widely used units of distance in astrophysics. However, you will occasionally see distances expressed in that favourite unit of the science fiction writer, the light-year, where one light-year is the distance travelled by light in one year: $1 \text{ ly} = 9.5 \times 10^{15}$ m. Consequently, one parsec is equivalent to about 3.3 light-years.

2.3.2 Measuring the velocities of stars

Stars are not static in space, rather they move around the Galaxy with speeds that are typically of the order of hundreds of kilometres per second. From an observational perspective, it is convenient to split the overall motion into transverse and radial components, as illustrated in Figure 2.5. It is only the transverse component of the velocity that can actually be *seen*, but since stars are extremely distant from us, a transverse speed of even (say) a hundred kilometres per second would still produce an extremely small angular shift on the sky over the course of one year.

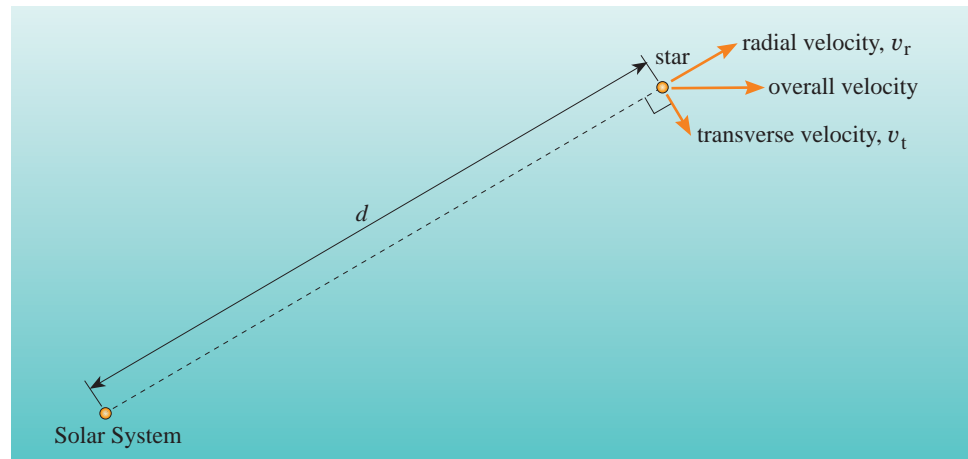


Figure 2.5 A star's motion through space can be split into transverse and radial components. The actual velocity of the star is the vector sum of these two components, and has a magnitude given by $v = \sqrt{v_t^2 + v_r^2}$.

Exercise 2.1 The star Tau Ceti has a transverse velocity component of 33 km s^{-1} and is situated $1.12 \times 10^{14} \text{ km}$ away. What is the angular shift of the star over the course of one year?

The angular shift with time is called the **proper motion** of the star and is usually represented by the Greek letter μ , μ . The largest proper motion measured is that of the nearby Barnard's star for which $\mu = 10.31 \text{ arcsec per year}$. For most stars though, the proper motion is so small as to be undetectable. The magnitude of the transverse velocity is related to the proper motion and distance d of a star by

$$v_t = d \tan \mu \quad (2.3)$$

The radial velocity component of a star's motion can be measured by considering its spectrum. Light propagates like a wave (see Section 5.10) and like any wave is subject to the Doppler effect. This effect is familiar to most people in the context of the change in pitch of a vehicle's siren as it races towards and then away from you. Just as the pitch of sound is shifted to a higher frequency as a vehicle approaches, and a lower frequency as it recedes, so the light emitted by an astronomical object is shifted to higher or lower frequencies too. Since the speed of light is constant, a shift to a higher frequency corresponds to a shift to shorter wavelength (and lower frequency corresponds to longer wavelength). The

magnitude of the shift depends on the relative speed of motion between source and observer – the higher the speed, the larger the shift. In particular, the magnitude of the radial velocity of a star is given by the following **Doppler shift** formula:

$$v_r = c \times \frac{\Delta\lambda}{\lambda} \quad (2.4)$$

where λ (the Greek letter *lambda*) is the wavelength of a spectral line emitted by a star (or observed at rest here on Earth), $\Delta\lambda$ is the shift in wavelength of that line as observed in the star's spectrum and c is the speed of light. Motion away from the Earth gives rise to a shift to longer wavelengths (a redshift), whereas motion towards the Earth gives rise to a shift to shorter wavelengths (a blueshift).

Exercise 2.2 A hydrogen absorption line in the spectrum of a star is observed to have a wavelength of 4863.5 Å compared to the 'rest' wavelength of this absorption line which is 4861.3 Å. What is the radial velocity of the star? ■

2.4 Spectra and temperatures

Much of this section assumes you have an understanding of the properties of light. If you wish to revise your understanding of this topic now, please jump ahead and read Sections 5.7 and 5.10 before returning to here to continue.

The visible **spectrum** of a star represents the light emerging from its photosphere – its outer layers – and can be used to determine a great deal about the physical properties of the star itself. The underlying continuum of a star's spectrum is approximately that of a so-called black-body (see Figure 2.6) and is what gives a star its overall colour. Stars with hotter photospheres will have black-body spectra which reach a peak at shorter wavelengths, or more towards the blue end of the visible spectrum.

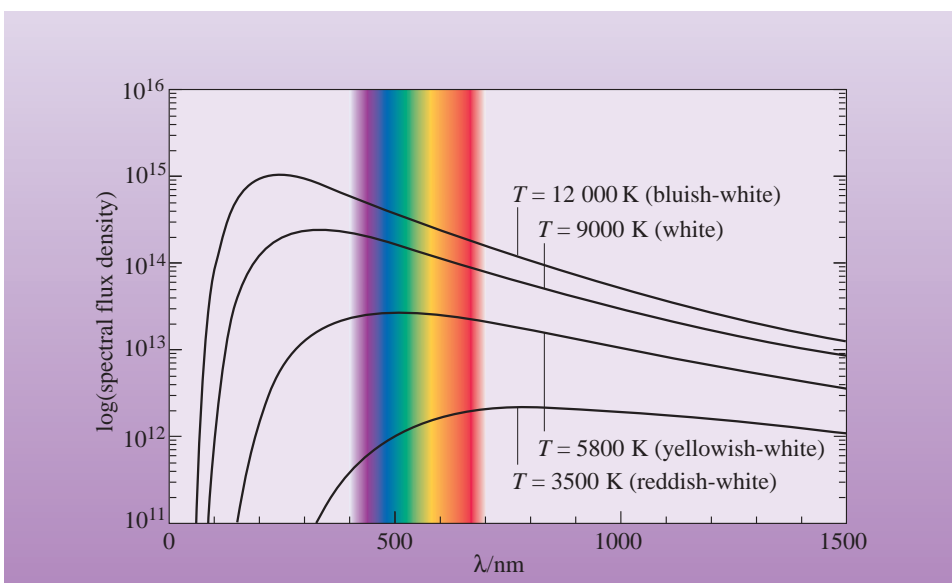


Figure 2.6 Black-body spectra at different temperatures. Notice that the vertical axis is plotted on a logarithmic scale.

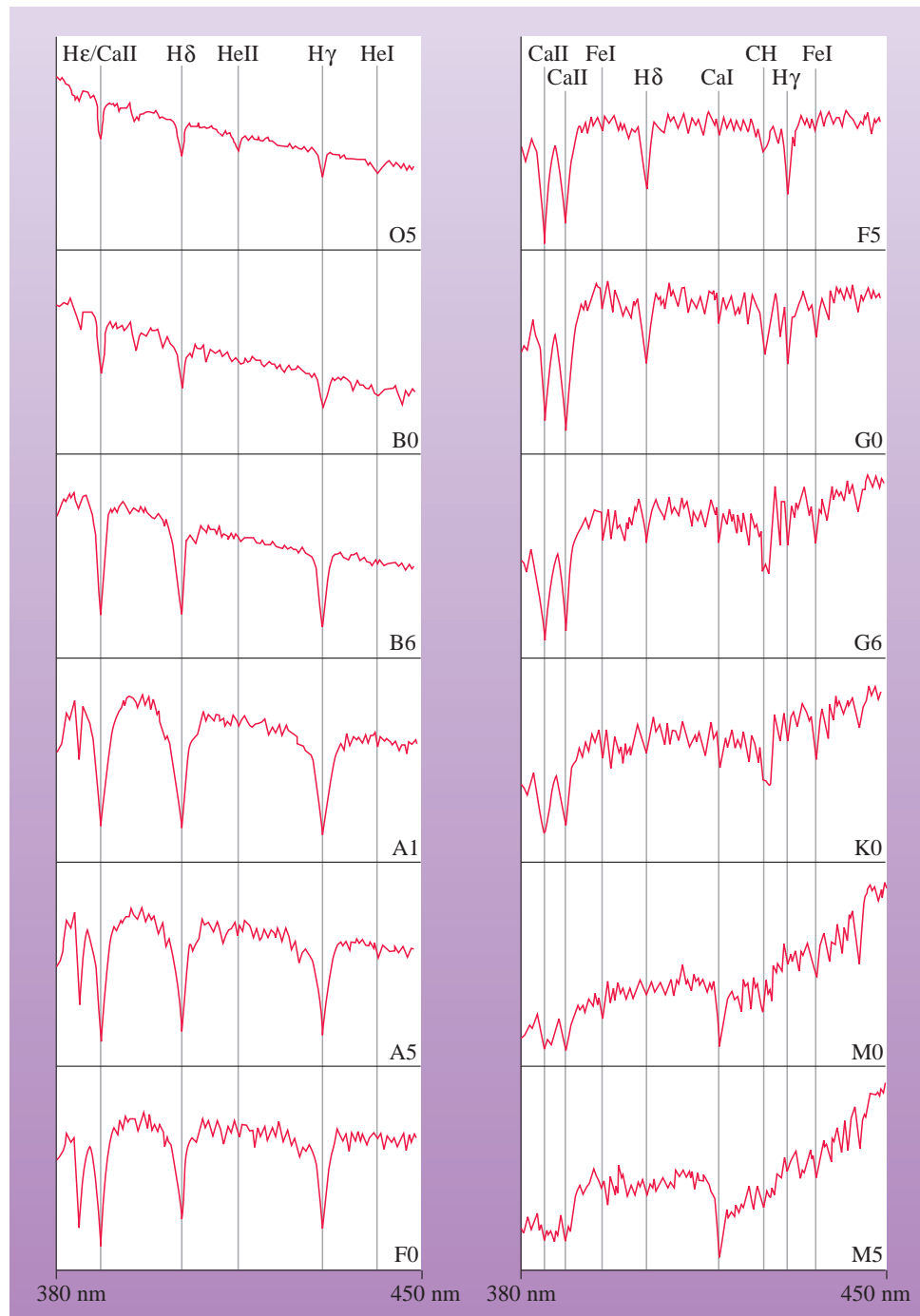


Figure 2.7 Graphs of stellar absorption spectra for different spectral classes. Hydrogen lines in the visible part of the spectrum are labelled $H\gamma$, $H\delta$, etc. (see Section 5.7).

Superimposed on top of the black-body continuum, though, is a series of **absorption lines** arising as a result of the *absorption* of light in the atoms present in the star's outer layers. These absorption lines may be used to accurately classify the star, since their relative strengths give a measure of the temperature of the photosphere (see Figure 2.7). For historical reasons, the seven major divisions denoting the **spectral classification** of stars are labelled as O, B, A, F, G, K, M, in

order of decreasing temperature (see Table 2.1). Each of these spectral classes may be subdivided, giving a more detailed sequence, e.g. ... B5, B6, B7, B8, B9, A0, A1, A2, ..., where the middle of a spectral class is subdivision 5.

Table 2.1 Temperatures of stars around the middle of each spectral class.

Spectra class	Photospheric temperature/K
O	40 000
B	17 000
A	9 000
F	7 000
G	5 500
K	4 500
M	3 000

As indicated by Figure 2.8, broadly speaking the visible spectra of the hottest stars (O-type and B-type) contain lines due to ionized and neutral helium; those of somewhat cooler stars (A-type and F-type) contain the strongest hydrogen lines; whilst those of even cooler stars (G-type and K-type) contain lines due to ionized metals (such as calcium and iron); and the visible spectra of the very coolest stars (M-type) contain molecular lines, in particular those of titanium oxide are prominent. The spectral class of the Sun is G2 which corresponds to a photospheric temperature of about 5800 K. Occasionally **emission lines** are seen in the spectra of stars too, and are generally an indication of energetic processes occurring in the regions that we see.

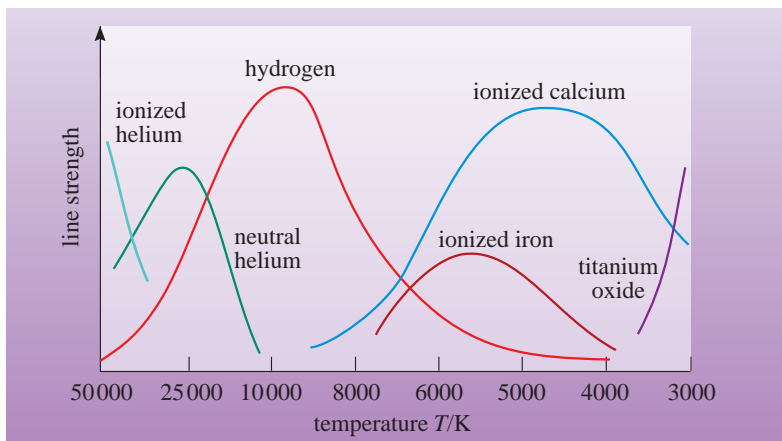


Figure 2.8 The strengths of various absorption lines versus photospheric temperature.

Exercise 2.3 (a) On the basis of Figure 2.8, what is the temperature of a star for which the hydrogen Balmer lines and (unspecified) lines due to neutral helium have equal strength? (b) Using Table 2.1, estimate the spectral class of such a star.



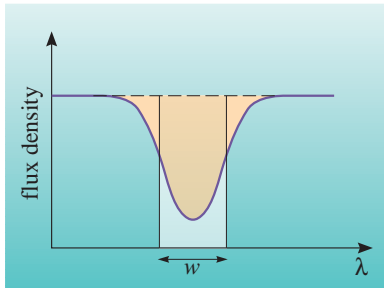


Figure 2.9 The equivalent width w of a spectral line is defined such that the area of the ‘rectangular’ profile is identical to the area of the true line profile.

The strength of a particular absorption or emission line is characterized by its so-called **equivalent width** (see Figure 2.9). The total strength of the line is proportional to its area, defined as the area of the graph enclosed by the line itself and an extrapolation of the continuum across the line. This area may be more conveniently represented by the width of a rectangle w which has the *same* area but a depth equal to that of the adjacent spectral continuum. Equivalent width therefore has the same units as those of wavelength, namely nanometres or ångströms if dealing with the visible part of the spectrum.

The actual width of a spectral line depends on a number of things, but an important consideration is the random motions of the atoms from which the line arises. If these atoms have high speed random motions – some of them moving towards us, some moving away from us, and others moving in all other directions – then the spectral line arising from each atom will be Doppler shifted either towards longer or shorter wavelengths (Figure 2.10) and the resulting wavelength will obey Equation 2.4. The spectral line we observe is the *sum* of these individual Doppler shifted lines, and the result is that a Doppler *broadened* spectral line is seen. **Doppler broadening** of the line can therefore give an indication of the range of speeds present in the atoms from which the line arises. For this reason, the actual width of a spectral line is often quoted in terms of a speed.

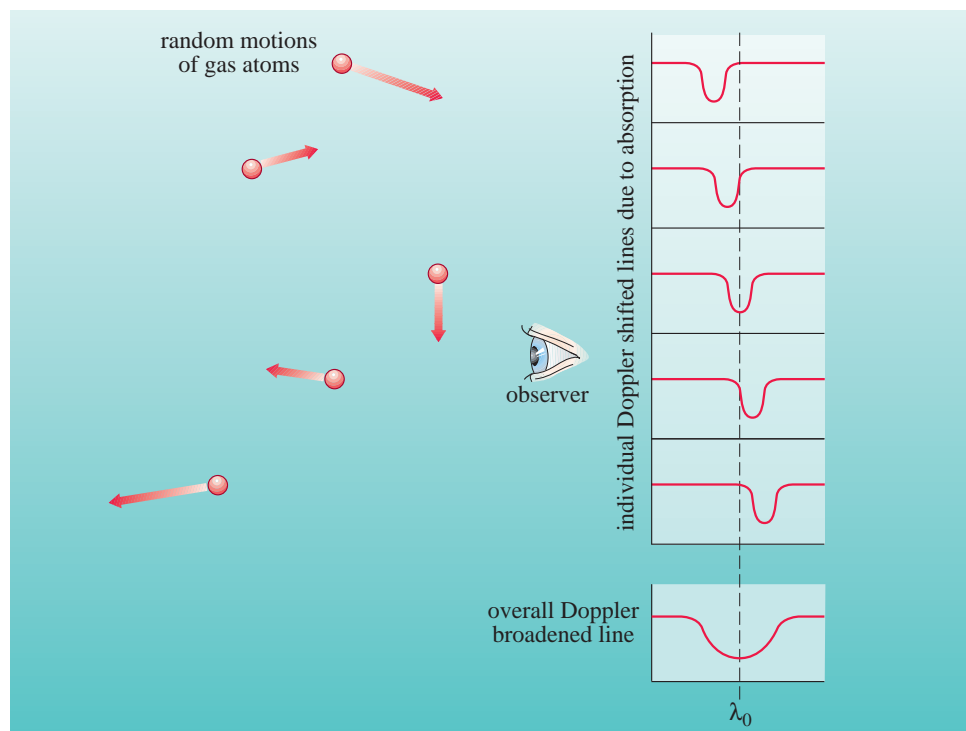


Figure 2.10 Random motions of the atoms from which a spectral line arises give rise to a range of Doppler shifts. The Doppler broadened spectral line is the sum of many individual Doppler shifted lines.

2.5 Luminosities and fluxes

The **luminosity** of a star is the total energy per second that it radiates. If the distance to a star is known, then in principle the *brightness* of the star (i.e. the total energy received per second per unit area, also known as the **flux**) can be used to calculate the star's luminosity. As the light from a star travels out into space, so it becomes spread over the surface of an imaginary sphere of radius d , the distance from the star. Since the surface area of a sphere of radius d is given by $4\pi d^2$, the luminosity and the flux are therefore related according to an inverse square law:

$$F = \frac{L}{4\pi d^2} \quad (2.5)$$

where F is the flux of the star (in, say, W m^{-2} or $\text{erg s}^{-1} \text{cm}^{-2}$) and L its luminosity (in W or erg s^{-1}). In practice, the situation is complicated by the fact that a certain proportion of a star's light is absorbed by intervening gas and dust, and so the relationship between flux and luminosity is not as simple as the above equation indicates.

Since a star's spectrum may be approximated by a black-body continuum, the flux escaping through the star's surface may be approximated by the Stefan–Boltzmann law

$$F \approx \sigma T^4 \quad (2.6)$$

Now, the surface of a star is a sphere, so using a similar geometrical argument to that in Equation 2.5, we may therefore estimate the radius R of a star using the relationship

$$L \approx 4\pi R^2 \sigma T^4 \quad (2.7)$$

where L and T are its luminosity and photospheric temperature respectively and σ is the Stefan–Boltzmann constant.

In fact, Equation 2.7 is used to define the **effective temperature** of a star as the temperature of a black-body source which has the same radius and luminosity as the star:

$$T_{\text{eff}} = \sqrt[4]{\frac{L}{4\pi\sigma R^2}} \quad (2.8)$$

Exercise 2.4 Use Equation 2.8 to determine the effective photospheric temperature of the Sun, i.e. the temperature of the photosphere assuming it radiates as a perfect black-body. (Assume $L_{\odot} = 3.83 \times 10^{26} \text{ W}$, $R_{\odot} = 6.96 \times 10^8 \text{ m}$, $\sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$.)

2.6 Astronomical magnitudes

In practice, fluxes and luminosities are not always used, and the related quantities of astronomical **apparent magnitude** and **absolute magnitude** are encountered instead. The relationship between apparent visual magnitude (represented by V or m_V) and flux in the visual band is shown in Figure 2.11. The brightest stars have

an apparent visual magnitude around -1 whilst the faintest stars visible to the naked eye have $m_V \sim 6$. The magnitude scale is itself logarithmic such that a difference of 5 magnitudes represents a ratio of $100\times$ in flux. So, the apparent magnitudes m_1 and m_2 of two stars with fluxes F_1 and F_2 are related by

$$m_1 - m_2 = 2.5 \log_{10}(F_2/F_1)$$

or $m_1 - m_2 = -2.5 \log_{10}(F_1/F_2)$ (2.9)

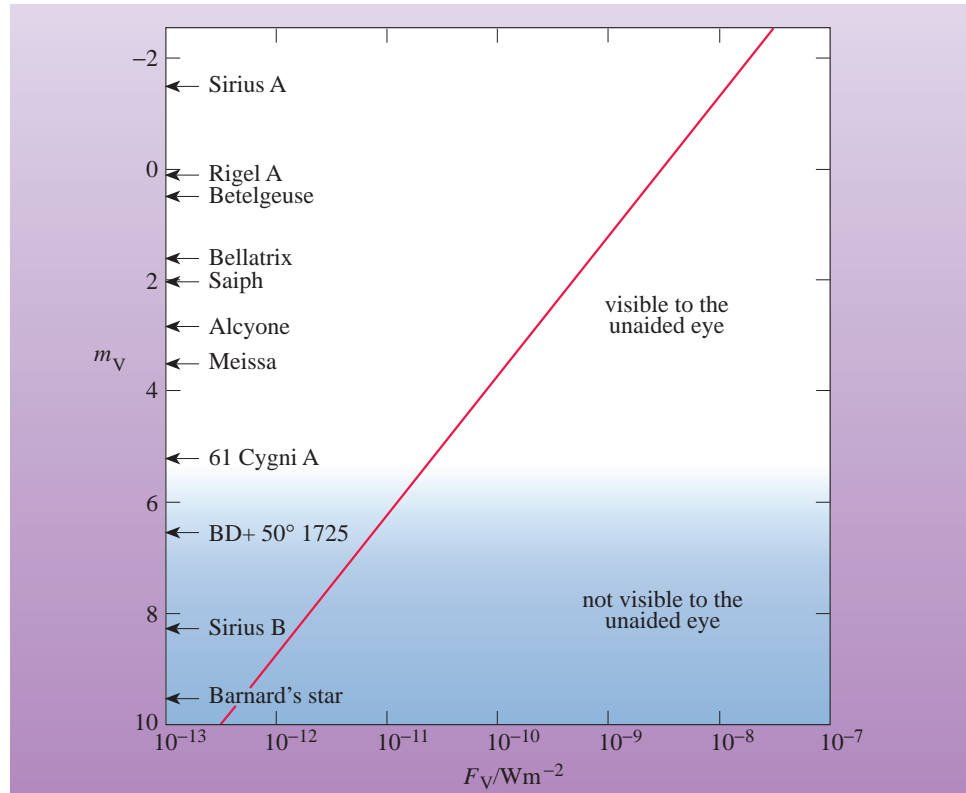


Figure 2.11 The relationship between flux in the V band and apparent visual magnitude. Approximate values of m_V are indicated for a number of stars.

- The bright star Rigel A has an apparent visual magnitude $m_V = 0.12$, whilst the faint star Ross 154 has an apparent visual magnitude $m_V = 10.45$. What is the ratio of the visual flux of Rigel A to that of Ross 154?
- Dividing each side of Equation 2.9b by -2.5 and then calculating 10 to the power of each side (recall Equation 1.23), $10^{(m_1-m_2)/-2.5} = F_1/F_2$, so, $F_1/F_2 = 10^{(0.12-10.45)/-2.5} = 10^{4.132} = 1.355 \times 10^4$. Therefore, viewed from Earth, Rigel A is about 13 600 times brighter than Ross 154.

The apparent magnitude (or flux) of a star is *not* an intrinsic property of the star itself – it also depends on the distance to the star and the amount of intervening absorbing material. By contrast, the absolute magnitude (or luminosity) of a star *is* an intrinsic property of the star. The absolute magnitude (represented by M) is defined as the value of the apparent magnitude that would be obtained at the standard distance of 10 pc from a star, in the absence of any intervening absorbing

matter. Consequently,

$$M = m + 5 - 5 \log_{10} d - A \quad (2.10)$$

where d is the distance to the star in pc and A is the amount of interstellar **extinction** expressed as an equivalent number of magnitudes. The **distance modulus** of a star (or other astronomical object) is defined as the difference between its apparent and absolute magnitudes, hence from Equation 2.10

$$\text{distance modulus} = m - M = 5 \log_{10} d - 5 + A \quad (2.11)$$

- The bright star Rigel A has an apparent visual magnitude $m_V = 0.12$ and is at a distance of $d = 280$ pc. Assuming there is negligible interstellar extinction in our line of sight to Rigel A, what is its absolute visual magnitude? What is its distance modulus?
- Using Equation 2.10, $M = 0.12 + 5 - (5 \log_{10} 280) + 0 = -7.12$. Substituting this value into Equation 2.11, $m - M = 0.12 - (-7.12) = 7.24$.

Absolute magnitudes are related to the luminosities of stars in a similar way to that shown in Equation 2.9 for apparent magnitudes and fluxes, namely

$$\begin{aligned} M_1 - M_2 &= 2.5 \log_{10}(L_2/L_1) \\ \text{or} \quad M_1 - M_2 &= -2.5 \log_{10}(L_1/L_2) \end{aligned} \quad (2.12)$$

Exercise 2.5 The apparent visual magnitudes of Rigel A and Ross 154 are 0.12 and 10.45 respectively, whilst their distances are 280 pc and 2.9 pc respectively. What is the ratio of the luminosities of Rigel A and Ross 154? (Assume that there is negligible interstellar extinction along the line of sight to each star.)

Apparent and absolute magnitudes can be quoted in any one of several regions of the spectrum. Commonly these are expressed as U, B, V, R and I, standing for, respectively, near ultraviolet, blue, visible, red, and near infrared, and are referred to as the *Johnson photometric* system. In addition, three further bands in the near infrared are referred to as J, H and K. The wavelength ranges corresponding to these regions are shown in Table 2.2. Subscripts on m or M indicate the waveband in question, alternatively, the apparent magnitudes are themselves represented by the symbols U, B, V, R, I, J, H and K .

Table 2.2 The standard photometric wavebands.

Waveband	Central wavelength	Width of band
U	365 nm	70 nm
B	440 nm	100 nm
V	550 nm	90 nm
R	700 nm	220 nm
I	900 nm	240 nm
J	1.25 μm	0.24 μm
H	1.65 μm	0.4 μm
K	2.2 μm	0.6 μm

The **bolometric magnitude** of a star is a measure of the total amount of radiation received from it, i.e. across the whole spectrum. The **bolometric correction** BC

is the difference between the bolometric and V-band magnitudes:

$$BC = m_V - m_{bol} \tag{2.13}$$

and is generally defined to be zero for stars of a similar temperature to the Sun.

2.7 Colours and extinction

The difference between two magnitudes (apparent or absolute) of a single object in different wavebands is referred to as an astronomical **colour**. Equations 2.9 and 2.12 show that a *difference* between two magnitudes is related to a *ratio* of two fluxes or luminosities. So, for instance, the quantity $(B - R)$ for a star represents the ratio of two fluxes in different parts of its spectrum (see Figure 2.12). Since stars with different temperatures will have different underlying spectral continua, they will also have different colours. The colour of a star is also therefore a measure of its photospheric temperature, although it is a somewhat cruder measurement than looking at the whole spectrum.

The amount of extinction between the observer and a star will affect its colour as well as its brightness. The extinction in a particular waveband is represented by the symbol A with a particular subscript, such as A_V , A_J , etc. A graph of a typical **extinction law** is shown in Figure 2.13, showing that the amount of extinction is less for the longer wavelength wavebands. For this reason, the effect of interstellar extinction on the spectrum of a star is to absorb the short wavelength part more strongly than the long wavelength part. As a result, the light from a star is said to be **reddened** – it appears relatively more bright in the red part of the spectrum than would be expected from its intrinsic spectrum. In general, the further away a star is, the more extinction it suffers from and therefore the more its spectrum is reddened.

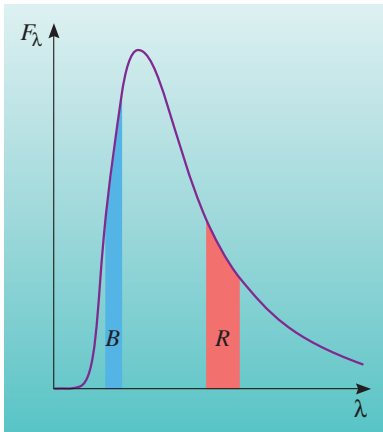


Figure 2.12 Measurement of the magnitudes of a star in two different parts of its spectrum (such as the B and R bands) gives a *colour* (such as $B - R$), which depends on the underlying spectral shape. The colour is therefore related to the photospheric temperature of the star.

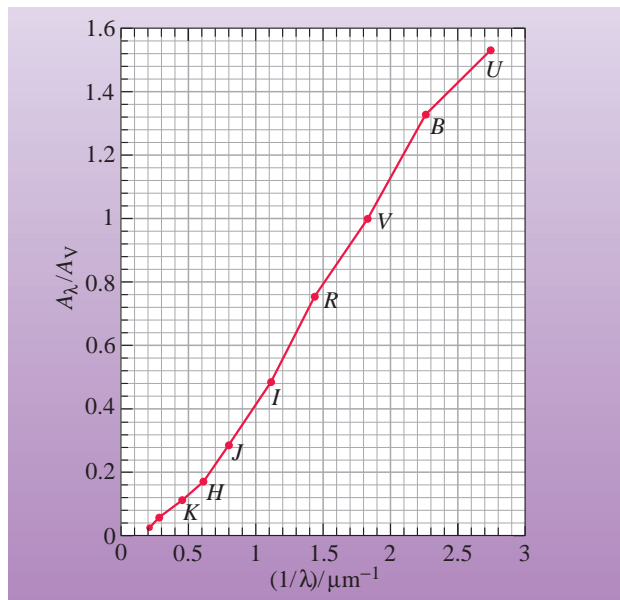


Figure 2.13 A graph of an extinction law, showing the extinction in a particular waveband, relative to that in the V band (i.e. A_λ/A_V), as a function of wavelength. Notice that it is conventional to plot extinction laws with $1/\lambda$ on the horizontal axis, so longer wavelengths are to the left.

- How is the apparent colour of a star ($m_B - m_V$) related to its absolute colour ($M_B - M_V$)?

- Using Equation 2.10, $M_B - M_V = (m_B + 5 - 5 \log_{10} d - A_B) - (m_V + 5 - 5 \log_{10} d - A_V)$. So $M_B - M_V = (m_B - m_V) - (A_B - A_V)$.

The difference between two extinction values in different bands is called a **colour excess** and is represented by the symbol E . So for instance

$$E(B - V) = A_B - A_V = (m_B - m_V) - (M_B - M_V) \quad (2.14)$$

The colour excess is therefore a measure of the amount of reddening for a particular source. In the absence of interstellar extinction, the apparent colour of a star is the same as its absolute colour.

2.8 The Hertzsprung–Russell diagram

The two quantities that can be measured (or inferred) for a large number of stars are their luminosity (or absolute magnitude) and their temperature (or spectral class or colour). A diagram on which stars are plotted according to these two quantities is known as a **Hertzsprung–Russell (H–R) diagram**, first drawn up independently by Ejnar Hertzsprung in 1911 and Henry Norris Russell in 1913.

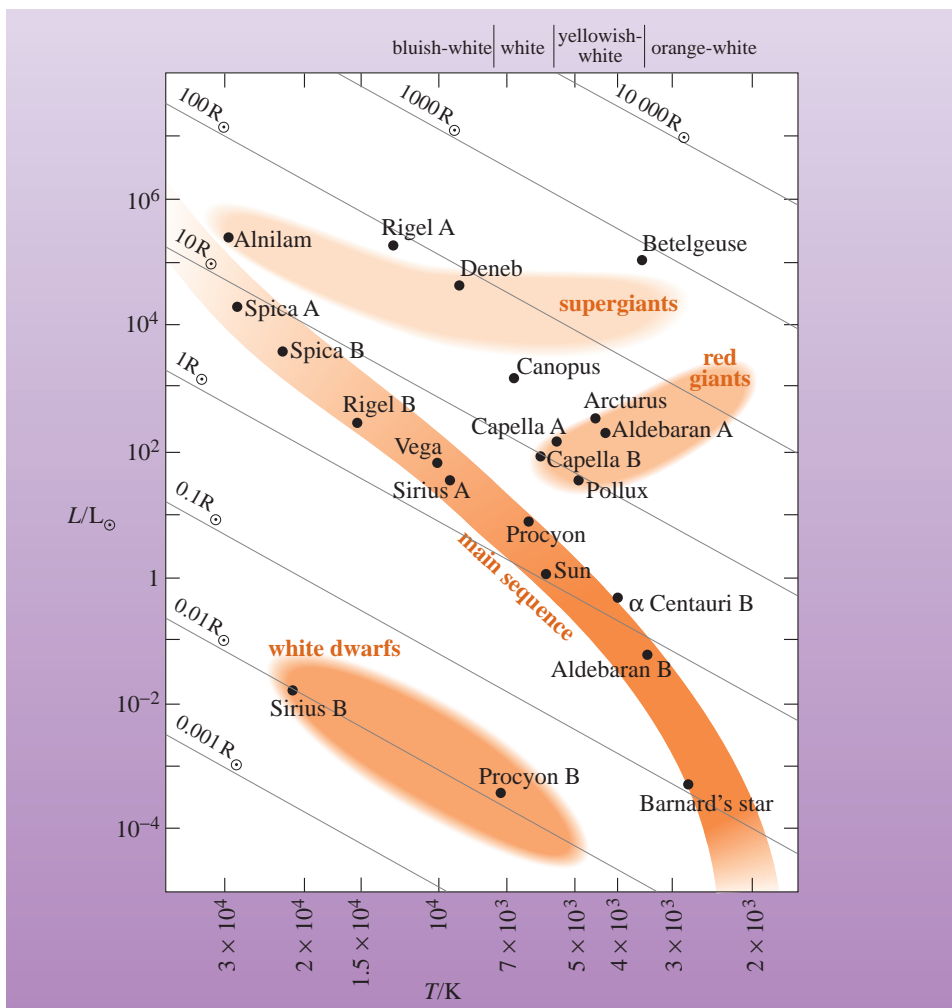


Figure 2.14 An H–R diagram showing where stars tend to concentrate and indicating the positions of a few well known stars.

The H–R diagram is the single most important diagram in stellar astrophysics – and an example is shown in Figure 2.14. On this H–R diagram, notice that both axes are logarithmic and that temperature increases to the *left* by convention. The majority of stars (about 90% of those observed) appear to lie along the so called **main sequence** which reflects the fact that this is where all stars spend the majority of their lives.

As well as the spectral classification introduced earlier (which essentially depends on a star’s photospheric temperature), stars are also assigned a particular **luminosity classification**. Stars on the main sequence are assigned luminosity class V (Roman numeral five), so that the full classification of the Sun is a G2V star. Other locations where many stars are seen are the red-giant branch (luminosity class III) and the supergiant branch (luminosity class I). Finally, white dwarfs are located in the bottom left of the diagram (luminosity class VII). The full list of luminosity classifications is given in Table 2.3

Table 2.3 Luminosity classification

Luminosity class	Description
Ia	bright supergiants
Ib	supergiants
II	bright giants
III	giants
IV	subgiants
V	main sequence (dwarfs)
VI	subdwarfs
VII	white dwarfs

Using Equation 2.7, lines of constant stellar radius have been plotted on Figure 2.14. This shows that giant (and supergiant) stars congregate in the upper right of the H–R diagram, whilst dwarf stars congregate in the lower left. It also shows that the ‘upper’ main sequence (hot, luminous stars) consists of stars which are larger than those on the ‘lower’ main sequence (cool, faint stars).

You will often see H–R diagrams plotted with absolute magnitude on the vertical axis instead of luminosity, and with spectral class on the horizontal axis instead of photospheric temperature, or indeed any combination of these axes. All such combinations are equivalent.

- An H–R diagram can be constructed for the stars in a given star cluster, by plotting their *apparent* visual magnitude up the vertical axis. This can then be used to compare their properties, despite the fact that apparent visual magnitude is not an intrinsic property of the stars themselves. Why is this?
- All the stars in a given star cluster are at about the same distance from the Earth and will suffer from similar amounts of interstellar extinction. Therefore their apparent magnitudes will all be related to their absolute magnitudes in the same way (i.e. $m - M = \text{constant}$ for all stars in the cluster). Their absolute magnitudes in turn are directly related to their luminosities.

Since an astronomical colour is also a measure of the photospheric temperature of a star, you will sometimes see H–R diagrams with apparent colour, such as $(m_B - m_V)$ or $(B - V)$, or absolute colour, such as $(M_B - M_V)$, plotted along

the horizontal axis. As noted above, in the absence of significant interstellar extinction (such as with relatively near by stars), the apparent and absolute colours are the same.

2.9 Masses of stars

The only systems in which the masses of stars can be measured directly are **binary stars** – systems in which two stars orbit around their common centre of mass, as shown in Figure 2.15. In fact, the majority of stars in the Galaxy are in binary (or higher multiple) systems, so this is feasible in many cases.

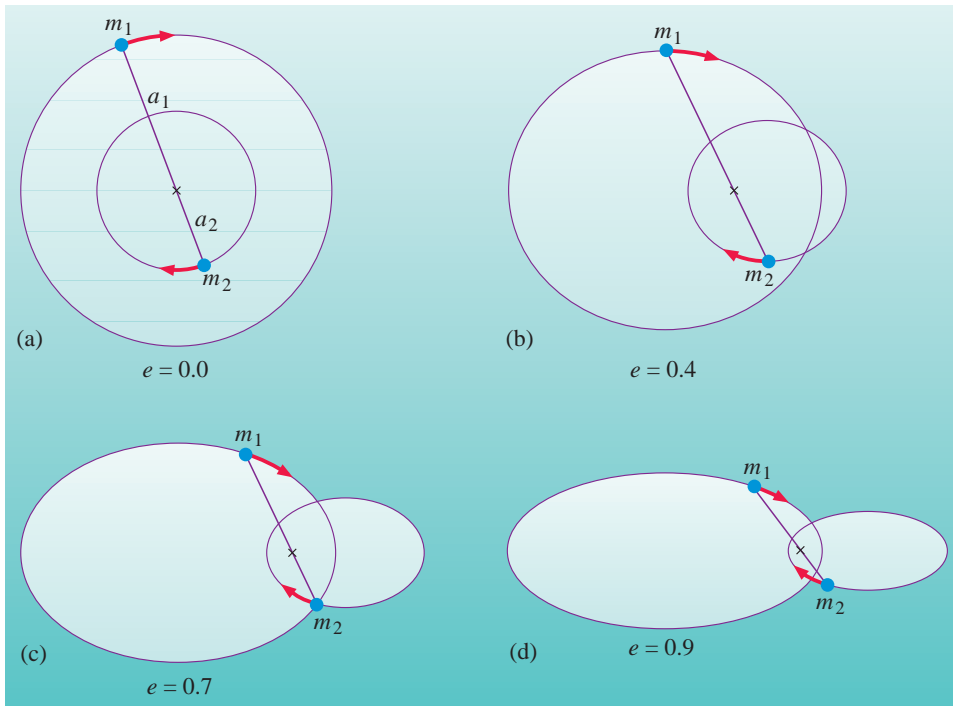


Figure 2.15 Two stars will orbit around their common centre of mass, marked with a cross, following elliptical paths. A special case of an ellipse with zero eccentricity is a circle, as shown in (a), the orbits shown in (b), (c) and (d) have successively larger eccentricities. The two stars lie along a straight line through their common centre of mass at all times during their orbits.

The measurement of masses relies on Kepler's third law, which is itself a consequence of Newton's law of gravitation. For two stars of mass m_1 and m_2 in circular orbits around their common centre of mass, we may write

$$\frac{(a_1 + a_2)^3}{P^2} = \frac{G(m_1 + m_2)}{4\pi^2} \quad (2.15)$$

where P is the orbital period of the binary, G is the gravitational constant and a_1 and a_2 are the distances of the two stars from their centre of mass. From the definition of centre of mass, the masses and distances are further related by

$$q = \frac{m_1}{m_2} = \frac{a_2}{a_1} \quad (2.16)$$

where q is referred to as the **mass ratio** of the system.

In the case of **visual binaries**, where the two stars are seen as separate points of light on the sky, direct measurements of P , a_1 and a_2 are possible and Equations 2.15 and 2.16 can be solved to find m_1 and m_2 .

Essential skill:
Solving for quantities in a binary star

Worked Example 2.1

In the visual binary system Alpha Centauri, the two stars are observed to have angular separations from their centre of mass of 8.0 arcsec and 9.7 arcsec. The system is located 1.31 pc away and the two stars travel around their common centre of mass in a circular orbit whose period is 80.1 years. What are the masses of the two stars? ($G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$ and $M_\odot = 1.99 \times 10^{30} \text{ kg}$)

(NB. Alpha Centauri is actually a triple system, but the third component orbits around the other two a great deal further away, and for the purposes of this calculation may be neglected.)

Solution

The physical separations of the two stars from the centre of mass are found using trigonometry. (Don't forget to convert the angles into degrees.)

$$a_1 = 1.31 \text{ pc} \times \tan(8.0 \text{ arcsec}) = 5.08 \times 10^{-5} \text{ pc} = 1.58 \times 10^{12} \text{ m}$$

$$a_2 = 1.31 \text{ pc} \times \tan(9.7 \text{ arcsec}) = 6.16 \times 10^{-5} \text{ pc} = 1.91 \times 10^{12} \text{ m}$$

So from Equation 2.16, the mass ratio is

$$m_1/m_2 = a_2/a_1 = 1.91 \times 10^{12}/1.58 \times 10^{12} = 1.21$$

and from Equation 2.15,

$$m_1 + m_2 = \frac{4\pi^2(a_1 + a_2)^3}{GP^2}$$

Putting in the numbers, remembering to convert the period into seconds,

$$\begin{aligned} m_1 + m_2 &= \frac{4\pi^2 \times (1.58 + 1.91)^3 \times 10^{36} \text{ m}^3}{(6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}) \times (80.1 \times 365 \times 24 \times 3600 \text{ s})^2} \\ &= 3.94 \times 10^{30} \text{ kg} = 1.98 M_\odot \end{aligned}$$

Now, we have two equations containing both m_1 and m_2 which can be solved. From the first, $m_1 = 1.21m_2$, and substituting this into the second, $(1.21m_2 + m_2) = 1.98 M_\odot$. So $m_2 = 0.90 M_\odot$ and therefore $m_1 = 1.08 M_\odot$. Clearly both stars have masses comparable to the Sun – one slightly more massive than the Sun and the other slightly less.

The situation can become a little more complicated than in Example 2.1 for a number of reasons. First, as you have seen, the orbits may not be circular but elliptical, and secondly, the plane of the orbit may not be perpendicular to the plane of the sky (Figure 2.16). In this case, we actually measure the projection of the separation of the stars, not their true separation. The angle between the plane of the orbit of a binary star and the plane of the sky is known as the **angle of inclination** of the orbit and usually denoted by i .

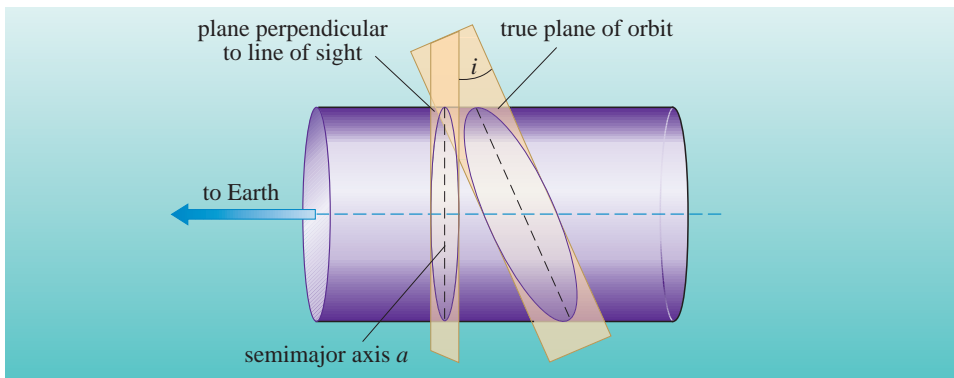


Figure 2.16 The plane of the orbit of a binary may be inclined at an angle i to the plane of the sky. An orbit that is seen ‘edge-on’ has $i = 90^\circ$, whilst an orbit that is seen ‘face-on’ has $i = 0^\circ$.

A more crucial problem is that most binary stars are not visual binaries: they appear through telescopes as a single point of light, because the angular separation between the stars is extremely small. However, they may appear as **spectroscopic binaries** in which the spectral lines from each component can be distinguished. As the two stars orbit each other, their spectral lines shift back and forth periodically as a result of the Doppler effect (see Figure 2.17). By measuring the motion of the Doppler shifted spectral lines, the projected orbital speeds of the two stars can be determined at various positions in the orbit. These values can then be plotted as a so-called **radial velocity curve**, as shown in Figure 2.18. For elliptical orbits, the radial velocity curves have more complex shapes, as shown in Figure 2.19, but the principle behind the measurement is identical.

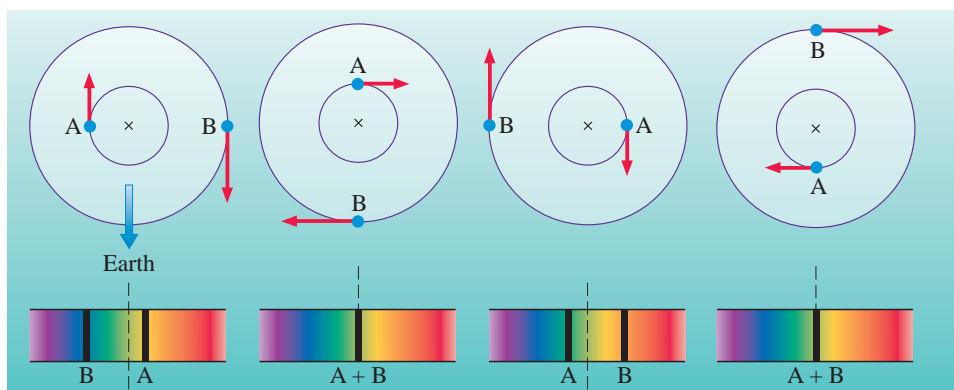


Figure 2.17 In a spectroscopic binary, the spectral lines from each star shift back and forth periodically as the stars move towards and away from the observer.

- Why is it impossible to construct radial velocity curves for stars whose orbits are face-on to the line of sight ($i = 0^\circ$)?
- There will be no component of velocity due to the orbital motion along the line of sight, so no Doppler shift. Such stars would show a constant radial velocity, namely that of the centre of mass of the system.

Figure 2.18 A simple radial velocity curve for the case of circular orbits seen edge-on ($i = 90^\circ$). The speed of the centre of mass of the binary is denoted by V , whilst the maximum speeds of the two stars with respect to their centre of mass are v_1 and v_2 . The individual radial velocity curves for the two stars trace out sinusoidal curves of amplitude v_1 and v_2 , both with period P .

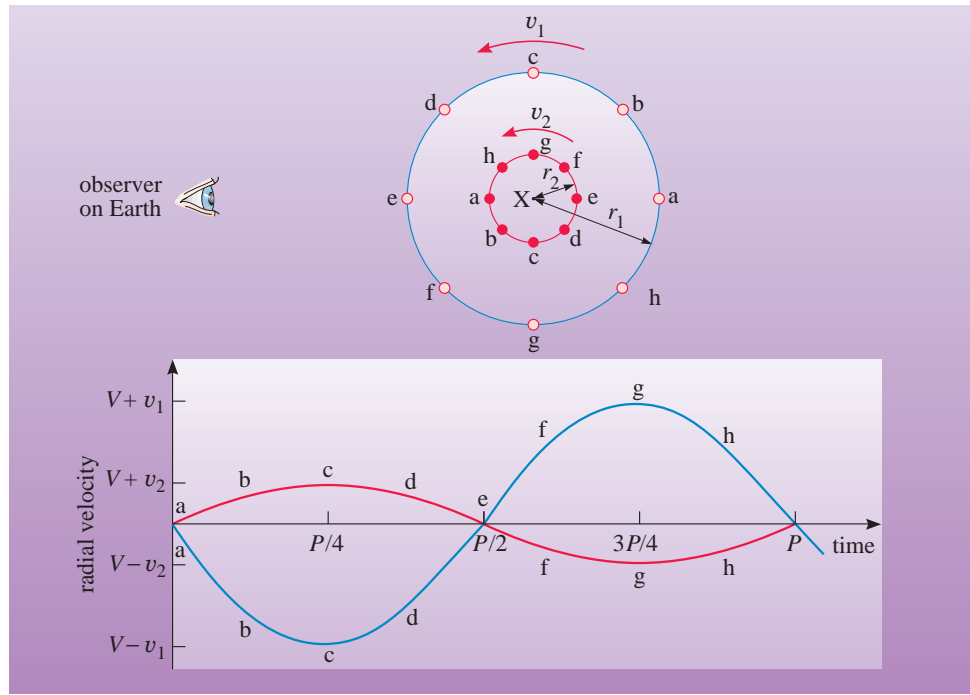
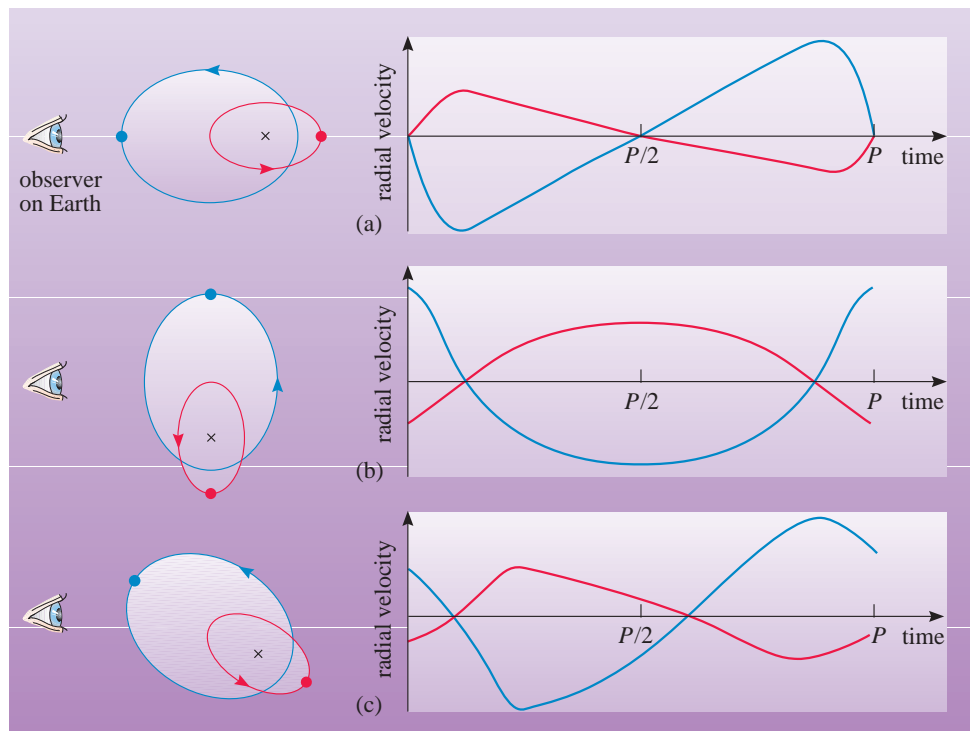


Figure 2.19 Radial velocity curves for elliptical orbits. The plane of the orbits on the sky is the same in each case ($i = 90^\circ$), but the orientation of the orbits varies between (a), (b) and (c).



For circular orbits seen at an angle of inclination i , the measured constant speed around the orbit is equal to the true speed multiplied by the sine of the angle of inclination:

$$v_{\text{measured}} = v_{\text{true}} \times \sin i \quad (2.17)$$

and the true orbital speed is just the circumference of the orbit (of radius a)

divided by the orbital period P :

$$v_{\text{true}} = 2\pi a/P \quad (2.18)$$

By combining these two equations, we may write the measured speeds of the two stars as

$$v_1 = (2\pi a_1 \sin i)/P \quad (2.19)$$

$$\text{and } v_2 = (2\pi a_2 \sin i)/P \quad (2.20)$$

Furthermore, by combining Equations 2.16 and 2.19, the mass ratio of the system is then simply the ratio of the two measured orbital speeds:

$$q = \frac{m_1}{m_2} = \frac{v_2}{v_1} \quad (2.21)$$

Finally, by combining Equation 2.19 with Kepler's law (Equation 2.15) we can define the so-called **mass function** of each star as follows:

$$f(m_1) = \frac{m_2^3 \sin^3 i}{(m_1 + m_2)^2} = \frac{4\pi^2 a_1^3 \sin^3 i}{GP^2} = \frac{Pv_1^3}{2\pi G} \quad (2.22)$$

$$f(m_2) = \frac{m_1^3 \sin^3 i}{(m_1 + m_2)^2} = \frac{4\pi^2 a_2^3 \sin^3 i}{GP^2} = \frac{Pv_2^3}{2\pi G} \quad (2.23)$$

The right-hand side of each of these equations contains measurable quantities (the orbital period and measured speeds of the stars, as well as constants), but the two equations can only be solved for the masses if we have other information about the angle of inclination of the plane of the orbit. Such information may be available if the system is also an **eclipsing binary** (Figure 2.20). In this case, the two stars periodically pass in front of one another, and the angle of inclination of the orbital plane must be close to 90° .

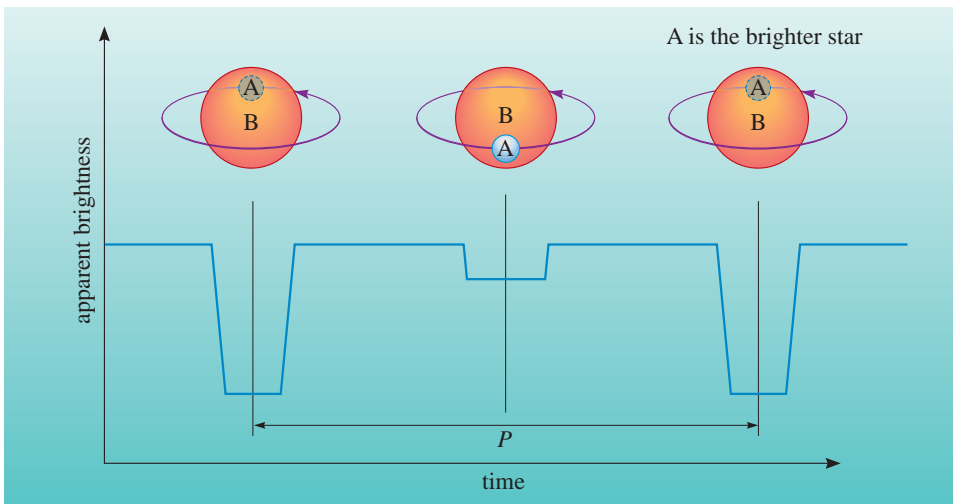


Figure 2.20 In an eclipsing binary, the stars periodically pass in front of one another causing dips in brightness. In the example shown here, the smaller star (A) is the brighter of the two so the deepest eclipse occurs when it passes behind the larger, fainter star (B).

Essential skill:

Solving for quantities in a binary star

Worked Example 2.2

The interacting binary star Centaurus X-3 consists of a neutron star and a giant companion star in circular orbits around their common centre of mass. The neutron star emits pulses of X-rays every 4.84 s and the companion star eclipses the neutron star once every 2.09 days. From the nature of this eclipse, the inclination of the orbit is estimated to be 70.0° . Measurements of the Doppler shift of lines in the spectrum of the companion show its observed orbital speed to be 24.4 km s^{-1} whilst measurements of the Doppler shift of the X-ray pulsations from the neutron star show its observed orbital speed to be 414 km s^{-1} . Calculate the masses of the two stars. ($G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$ and $M_\odot = 1.99 \times 10^{30} \text{ kg}$.)

Solution

Using Equation 2.21, the mass ratio of the system is simply

$$q = m_1/m_2 = v_2/v_1 = 24.4 \text{ km s}^{-1}/414 \text{ km s}^{-1} = 0.0589$$

Using Equation 2.22, the mass function of the neutron star is

$$\begin{aligned} f(m_1) &= \frac{m_2^3 \sin^3 i}{(m_1 + m_2)^2} = \frac{Pv_1^3}{2\pi G} \\ &= \frac{(2.09 \times 24 \times 3600 \text{ s}) \times (414 \times 10^3 \text{ m s}^{-1})^3}{2\pi \times 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}} \\ &= 3.06 \times 10^{31} \text{ kg} = 15.4 M_\odot \end{aligned}$$

Now, the left-hand side of this equation can be rearranged by dividing top and bottom by m_2^2 to give

$$\frac{m_2 \sin^3 i}{(m_1/m_2 + 1)^2} = 15.4 M_\odot$$

and then substituting for the mass ratio from above, this becomes

$$\frac{m_2 \sin^3 i}{(0.0589 + 1)^2} = 15.4 M_\odot$$

so $m_2 = 15.4 M_\odot \times 1.121/(\sin 70.0^\circ)^3 = 20.8 M_\odot$ and therefore $m_1 = 0.0589 \times 20.8 M_\odot = 1.23 M_\odot$. So the mass of the companion star is 20.8 times that of the Sun, whilst the mass of the neutron star is 1.23 times that of the Sun.

From measurements such as those discussed above, stellar masses have been found to lie in the range from about $0.1 M_\odot$ to about $50 M_\odot$. The masses of stars lying in various parts of the H–R diagram are shown in Figure 2.21. The stars on the upper main sequence (hot, luminous, large stars) are clearly more massive than those on the lower main sequence (cool, faint, small stars). Table 2.4 summarizes the masses, radii, luminosities and effective temperatures of stars on the main sequence.

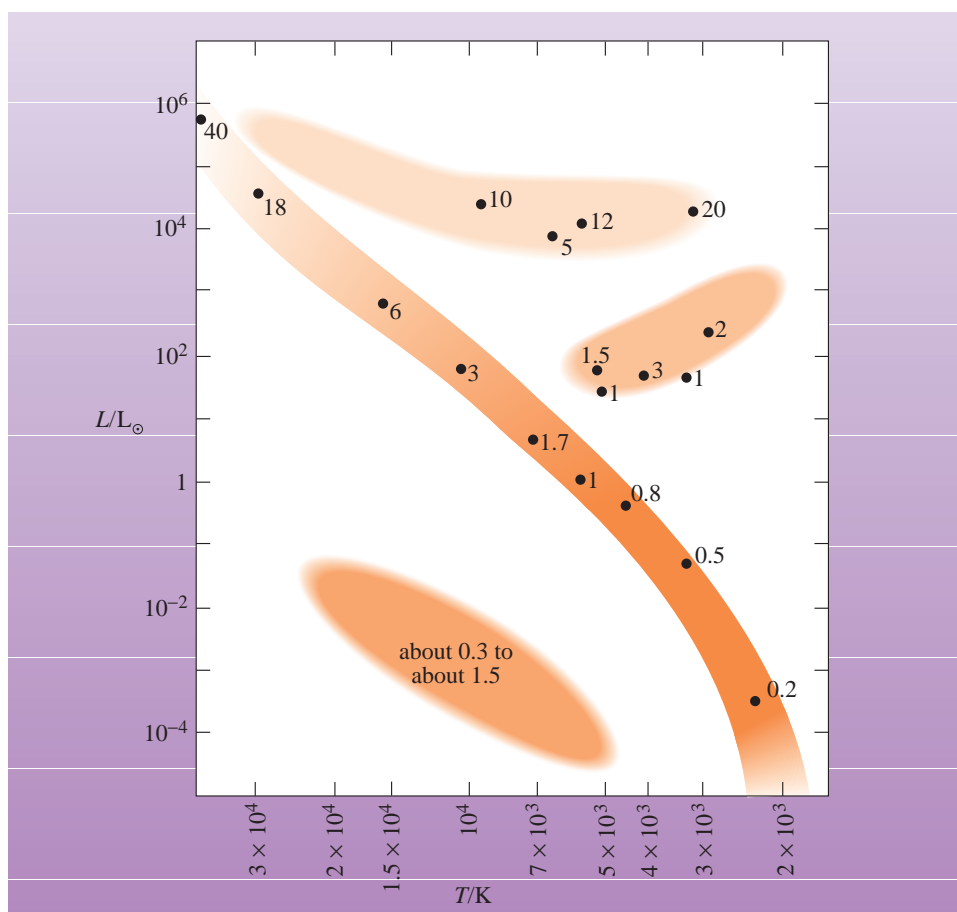


Figure 2.21 The H–R diagram showing stellar masses (M/M_{\odot}) found in various locations.

Table 2.4 Physical characteristics of stars on the main sequence.

Mass/ M_{\odot}	Radius/ R_{\odot}	Luminosity/ L_{\odot}	Effective temperature/K
0.1	0.13	0.001	2500
0.5	0.25	0.03	3800
1.0	1.0	1.0	6000
1.5	1.2	5.0	7000
3.0	2.4	60	11 000
5.0	3.5	450	14 500
15	7.0	17 000	28 000
25	10	80 000	35 000
40	18	500 000	38 000

Exercise 2.6 Use Equation 2.7 to verify that the luminosity of a $5 M_{\odot}$ star with radius $3.5 R_{\odot}$ and effective temperature $14\,500$ K is indeed about $450 L_{\odot}$ as shown in Table 2.4. (Assume $L_{\odot} = 3.83 \times 10^{26}$ W, $R_{\odot} = 6.96 \times 10^8$ m and $\sigma = 5.67 \times 10^{-8}$ W m $^{-2}$ K $^{-4}$.)

2.10 Life cycles of stars

From one day to the next, the brightness, temperature and size of the Sun seem to be reasonably constant. However, stars do indeed evolve and change, but for the majority of stars these changes occur on such long timescales that they cannot be observed directly. The H–R diagram provides us with a snapshot in time of stars in various stages of their evolution. Densely populated regions of the H–R diagram indicate locations where stars spend a large portion of their lives, whilst sparsely populated regions may indicate combinations of temperature and luminosity that stars pass through relatively quickly.

Observational and theoretical evidence points to dense interstellar molecular clouds as being the place where star formation begins. An external trigger mechanism is believed to cause a cloud to start contracting under the influence of gravitational forces. As contraction of a dense cloud continues, it is thought that the cloud fragments into smaller parts, each of which may continue to contract further. This gravitational contraction is accompanied by a rise in temperature throughout the fragment, though this is moderated by the escape of radiation. Eventually the contracting fragment may be considered to be a **protostar**. Some protostars are seen to exhibit bipolar outflows as well as discs of material which may be planetary systems in the process of formation.

When the temperature in the core of the protostar rises sufficiently, nuclear fusion reactions begin. This provides the energy source to prevent further contraction and at this stage the protostar has reached the region of the H–R diagram referred to as the main sequence. In particular, the track along which newly formed stars lie is known as the **zero-age main-sequence** (ZAMS). The time for a protostar to reach this stage is generally less than about 10^8 years – the more massive the fragment, the shorter the timescale (see Table 2.5).

Nuclear fusion is the source of energy that powers main sequence stars, and the main sequence represents the stable configuration of stars of different mass but similar composition, converting hydrogen into helium. The processes all rely on the fact that the mass of a single nucleus of helium-4 is *less than* the mass of the four protons (hydrogen nuclei) from which it was formed. The mass deficit is converted into energy and consequently, for each helium-4 nucleus that is formed, about 25 MeV of energy is released into the core of a star in the form of electromagnetic radiation and kinetic energy. The detailed nuclear reactions that are responsible for converting hydrogen into helium depend on the core temperature, and therefore the mass of a star. For stars less than about $1.5 M_{\odot}$ (lower main sequence stars) the **proton–proton chain** dominates (see Figure 2.22), whilst for more massive stars (upper main sequence) reactions involving carbon, nitrogen and oxygen as catalysts are dominant and comprise the **CN cycle** (see Figure 2.23).

Table 2.5 The time for protostars of various masses to reach the main sequence.

Mass/ M_{\odot}	Time to reach main sequence/yr
0.5	1.5×10^8
1.0	5.0×10^7
1.5	1.8×10^7
3.0	2.5×10^6
5.0	5.8×10^5
9.0	1.5×10^5
15.0	6.2×10^4

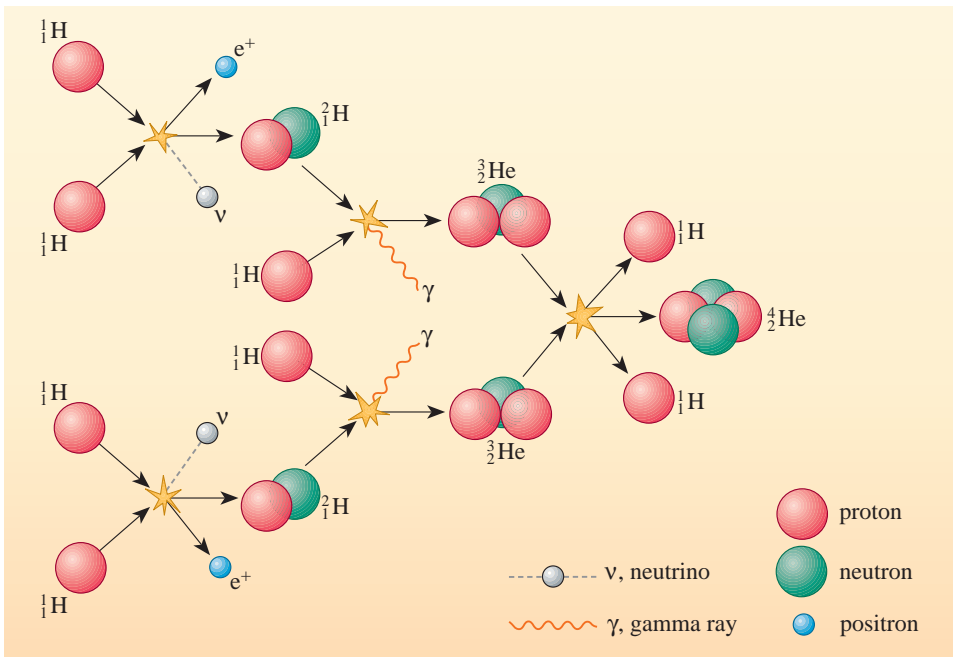


Figure 2.22 The proton–proton chain. Two hydrogen nuclei (${}^1_1\text{H}$ or protons) combine to form a nucleus of deuterium. The deuterium nucleus reacts with another proton to form a nucleus of helium-3. Finally two helium-3 nuclei react to form a nucleus of helium-4 with the ejection of two protons. The net result is that four protons have been converted into a nucleus of helium-4.

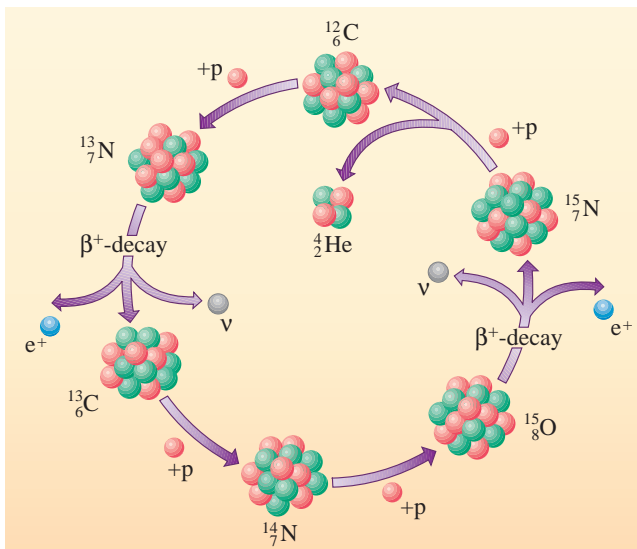


Figure 2.23 The CN cycle. A nucleus of carbon-12 captures a proton to form nitrogen-13. This nucleus undergoes beta-plus decay (β^+ -decay) to form carbon-13. Two more protons are captured in turn to produce nitrogen-14 then oxygen-15, before another β^+ -decay produces a nucleus of nitrogen-15. This nucleus captures one final proton and converts into a nucleus of carbon-12 with the ejection of a helium-4 nucleus. The net result is that four protons have been converted into a nucleus of helium-4, with the carbon-12 acting as a catalyst, since it is returned at the end of the cycle. (Nuclear decay processes are discussed in Section ??.)

The lifetime of a star on the main sequence decreases rapidly for increasing mass, as shown in Table 2.6. Stars of mass much less than about $0.1 M_{\odot}$ never reach sufficiently high core temperatures to trigger hydrogen fusion reactions and so instead they become **brown dwarfs**. At the other extreme, stars of mass greater

Table 2.6 The lifetime of a star on the main sequence.

Mass/ M_{\odot}	Lifetime on main sequence/yr
0.1	1.0×10^{12}
0.5	2.0×10^{11}
1.0	1.0×10^{10}
1.5	3.0×10^9
3.0	5.0×10^8
5.0	1.5×10^8
15.0	1.5×10^7
25.0	6.0×10^6
40.0	2.0×10^6

than about $100 M_{\odot}$ (the upper limit is very uncertain) are not stable because their radiation pressure is so great as to overcome the gravitational forces holding them together. The least massive stars are in fact the most common, and a graph illustrating the relative number of stars that are born with a given mass is shown in Figure 2.24. Note, however, that recent observations indicate that this graph may turn over below about $0.5 M_{\odot}$ (i.e the blue line on Figure 2.24). In other words the commonest stars have a mass of about half that of the Sun, and there are relatively fewer low-mass stars and brown dwarfs.

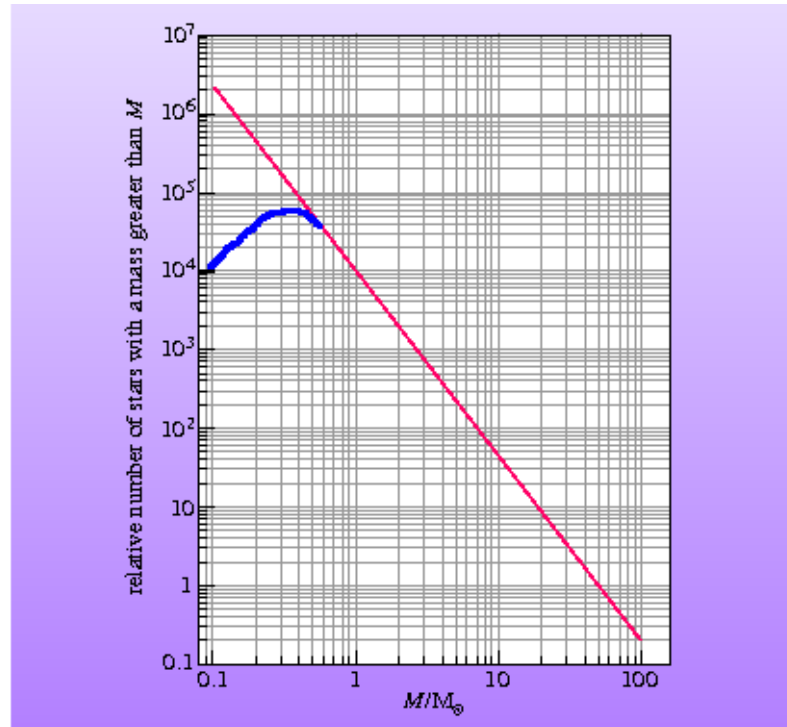


Figure 2.24 The relative number of stars that are born with a given mass. Notice that both axes of this histogram are logarithmic. Roughly speaking for every one star born with a mass around $50 M_{\odot}$, there will be about 50 with a mass of more than $10 M_{\odot}$, 200 with a mass of $5 M_{\odot}$ or more, 2000 with a mass of at least $2 M_{\odot}$, 10 000 with a mass above $1 M_{\odot}$, and 50 000 stars with a mass of at least $0.5 M_{\odot}$. The red line is the conventional picture, the blue line indicates the results of recent observations.

When the hydrogen in the core of a main sequence star is exhausted, the core contracts and its temperature rises. The star begins to move away from the main sequence on the H–R diagram. The transition to the next phase is accompanied by a contraction and heating of the core, but a swelling of the diameter of the star by typically a factor of ten, and a cooling of the surface. The star becomes a **red giant** and moves to rapidly occupy the upper right of the H–R diagram (see Figure 2.25). At a core temperature of around 10^8 K, helium burning is initiated, with hydrogen continuing to burn in a shell surrounding the core. Helium fusion occurs via the **triple-alpha process** in which three helium nuclei fuse to form a nucleus of carbon-12. In stars of mass less than about $2.5 M_{\odot}$, the electrons in the core first become degenerate – a state in which the pressure is nearly independent of temperature. As a result, an unstable ‘run-away’ process develops and produces

an explosive release of energy in the core of the star. This **helium flash**, as it is known, occurs on a timescale of only a few hours. After depletion of helium in the core, helium burning continues in a shell surrounding the core, with hydrogen burning still occurring in another shell further out. During this phase, periodic shell helium flashes occur, each accompanied by a further swelling of the red giant. The giant phase as a whole lasts for approximately 10% of the main sequence lifetime of a particular star.

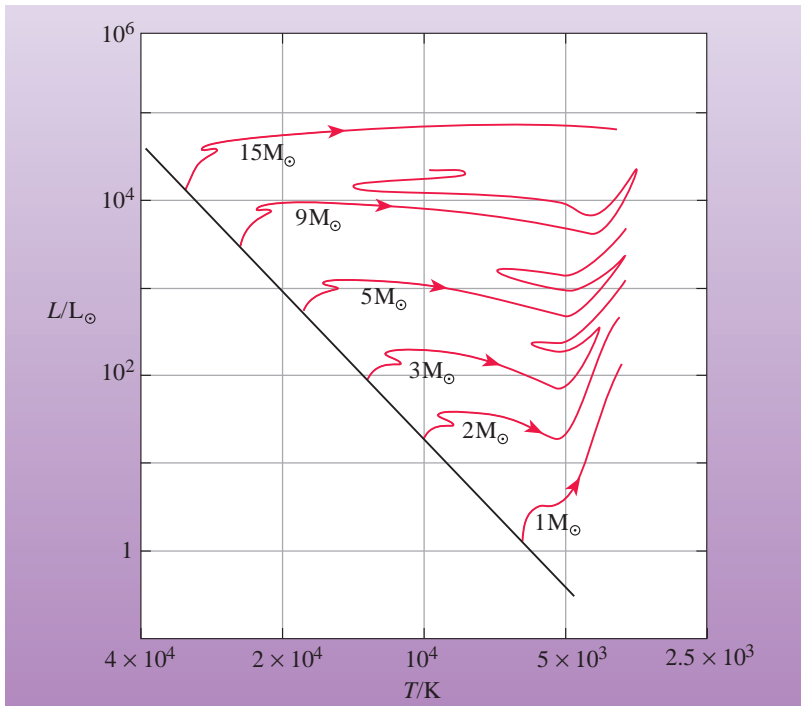
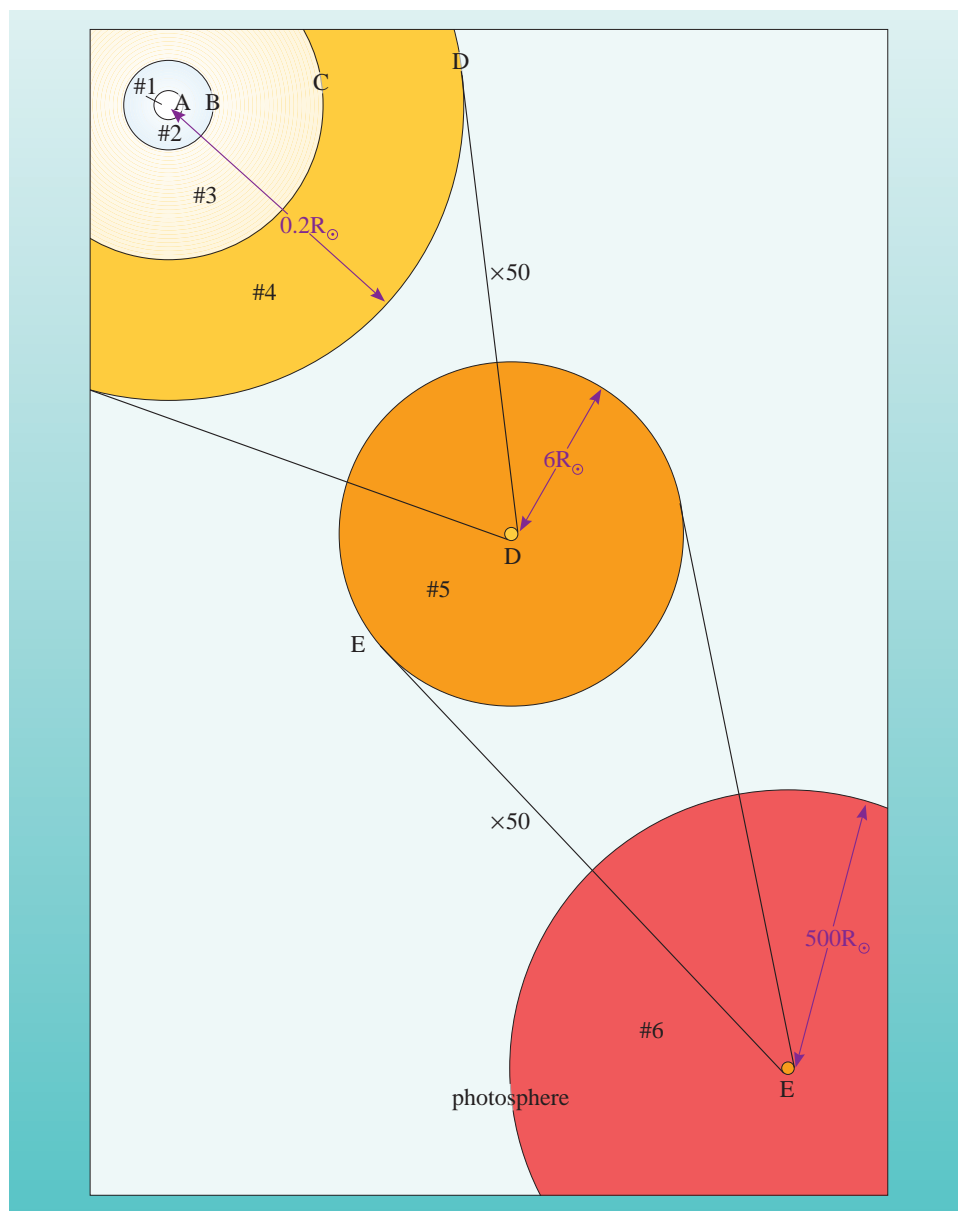


Figure 2.25 Theoretical evolutionary tracks on the H–R diagram for stars of various masses.

During and after the giant phase, most stars pass through a region on the H–R diagram called the **instability strip**. Here a star undergoes pulsations that lead to a regular variation in the star’s luminosity. At around this stage, some stars eject a shell of material, of mass $0.1 M_{\odot}$ to $0.2 M_{\odot}$, called a **planetary nebula**. In relatively low-mass stars, like the Sun, carbon nuclei are the furthest that nuclear fusion can progress. Therefore, after depletion of helium in the core, the core again collapses until electron degeneracy sets in. The pressure resulting from this quantum-mechanical effect is able to halt collapse and the result is a **white dwarf** star. With no further nuclear reactions possible, the white dwarf slowly cools and eventually disappears from view.

More massive stars have a more interesting fate. After the triple-alpha process, further nuclear reactions are initiated in the core, each one building heavier nuclei than the one that preceded it. As each new reaction sequence is initiated, a new ‘flash’ occurs as the star readjusts its internal structure, and the star moves around the H–R diagram. The star will eventually swell into a supergiant and its internal structure will somewhat resemble that of an onion (Figure 2.26) consisting of concentric shells burning ever heavier nuclei the closer to the centre they are found.

Figure 2.26 The ‘onion shell’ structure of an evolved supergiant star with a mass of $15 M_{\odot}$ and a radius of $500 R_{\odot}$. Notice this diagram is drawn approximately to scale with each of the three parts magnified by $\times 50$ compared with that further out. The innermost zone (#1) contains nuclei of silicon and sulfur with silicon burning to produce iron and nickel. Surrounding this, oxygen burns to silicon in shell A and zone #2 contains oxygen, magnesium and silicon. Neon burns to oxygen and magnesium in shell B and zone #3 contains oxygen, neon and magnesium. Carbon burns to neon and magnesium in shell C and zone #4 contains carbon and oxygen. Helium burns to carbon and oxygen in shell D and zone #5 contains helium. Finally hydrogen burns to helium in shell E and zone #6 contains material with essentially the original proportions of hydrogen and helium from which the star formed.



Nuclei of iron, nickel and cobalt represent the furthest limit of nuclear fusion. To progress to yet more massive nuclei, more energy must be put in than is returned by the fusion reaction. Consequently these reactions are not energetically favourable for the star. With no further source of energy, the iron core of a massive star will therefore collapse in a matter of seconds. The density in the core of the star becomes comparable to that of an atomic nucleus, and the collapse is only halted by the quantum-mechanical effect known as neutron degeneracy pressure. As the collapse ceases, a shock wave is launched through the outer layers of the star which causes the outer layers to be explosively expelled. This is called a type II **supernova** explosion. In the explosion, rapid nuclear processes take place producing elements even heavier than iron, and the atoms formed are then thrown out into space. The remnant of the core left behind is a **neutron star**, whilst the expanding cloud of debris may be observable for a few thousand years as a glowing **supernova remnant**.

2.11 Stellar end-points

Two possible end-points of stellar evolution have been mentioned: low-mass stars will end their lives as white dwarfs, whilst more massive stars will end up as neutron stars after undergoing a supernova explosion. A comparison between the properties of these types of object is shown in Table 2.7.

Table 2.7 A comparison of the properties of white dwarfs and neutron stars.

	White dwarf	Neutron star
mass	$< 1.4 M_{\odot}$	$\approx 1.4 - 3.0 M_{\odot}$
radius	$\approx 10^4$ km	≈ 10 km
density	$\approx 10^9$ kg m ⁻³	$\approx 10^{18}$ kg m ⁻³

In both white dwarfs and neutron stars, angular momentum is conserved when the core of a star collapses, and so newly formed white dwarfs and neutron stars are likely to rotate very rapidly. **Pulsars** are rapidly rotating, highly magnetized, neutron stars which produce beamed radio emission. As the star rotates, the beam is swept around the sky. If the orientation is such that the beam sweeps across the Earth, regular pulses of radio emission can be observed repeating at the pulsar rotation period (typically a few milliseconds to the order of seconds).

For stellar cores with a mass in excess of around $3 M_{\odot}$, it is believed that neutron degeneracy pressure is insufficient to prevent even further collapse. The result is a **black hole** – an object with such an intense gravitational field that its escape speed exceeds the speed of light (see Section 5.4 for a discussion of escape speed).

White dwarfs, neutron stars and black holes, collectively referred to as compact objects, are observed in a variety of binary star systems. If the system has evolved such that the two stars in the binary are relatively close together, it is possible for material to transfer from the companion (normal) star onto the compact object with the result that high-energy electromagnetic radiation is emitted from the system. The evolution of each of the stars in such a binary will in general be modified by the presence of the other star and a huge variety of phenomena are observed in these **accretion-powered compact binaries**.

2.12 Planetary structure

When stars explode as supernovae, they seed the interstellar medium with nuclear-processed material (elements heavier than hydrogen and helium) which can be incorporated into newly forming stellar systems and provide the raw material for the formation of planets. In this section we briefly summarise the structure of the planets in our Solar System as an introduction to the likely planets that might be found around other stars.

2.12.1 Terrestrial planets

The **terrestrial planets** of the Solar System include Mercury, Venus, the Earth and Moon, and Mars. In addition, some of the larger satellites of the outer

planets, such as Io, Europa, Ganymede, Callisto and Titan, are considered to be terrestrial-like bodies and have similar composition or structure in some cases.

Detail of the internal structure of Earth is provided by both direct and indirect evidence. Composition of the lithosphere, that is the crust and upper mantle, can be determined by examination of key rock types. The structure and composition of the deep mantle and core is revealed from the properties and response of seismic waves that pass through the planet. Similar studies of other terrestrial planets may be carried out from planetary landers or satellites.

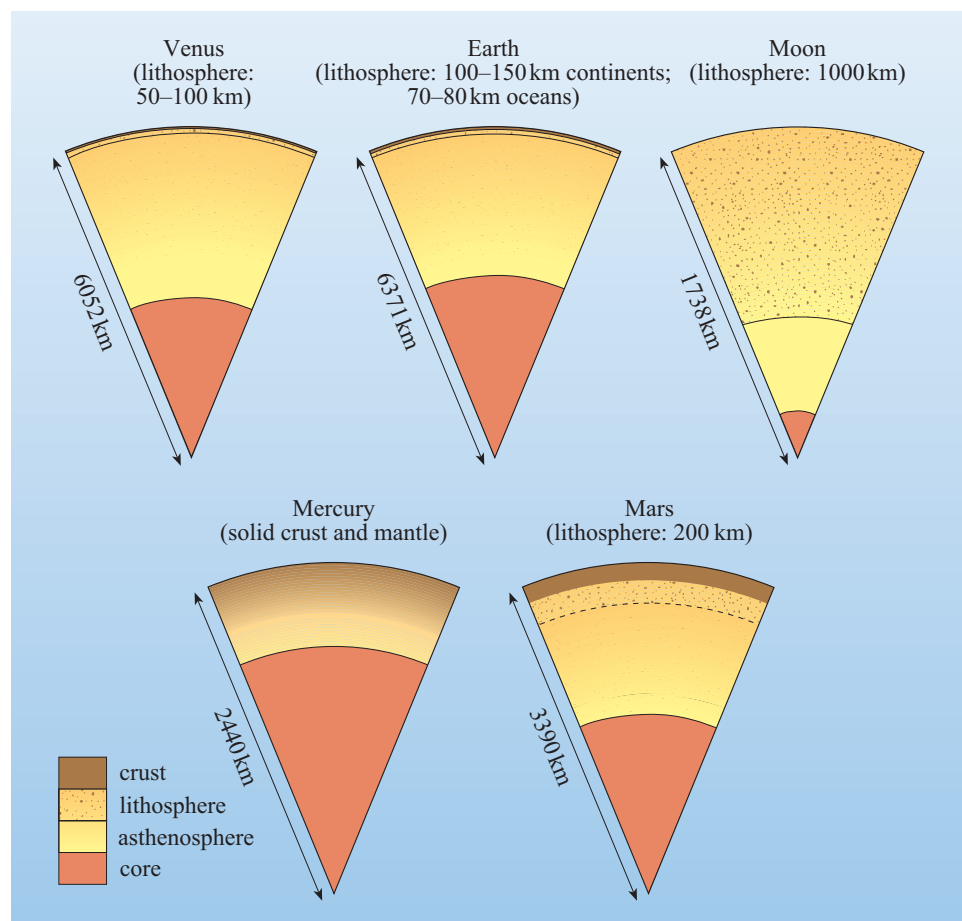


Figure 2.27 The internal structure of terrestrial planets.

Such studies allow the definition of various zones within a planet's structure, as shown in Figure 2.27. The **asthenosphere** of a planet is the zone where pressure and temperature are sufficiently high to allow a material to flow, even in its solid state. Consequently, in this zone the internal heat is mainly transferred surfacewards by convection. The **lithosphere** of a planet is defined as the rigid outermost layer that cannot convect. Instead, the internal heat is carried through it by conduction. On geologically active worlds, such as Earth and Io, internal heat can also be transported by advection through the lithosphere by volcanic processes. In addition, heat is also transferred to Earth's surface by wholesale recycling of the lithosphere. This results in a phenomenon known as plate tectonics, which helps describe the movement of lithospheric plates. The Earth's lithosphere consists of oceanic and continental crust, and the uppermost part of

the mantle. Convective and conductive processes of heat transfer can occur equally well in ice-dominated bodies as well as silicate bodies.

The accretion and final assembly of Earth and other terrestrial planets is thought to have followed a similar pattern of evolution. As a consequence, their layered structure must have been the result of differentiation and element partitioning that operated during and after their assembly from colliding planetary embryos. The differentiation came about as a result of melting following energy release during these collisions.

There are several possible sources of heating for a terrestrial planet. **Primordial heat** is that retained from processes operating in the early stages of planetary evolution, and represents one of the important heat sources within terrestrial-like bodies. Primordial heat includes that derived from the collision and assembly of planetary embryos, and that delivered to the surface by incoming impactors after the planet had assembled. It also includes heat released by the separation of denser components during core formation.

Internal heat generation within terrestrial planets such as Earth is mainly **radiogenic heating** which is the result of radioactive decay of ^{235}U , ^{238}U , ^{232}Th and ^{40}K in their silicate-rich mantle and crustal layers. The amount of radioactive decay was greater early in a planet's evolution because there would have been considerably more radioactive elements present. This radiogenic heat would have been augmented by the decay of short-lived isotopes such as ^{26}Al in those early stages.

Tidal heating is a further process whereby internal heat can be generated over the long term. In some instances, notably large orbiting satellite bodies such as the moons of Jupiter, tidal heating becomes the dominant process of generating internal heat.

Those bodies where heat is either more efficiently retained or continually generated in significant amounts are geologically active. In larger bodies, radiogenic sources are likely to continue to be important because they contain a greater mass of radiogenic elements to begin with, and because cooling is less efficient due to a lower surface area to volume ratio. The heat retained or generated within a terrestrial-like planetary body represents a key control in shaping the nature of the planetary surface. It also controls the rapidity and extent to which planetary resurfacing has occurred, or continues to occur.

2.12.2 Giant planets

The **giant planets** of the Solar System are Jupiter, Saturn, Uranus and Neptune. Data on the interiors of the giant planets can be obtained from measurements of density, gravitational field, magnetic field, emitted heat and atmospheric composition.

Jupiter and Saturn probably do not have a definite liquid or solid surface. Current models of Jupiter and Saturn distinguish five layers, as shown in Figure 2.28. The two innermost layers constitute a core of rocky and icy materials. This core is surrounded by layers that are mostly hydrogen and helium, which account for most of the planet's mass. The layer adjacent to the core in Jupiter and Saturn is

predicted to contain hydrogen in a metallic state. The deep interiors of both Jupiter and Saturn are very hot (over 15 000 K in the case of Jupiter).

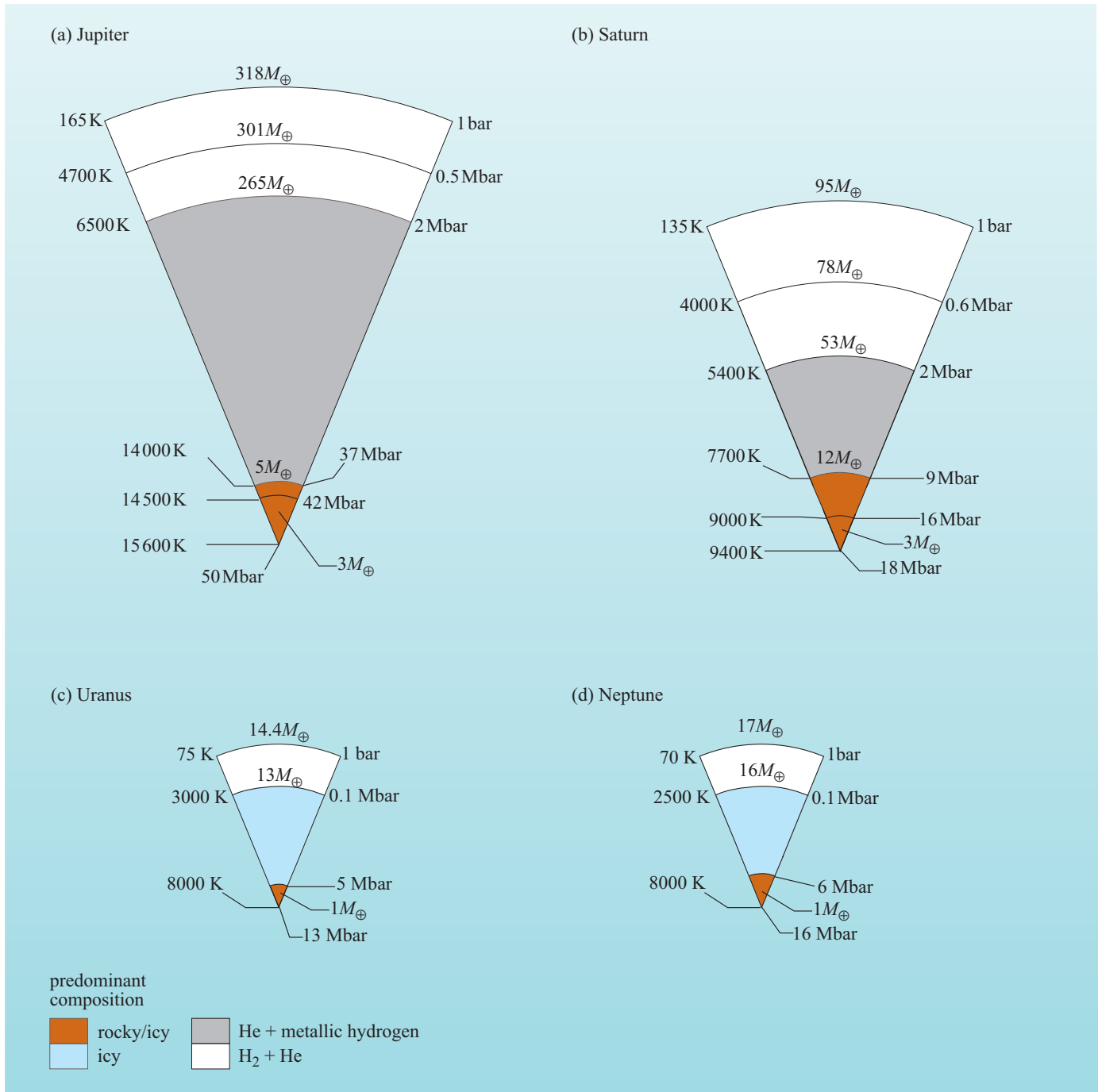


Figure 2.28 The internal structure of giant planets.

Uranus and Neptune may not have a definite liquid or solid surface. They may have rocky cores but current models suggest that rocky and icy materials are not completely differentiated. Surrounding the core is a mantle of mainly icy materials and around this is a layer of mainly hydrogen and helium. Overall, these two planets are less dominated by hydrogen and helium than Jupiter and Saturn

and the layers are probably less differentiated in composition.

Jupiter, Saturn and Neptune emit more energy than they receive from the Sun (heat excess). The heat excess of Jupiter is thought to be due to the continuing escape of original accretional heat and heat of differentiation. Saturn's heat excess is thought to have an additional contribution from helium droplets separating out from metallic hydrogen and sinking. Neptune and Uranus are both thought to have internal energy arising from continuing heat of differentiation. The cause of the heat excess for Neptune is still debatable. The cause of the lack of heat excess for Uranus may be associated with its unusual spin-axis inclination.

The magnetic fields of Jupiter and Saturn are believed to originate in the shell of liquid metallic hydrogen. The magnetic fields of Uranus and Neptune are thought to originate in a shell of liquid icy material containing the ions H_3O^+ , OH and NH_4^+ .

The atmospheres of the giant planets have hydrogen, H_2 , and helium as their major components. Other molecules detected are reduced forms of the heavier elements, for example CH_4 and NH_3 . Most of the molecules in the atmospheres are detected by IR or UV spectroscopy. The Galileo probe used mass spectrometry to obtain the relative abundances of molecules in the region of atmosphere it entered. The outermost cloud layer can be identified as ammonia on Jupiter and Saturn. Clouds of methane have been observed in the atmospheres of Neptune and Uranus.

Models assuming chemical equilibrium can predict the composition of the lower cloud layers, but these compositions have not been positively identified by observation. The Galileo probe detected a very tenuous cloud which could be part of the ammonium sulfide cloud layer.

The variation of temperature with depth on the giant planets divides the atmospheres into two regions. In the lower part (the **troposphere**) the temperature decreases the further out from the centre we go. The decrease is close to the adiabatic lapse rate, except for Uranus where the rate of decrease is slower. In the upper layers of the atmosphere (the **thermosphere**) the temperature increases with distance from the centre. Wind velocities on the giant planets are measured, remotely, by tracking the movement of cloud features. These measurements will give a velocity that includes the rotation speed of the planetary interior and so this has to be subtracted. The rotation speed of the interior can be measured from radio bursts.

On Jupiter and Saturn, there is evidence for a series of deep convection cells giving rise to the observed pattern of wind velocities. On Jupiter, major changes in wind velocity correlate with the boundaries between different coloured bands. There is no such correlation on Saturn.

Jupiter and Saturn have positive wind velocities at the equator. Equatorial wind velocities can be very high; up to 500 m s^{-1} on Saturn. The Galileo probe measured wind velocities on Jupiter directly. These were higher than the values obtained by Voyager but were only for one latitude. Neptune has a large negative equatorial wind velocity, and extrapolation suggests that Uranus has a negative equatorial wind velocity too.

Jupiter, Saturn, Uranus and Neptune all have large magnetospheres produced when the solar wind and IMF interact with the planetary magnetic fields.

The main features of the magnetospheres are similar to those of the Earth's magnetosphere. It contributes to Jupiter's magnetosphere. The large angles between the magnetic and rotation axes of Uranus and Neptune cause the magnetic field lines to vary substantially with time.

2.13 Extrasolar planets and how to find them

To complete this brief summary of planetary science, we turn to the astrophysics of planets that orbit other stars. These are variously known as **extrasolar planets** or simply **exoplanets**. Until 1995, no planets had been discovered around main sequence stars (other than the Sun!). The detection of exoplanets was made possible through improvements in telescope and detector technology, and by the fact that many exoplanetary systems do not look much like our Solar System.

In principle, exoplanets can be detected by radiation reflected from the star around which they are in orbit. However, in the vast majority of cases, current technology cannot resolve the star and planet because their difference in brightness is so huge. Most of the systems in which it is currently possible to image a planet directly are those where the planet and star are widely separated, where the star is very faint (i.e. a brown dwarf) and where the planet is very massive (a few Jupiter masses), and where the observation is performed in the infrared where the contrast is less extreme.

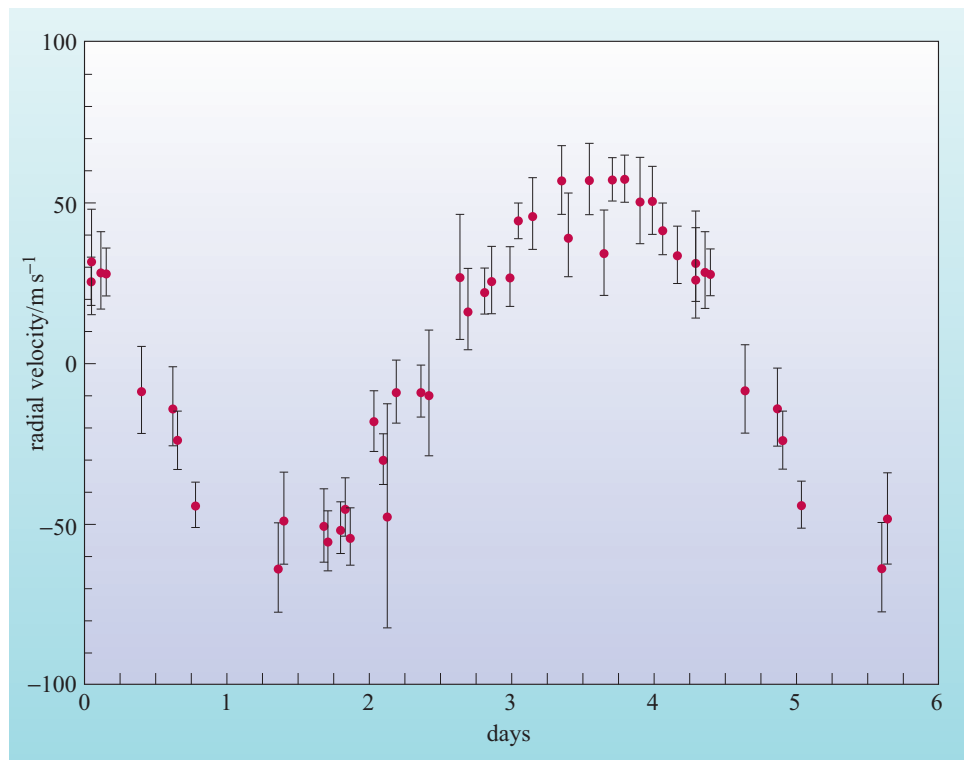


Figure 2.29 Measurements of the radial velocity of the star 51 Pegasi over one orbit. (S. Korzennik, Harvard University, Smithsonian Center for Astrophysics).

The technique by which the first exoplanet was found relies on the Doppler effect for its success. The presence of a planet in orbit around a star will cause the star to execute a (relatively small) orbit around the centre of mass of the system that can be detected by the technique of Doppler spectroscopy. By carefully monitoring

the wavelength of spectral lines from the photosphere of the star, its apparent motion towards and away from us can be detected, as the wavelength periodically shifts back and forth. The velocities induced by the presence of a planet are typically rather small – ‘stellar wobbles’ of only a few tens of metres per second or less are typical. The radial velocity curve for the first star to be found to host an exoplanet is shown in Figure 2.29. The amplitude of the radial velocity curve allows the mass ratio between the star and planet to be found, subject to an unknown factor of $\sin i$, which accounts for the inclination of the orbit to our line of sight. Estimating the mass of the star (from its spectral type) then allows an upper limit to the mass of the planet to be calculated. This technique requires individual stars to be targeted for observations and as such it is not necessarily very cost-effective in terms of discovering exoplanets.

A more economical technique for detecting exoplanets is looking for the occultation of the host star when an exoplanet transits in front of it. For such observations, many thousands of stars may be targeted in each exposure. For those systems in which the orbit is virtually edge-on to our line of sight, a Jupiter sized exoplanet will occult around 1% of the light from a Sun-like star, causing a periodically repeating small dip in the star’s lightcurve. The first star to be seen to exhibit such a dip was HD 209458. The transit lightcurve of this star is shown in Figure 2.30. Careful measurement of the characteristic U-shaped dip allows the radius of the transiting exoplanet to be determined. By carefully examining the lightcurves of millions of stars, using both ground-based and space-based surveys, several thousand exoplanets have now been found by this technique.

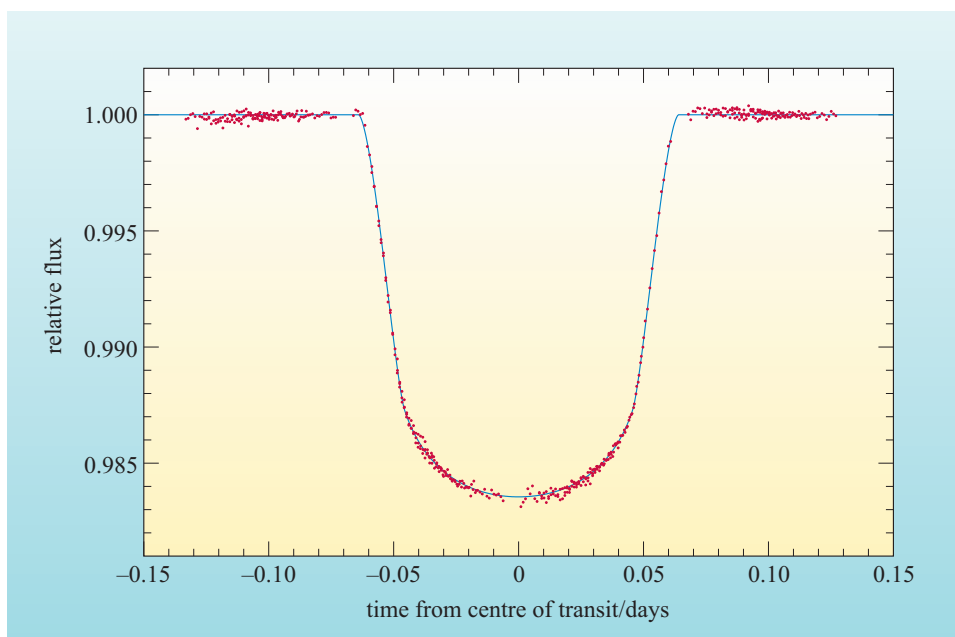


Figure 2.30 A lightcurve showing a planetary transit for the star HD 209458, as observed by the Hubble Space Telescope.

Any planet found to produce transits, will necessarily have an orbit close to edge-on ($i = 90^\circ$), and can subsequently be followed up with radial velocity spectroscopy, if it is bright enough. For such systems both the radius and the mass can be found (from the transit and from the radial velocity curve respectively), so allowing the density of the planet to be calculated.

Another technique which has successfully discovered around a hundred

exoplanets relies on the phenomenon of **gravitational microlensing**. In his theory of General Relativity, Einstein proposed that space is curved in the presence of massive objects. This distortion of space causes the path of light in the vicinity of a massive object to change. Gravitational microlensing is the name given to the effect whereby a background star apparently brightens and then fades over the course of a few weeks when a foreground star passes directly in front of it. If the foreground star has a planet in orbit around it, and if the position of the planet happens to line up with the background star too, then an additional microlensing ‘spike’ may be seen superimposed on the stellar microlensing profile (see Figure 2.31). The amplitude and duration of the spike can give information about the mass of the planet.

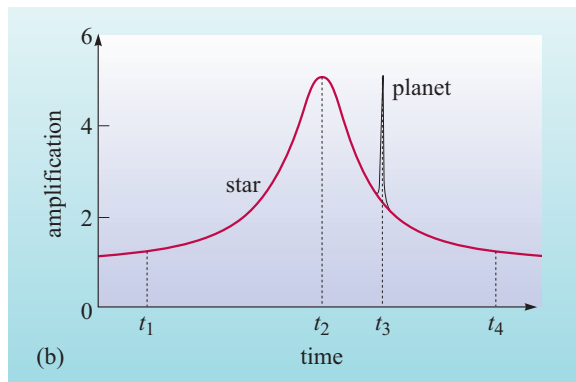


Figure 2.31 If a planet is in orbit around the lensing star and happens to line up with the lensed star, then an extra ‘spike’ is produced superimposed on the lightcurve.

The first exoplanets discovered tended to be very close to their host stars (and hence had very short orbital periods – of the order a few days) and had masses similar to that of Jupiter. These ‘hot Jupiters’, as they became known, were the first to be found, simply because they are the easiest to detect. The larger the planet and the smaller its orbit, the larger the amplitude of the star’s radial velocity curve and the deeper the transit. Also, shorter period orbits are easier to detect because one does not have to observe for so long in order to find them.

Most of the Sun-like stars within a hundred parsec or so of the Sun have been observed by Doppler spectroscopy for long enough to reveal any planets of the order of Jupiter’s mass in orbits with periods less than a decade or so. Furthermore, transit surveys with baselines of a few years are underway to monitor all of the millions of bright ($m_V < 13$) stars in the sky. As a result of these programmes, several thousand exoplanets have now been discovered and it seems that most stars are likely to host planets.

At the time of writing, the known planets range from around the mass of the Earth to around 20 Jupiter masses, with around half of them in orbits that are closer to their star than the Earth is to the Sun. Many systems are known to host two or more planets, and planets are also found in binary or triple star systems. The science of exoplanets is currently the most rapidly moving field in astrophysics with new discoveries being reported every month.

2.14 Astronomical telescopes

This final section of Chapter 2 considers how the characteristics of astronomical telescopes affect what we can learn about the stars and planets (and for that matter

galaxies and clusters) that are observed with them. We concentrate here on telescopes that operate in the *optical* region of the electromagnetic spectrum, but end the section with a brief mention of telescopes that operate in other wavebands.

2.14.1 Telescope characteristics

One of the key benefits of using a telescope is that it enables fainter objects to be detected than with the naked eye alone. The **light-gathering power** of a simple telescope used with an eyepiece is defined as

$$\text{light-gathering power} = (D_o/D_p)^2 \quad (2.24)$$

where D_o is the diameter of the objective (or primary) lens (or mirror) of the telescope and D_p is the diameter of the eye's pupil, assuming that all the light passing through the objective enters the eye. This is proportional to the light-gathering area of the objective lens or mirror of the telescope. Clearly, the larger the aperture the more light is collected and focused into the image, and therefore fainter stars can be detected.

The **field-of-view** of a telescope is the angular area of sky that is visible through an eyepiece or can be recorded on a detector, expressed in terms of an angular diameter. When a telescope is used with an eyepiece, the angular field-of-view is equal to the diameter of the field stop (i.e. the diameter of the aperture built into the eyepiece) divided by the effective focal length of the objective mirror or lens. In symbols:

$$\theta = D/f_o \quad (2.25)$$

where the angular diameter of the field-of-view θ is in radians. When a telescope is used with a detector in place of an eyepiece, the determining factor here is the linear size of the detector itself, rather than the field-stop diameter.

The **angular magnification** indicates by what factor the angular dimension (e.g. angular diameter) of a body is increased. So if you were to observe Saturn through a telescope, you would be benefiting from a high angular magnification which makes the image appear larger even though it is squeezed into the tiny space of your eyeball. The angular magnification M of an astronomical telescope, used visually, is defined as the angle subtended by the image of an object seen through a telescope, divided by the angle subtended by the same object without the aid of a telescope. By geometry, this can be shown to be equivalent to

$$M = f_o/f_e \quad (2.26)$$

where f_o is the effective focal length of the objective lens or mirror system and f_e is the focal length of the eyepiece lens.

Notice that the angular magnification and field-of-view of a telescope both depend on the focal length of the objective lens or mirror. However, increasing f_o will *increase* the angular magnification but *decrease* the field-of-view, and vice versa.

The nearest equivalent definition to angular magnification that is applicable to telescopes used for imaging onto a detector is the **image scale** (sometimes called the *plate scale*). Because of the importance of angular measures, the image scale quoted by astronomers indicates how a given angular measure on the sky corresponds to a given physical dimension in an image. The most common

convention is to state how many arcseconds on the sky corresponds to 1 mm in the image. Fortunately, it is very easy to calculate the image scale for any imaging system, as it depends on only one quantity: the focal length f_o of the imaging system. The image scale I in arcseconds per millimetre is given by

$$I/\text{arcsec mm}^{-1} = \frac{1}{(f_o/\text{mm}) \times \tan(1 \text{ arcsec})} \quad (2.27)$$

Note that as the image on the detector becomes *larger*, the numerical value of I becomes *smaller*. Also, since there are 206 265 arcseconds in 1 radian of angular measure, and since $\tan \theta \sim \theta$ for small angles measured in radians, Equation 2.27 can be rewritten as

$$I/\text{arcsec mm}^{-1} = \frac{206\,265}{(f_o/\text{mm})} \quad (2.28)$$

A final important characteristic of astronomical telescopes is their angular resolution. The image of a point-like source of light (such as a distant star) obtained using a telescope will never be a purely point-like image. Even in the absence of aberrations and atmospheric turbulence to distort the image, the image of a point-like object will be extended due to diffraction of light by the telescope aperture. The bigger the aperture, the smaller is the effect, but it is still present nonetheless. The intensity of the image of a point-like object will take the form shown in Figure 2.32. The structure shown here is referred to as the **point spread function** (PSF) of the telescope. Lens or mirror aberrations and atmospheric turbulence will each cause the width of the PSF to broaden, and may cause its shape to become distorted too. However, in the ideal case when neither aberrations nor turbulence is present, the telescope is said to be diffraction-limited, and its PSF has the form shown. The width of the PSF, in this idealized case, is inversely proportional to the aperture diameter of the telescope.

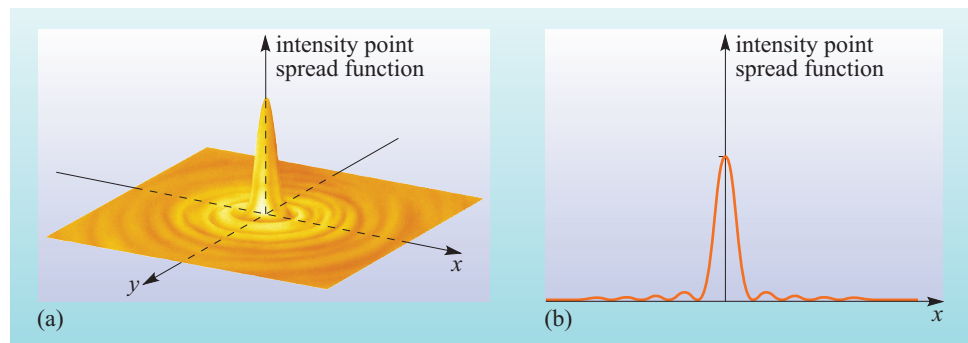


Figure 2.32 The image of a point-like object is not point-like even under ideal conditions. (a) The vertical direction represents image intensity. The point spread function of a point-like object under ideal conditions consists of a central peak surrounded by concentric ripples. The two-dimensional PSF has the circular symmetry of the telescope aperture. (b) A slice through (a) along one axis.

Using the idea of the diffraction-limited PSF, we can also define the (theoretical) limit of angular resolution for an astronomical telescope. This is the minimum angular separation at which two equally bright stars would just be distinguished by an astronomical telescope of aperture D_o (assuming aberration-free lenses or mirrors and perfect viewing conditions). As shown in Figure 2.33b, at a certain

separation, the first minimum of the PSF of one star will fall on the peak of the PSF of the other star. At this separation, the two stars are conventionally regarded as being just resolved.

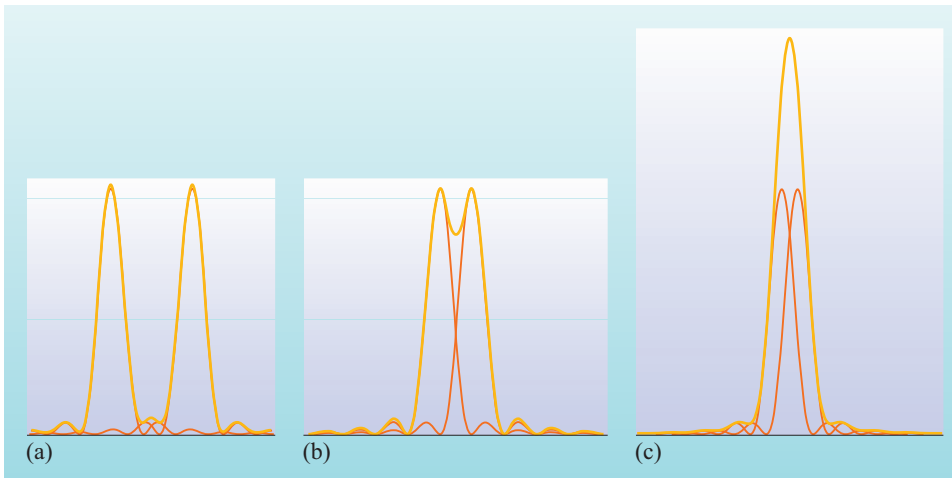


Figure 2.33 The images of the two stars in (a) are clearly resolved, whereas those in (c) are unresolved. In (b), the first minimum of one PSF coincides with the peak of the other PSF. At this separation the stars are said to be just resolved.

The angular separation corresponding to the situation in Figure 2.33b is given by

$$\alpha_c = 1.22\lambda/D_o \quad (2.29)$$

where α_c is the limit of angular resolution measured *in radians* and λ is the average wavelength of light contributing to the image. As noted above, the limit of angular resolution arises due to diffraction of light by the telescope aperture and represents a fundamental limit beyond which it is impossible to improve.

Exercise 2.7 (a) Calculate the ratio of the light-gathering power of a telescope of diameter $D_o = 5.0$ m to that of one with a diameter of 1.0 m. (b) Compare the (theoretical) limits of angular resolution of these two telescopes (at the same wavelength).

In practice, for ground-based astronomy, the limit on angular resolution is usually that imposed by **astronomical seeing** rather than the diffraction limit of the telescope. Astronomical seeing describes the effects of the blurring due to turbulence in the Earth's atmosphere. At the very best locations on the Earth, such as the top of a high mountain, the astronomical seeing may result in a point spread function whose central peak has a full width at half maximum of around 0.5 arc seconds.

Optical telescopes are often equipped with **spectrometers** which disperse the light in order to display a spectrum in the image plane. The heart of a spectrometer is usually either a prism or a diffraction grating. In the latter case, the **diffraction equation** may be written as

$$d \sin \theta_n = n\lambda \quad (2.30)$$

where λ is the wavelength of light, d is the spacing between adjacent lines of the grating, n is the diffraction **order** and θ_n is the angle through which light of the given wavelength is diffracted in the given order of the spectrum.

- Imagine that you have a spectrometer whose grating has 500 lines per mm, and is set up with the light falling on it at normal incidence. Calculate the angles at which light of wavelength 400nm and 600nm will be diffracted in the first spectral order.
- The line spacing is $d = 1/(500 \text{ mm}^{-1}) = 0.002 \text{ mm} = 2 \mu\text{m}$ and the spectral order, $n = 1$. Equation 2.30 then gives $\sin \theta = 1 \times \lambda/2 \mu\text{m}$. So for $\lambda = 400 \text{ nm}$ or $0.4 \mu\text{m}$, $\theta = 11.5^\circ$ and for $\lambda = 600 \text{ nm}$ or $0.6 \mu\text{m}$, $\theta = 17.5^\circ$
- How would the angular dispersion of the spectrometer above change if it were operated such that the second order spectrum was observed?
- The sine of the angle of diffraction would be doubled at each wavelength. So the short wavelength light would now have a diffraction angle of 23.6° and the long wavelength light would now have a diffraction angle of 36.9° . The overall dispersion of the spectrum would therefore *increase* from $(17.5^\circ - 11.5^\circ) = 6.0^\circ$ to $(36.9^\circ - 23.6^\circ) = 13.3^\circ$.

2.14.2 Telescopes in other parts of the electromagnetic spectrum

Although the discussion above has concentrated on the characteristics of telescopes that operate in the optical region of the electromagnetic spectrum, similar principles apply also to those that work in the near infrared or near ultraviolet wavebands too. The Earth's atmosphere is reasonably transparent to optical light as well as some wavelengths of infrared light, although many regions of the infrared spectrum are absorbed in the Earth's atmosphere, principally by water vapour. It is for this reason that ground-based infrared observatories are sited at high altitude, dry sites around the world. The Earth's atmosphere does not transmit much ultraviolet light, and for this reason ultraviolet astronomy is carried out using space-based satellite observatories.

In fact the only other regions of the electromagnetic spectrum to which the Earth's atmosphere is transparent are in the microwave and radio wavebands. Here the telescopes are of rather different nature, comprising large parabolic dishes which focus the incoming radio waves onto detectors placed at their focus. As with optical telescopes, the angular resolution of such a telescope also depends on the diameter of the dish. However, with radio telescopes it is possible to construct vast arrays of individual telescopes that operate as a single instrument. Such arrays can operate with baselines between individual dishes of anything from a few hundred metres to thousands of kilometres and as a result obtain angular resolutions of milli-arcseconds or better.

For high energy regions of the electromagnetic spectrum, space-based astronomy is the solution. Many satellite observatories have been built that operate in the ultraviolet, X-ray and gamma-ray regions of the spectrum. Ultraviolet telescopes can use conventional focussing optics, however, this does not work for X-rays or gamma-rays. Many X-ray observatories focus X-rays using gold-plated, nested, conical mirrors. The X-rays reflect off these mirrors at shallow angles (referred to as grazing incidence) and are focussed onto imaging detectors which count individual photons. Gamma-ray detectors also count individual high energy photons, but there is no way of focussing these to form images. One ingenious

technique to form gamma-ray images of the sky uses a grid of opaque and transparent cells in front of the telescope to cast a ‘shadow’ of the gamma-ray sources onto a detector from which the spatial distribution of the gamma-ray emitting sources on the sky may be reconstructed.

This is only an extremely brief summary of non-optical astronomy but it serves as a reminder that there is far more information to be gleaned about astronomical objects than using optical telescopes alone. For instance, infrared astronomy is becoming increasingly important as we look to the very early Universe, where the light from distant galaxies is hugely redshifted. It is to the study of galaxies and the Universe as a whole that we turn in the next Chapter.

Summary of Chapter 2

1. Although astronomers do use SI units, they also use cgs units and frequently employ non-SI units such as the parsec or astronomical unit (both for distance), the ångström (for wavelength), and the erg or electronvolt (both for energy).
2. Astrophysical quantities are often expressed in terms of units relative to the Sun (such as L_{\odot} , M_{\odot} or R_{\odot}). Variables may also be expressed in terms of a power of ten and a particular unit. For example, $R_6 = 3.0$, where $R_6 = R/10^6$ m, implies that the radius in question $R = 3.0 \times 10^6$ m.
3. A variety of systems are used for naming stars. Some are based on simple letter designations in order of brightness, others on numerical order in a catalogue, and yet others are based on the coordinates of a star on the sky.
4. Positions of celestial objects are specified in terms of their right ascension (α), measured in hours, minutes and seconds, and their declination (δ), measured in degrees, arc minutes and arc seconds.
5. The magnitude of the tangential velocity of a star is related to its distance (d) and proper motion (μ) by $v_t = d \tan \mu$. The magnitude of the radial velocity is obtained from the Doppler shift as $v_r = c \Delta\lambda/\lambda$.
6. The distance d to a star is related to its trigonometric parallax (π) by $d/\text{pc} = 1/(\pi/\text{arcsec})$.
7. The spectra of stars are essentially black-body continua with spectral lines superimposed. The spectral classification sequence (O B A F G K M) reflects a progression in temperature from hot ($\sim 40\,000$ K) to cool (~ 3000 K) stars. O/B-type stars have the strongest helium lines, A/F-type stars the strongest hydrogen lines, G/K-type stars the strongest lines from ionized metals, and M stars have molecular lines.
8. The luminosity classification of stars includes class I (bright supergiants), class III (giants), class V (main sequence dwarfs) and class VII (white dwarfs).
9. Hot stars on the main sequence have relatively large mass, radius and luminosity. Cool stars on the main sequence have relatively small mass, radius and luminosity. A star’s luminosity L , radius R and effective temperature T_{eff} are related by $L = 4\pi R^2 \sigma T_{\text{eff}}^4$ where σ is the Stefan-Boltzmann constant.

10. The equivalent width of a spectral line is a measure of its strength. The actual width of a spectral line may be the result of Doppler broadening and can indicate the range of speeds of the atoms in which the line originated.
11. Apparent and absolute magnitudes may be measured in any one of several wavebands, conventionally labelled U B V R I J H K from the near ultraviolet to the near infrared. Apparent magnitude is related to flux by

$$m_1 - m_2 = 2.5 \log_{10}(F_2/F_1) = -2.5 \log_{10}(F_1/F_2)$$

whilst absolute magnitude is related to luminosity by

$$M_1 - M_2 = 2.5 \log_{10}(L_2/L_1) = -2.5 \log_{10}(L_1/L_2)$$

Apparent and absolute magnitudes are related by

$$M = m + 5 - 5 \log_{10} d - A$$

where d is the distance to the star in parsec and A is the interstellar extinction.

12. In the absence of extinction, flux and luminosity are related by $F = L/4\pi d^2$.
13. An astronomical colour is the difference between two apparent or absolute magnitudes in different wavebands, and so is equivalent to the ratio of two fluxes or luminosities in different parts of the spectrum.
14. Interstellar extinction is progressively less at longer wavelengths, so the optical spectra of astronomical objects are generally reddened.
15. The Hertzsprung–Russell diagram plots the positions of stars according to their luminosity (or absolute magnitude) and their temperature (or spectral class or colour). The main features of H–R diagrams are a main sequence running from top left (high luminosity and temperature) to bottom right (low luminosity and temperature); giant branches occupying the top right of the diagram; and degenerate stars occupying the bottom left.
16. Masses of stars can in general only be directly measured in binary systems by using Kepler’s third law

$$\frac{(a_1 + a_2)^3}{P^2} = \frac{G(m_1 + m_2)}{4\pi^2}$$

and the definition of mass ratio $q = m_1/m_2 = a_2/a_1 = v_2/v_1$.

17. Stars are formed by the collapse of fragments of dense molecular clouds. When the central regions become hot enough for nuclear fusion to be initiated, the star is born on the zero-age main-sequence of the H–R diagram. The star remains on the main sequence whilst undergoing hydrogen fusion in its core by the proton–proton chain or the CN cycle. When hydrogen in the core is exhausted, helium fusion may begin, and occurs by the triple-alpha process. In low-mass stars this happens by way of an explosive helium flash. Other nuclear fusion reactions are subsequently possible in massive stars. When nuclear fuel is exhausted the star ends its life in one of several ways, depending on its mass. Low-mass stars shed their outer layers as planetary nebulae and the core becomes a white dwarf. Massive stars explode as supernovae and the core collapses to form a neutron star or black hole.

18. During the formation of stellar systems, planets form from the nuclear-processed material expelled into the interstellar medium by earlier generations of supernovae. The Solar System is broadly separated into terrestrial planets (predominantly rocky and metallic in composition) in the inner region, and giant planets (predominantly gaseous) in the outer regions. The terrestrial planets have metallic cores overlaid by rocky mantles; some of them have atmospheres. The giant planets have small rocky cores, overlaid by either a thick layer of fluid helium and metallic hydrogen (Jupiter and Saturn) or liquid icy materials including water, ammonia and methane (Uranus and Neptune). Above this in each case is a thick atmosphere composed predominantly of hydrogen and helium.
19. Exoplanets (or extrasolar planets) are known to exist around (probably) most Sun-like stars. They have been detected mainly by the techniques of Doppler spectroscopy (measuring the radial velocity of the host star), transit photometry (measuring the dip in a star's lightcurve due to occultation by the planet) and gravitational microlensing (measuring the additional brightening caused by a planet in orbit around a foreground lensing star).
20. The main parameters of an optical telescope are its light-gathering power, its field-of-view, its angular magnification or image scale and its limit of angular resolution. Increasing the size of the objective lens or mirror of a telescope increases its light gathering power and improves its angular resolution. Increasing the focal length of the objective lens or mirror decreases the telescope's field-of-view and increases its angular magnification.
21. The performance of a grating spectrometer is governed by the diffraction equation $d \sin \theta_n = n\lambda$.

This page is intentionally left blank to ensure that subsequent chapters begin on an odd-numbered page.

Chapter 3 Galaxies and the Universe

Introduction

This chapter will allow you to revise and consolidate your knowledge of cosmology and the astrophysics of galaxies. If you have recently completed the OU's Level 2 astronomy course (S282), then a large part of this chapter will be familiar to you, but perhaps not all of it.

3.1 The Milky Way – our galaxy

A **galaxy** is simply a collection of stars and clouds of dust and gas that are bound together by their mutual gravitational attraction. Our own galaxy (sometimes called the Galaxy or the **Milky Way**) is a typical example of a so-called spiral galaxy. As with stars, virtually the only thing astronomers can measure is the electromagnetic radiation emitted by the stars and gas of which the galaxy is composed.

The Milky Way consists of three major, directly detectable structural components (Figure 3.1): a disc, a surrounding halo and a central nuclear bulge. Pervading the Galaxy, between the stars, is the gas and dust of the interstellar medium. These components are embedded in a massive cloud of **dark matter**, currently detectable only through its gravitational influence. The directly detectable matter consists mainly of stars ($\sim 90\%$ of the Galaxy's visible mass), gas ($\sim 10\%$) and dust ($\sim 0.1\%$). There are about 10^{11} stars in all, with (very roughly) a total mass of $10^{11} M_{\odot}$. The gas is almost entirely hydrogen and helium, in a 3:1 ratio by mass. The hydrogen occurs in the form of molecules (H_2), atoms (referred to as H or HI) or ions (referred to as HII) according to local conditions. The disc of the Milky Way is about 30 kpc in diameter and 1 kpc thick. The nuclear bulge is roughly spherical and has a diameter of about 6 kpc. The halo is also thought to be roughly spherical; its size is difficult to determine but estimates of 40 kpc or so are common.

The stars of the Milky Way may be divided into a number of populations, each of which predominates in a particular region of the Galaxy. The youngest stars, those of extreme **population I**, are found mainly in the spiral arms of the disc. The oldest stars, those of extreme **population II**, are found mainly in the ancient globular clusters of the halo. The Sun is one of the intermediate population I stars, located in the disc between 7.5 kpc and 10 kpc from the centre of the Galaxy.

The **disc** of the Galaxy is in a state of differential rotation, with stars in the vicinity of the Sun taking about 2×10^8 years to make a complete orbit of the Galactic centre. The disc is thought to be threaded by bright spiral arms, and there is also a central bar. The spiral arms are sites of active star formation. Attempts to trace the arms make use of young, short-lived features of the disc such as bright **HII regions**, young **open clusters** of stars and associations of O and B type stars. It is thought that the large-scale patterns of star formation that highlight the spiral arms might be caused by density waves – relatively slow moving patterns of density enhancement that rotate, as if rigidly, around the Galactic centre.

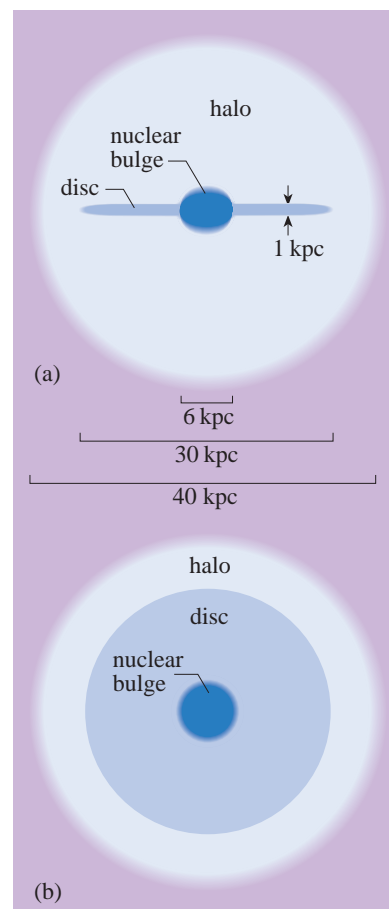


Figure 3.1 The structure of our galaxy showing (a) edge-on and (b) face-on views of the major structural components – the disc, the halo and the nuclear bulge.

Differentially rotating disc material, particularly giant molecular clouds, entering such regions of enhanced density would undergo collisions with gas already there. Such collisions may trigger the birth of stars which, given the size of the giant clouds, would be expected to form in clusters.

The Galactic **halo** is the most extensive of the directly detectable structural components of the Milky Way. It is an ancient and relatively inactive part of the Galaxy. The main constituents of the halo are old stars of population II, containing few elements other than hydrogen and helium. Stars with this composition are said to have low **metallicity**, where, in astronomical terms, all elements other than hydrogen and helium are referred to as ‘metals’. The total mass of the stars in the halo is about $10^9 M_{\odot}$ (or 1% of that of the Galaxy as a whole) and about 1% of the halo stars are contained in **globular clusters**. These are tight spherical swarms of stars up to 50 pc across and each containing $10^5 - 10^6$ members.

The **nuclear bulge** is the most enigmatic part of the Milky Way due to the high level of optical extinction between us and it. The bulge seems to be a spheroid of equatorial diameter about 6 kpc. Its outer regions rotate at about 100 km s^{-1} and its total mass is around $10^{10} M_{\odot}$. The bulge consists mainly of population II stars, though their metallicities seem to be unusually high in many cases. Within the bulge a number of different features can be identified, including the radio source Sagittarius A. This complex radio source surrounds the centre of our galaxy, and at its heart lies the object referred to as Sgr A*. Detailed infrared studies over the last few years have revealed the motions of several individual stars orbiting around Sgr A*. By measuring the speeds of these stars, and their distances from the orbital centre, it has been deduced that Sgr A* is a black hole with a mass of around $4 \times 10^6 M_{\odot}$. This black hole is not currently accreting much matter and so may be regarded as ‘quiescent’, unlike the black holes in active galaxies which will be described in later sections.

3.2 Other galaxies

Note that much of the terminology discussed in Chapter 2 in relation to stars, also applies to individual galaxies. To refer to the position of a galaxy, its right ascension and declination are used, its relative velocity may be measured by the Doppler shift of its spectral lines and its distance may be referred to in parsecs (or more likely megaparsecs or gigaparsecs). Moreover the spectrum of a galaxy will often look rather similar to the spectrum of a star, since the light from a galaxy is primarily the sum of the light from its constituent stars. We can also talk about the flux, luminosity, apparent magnitude, absolute magnitude and colour of a galaxy, in different wavebands, in a similar manner to that which is used for individual stars.

3.2.1 Classification of galaxies

Mainly according to their shape, most galaxies can be assigned to one of four different classes: elliptical, lenticular, spiral and irregular. In the modified form of the **Hubble classification** scheme (Figure 3.2), the spirals and lenticulars can be subdivided into barred and unbarred subclasses, whilst the spirals and ellipticals can be further subdivided into a number of Hubble types.

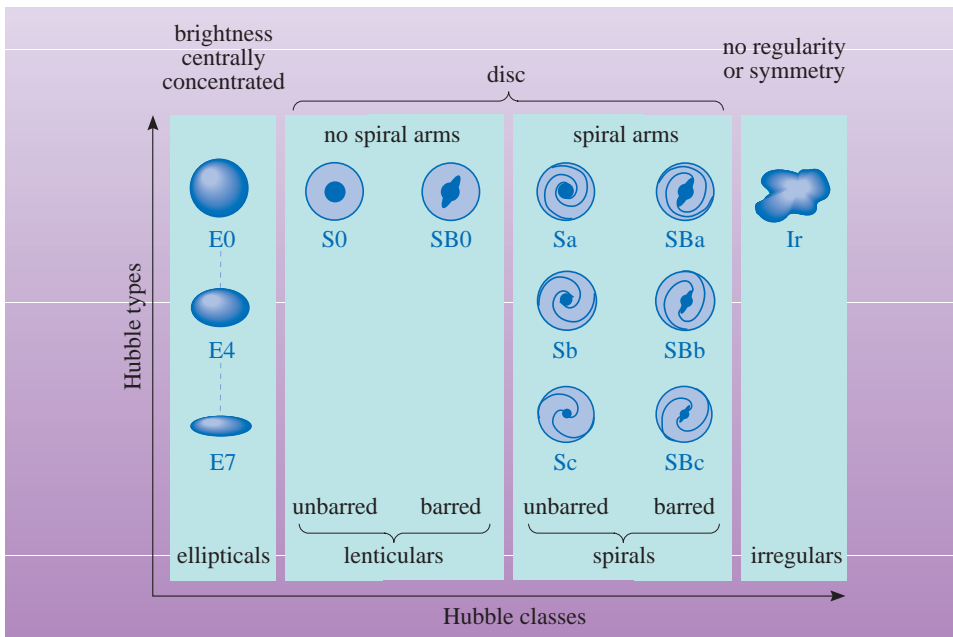


Figure 3.2 The Hubble classification scheme for galaxies.

Elliptical galaxies are essentially ellipsoidal distributions of old (population II) stars. Their three-dimensional shape is difficult to determine, but some at least appear to be triaxial ellipsoids with very little rotation. The largest galaxies are the cD galaxies – giant ellipticals which may have been formed in mergers and which are often found close to the centres of clusters of galaxies. Whereas most ellipticals are almost devoid of cold gas and dust, these giant elliptical galaxies can contain $10^9 M_{\odot}$ of hot gas ($> 10^6$ K). **Lenticular galaxies** appear to be an intermediate class between the most flattened of elliptical galaxies and the most tightly wound spirals. They show clear signs of a disc and a central bulge, but they have no spiral arms and little cold interstellar gas. **Spiral galaxies** have a disc, a central bulge and often a central bar. Within this class, spiral arms may be more or less tightly wound and the bulge may be more or less prominent in relation to the disc. **Irregular galaxies** are generally chaotic and asymmetric, though some exhibit a bar and others show traces of spiral structure. Amongst the nearest galaxies to our own, the commonest type by far are **dwarf galaxies** – a class which includes dwarf ellipticals, dwarf spheroidals and dwarf irregular galaxies. They contain only a few million stars and are difficult to observe because they look very similar to typical fields of foreground stars. As with stars, where the smallest are the most common, dwarf galaxies are almost certainly the most abundant type of galaxy in the Universe.

Galaxy names

The only galaxies beyond our own which are visible to the naked eye are two irregular satellite galaxies of the Milky Way known as the Small Magellanic Cloud and the Large Magellanic Cloud, and the spiral galaxy known as the Andromeda Nebula (or more properly, Andromeda Galaxy). As their names suggest, the two Magellanic clouds were first brought to the attention of Western science by Ferdinand Magellan (in 1519) – they are

visible only from the Southern Hemisphere and so had gone unnoticed by Europeans until then. The Andromeda Nebula takes its name simply from the constellation in which it lies. Other prominent galaxies are also named simply after their host constellation.

With the advent of telescopic observations, several other nebulae were discovered which were subsequently identified as galaxies. The first extensive catalogue of such objects was that compiled by Charles Messier in 1784. His original list of 103 objects included the Andromeda Galaxy (as M31) as well as other galaxies such as the Whirlpool Galaxy (M51) and the Sombrero Galaxy (M104). These names exemplify the way in which the brighter or visually spectacular galaxies are named – by a descriptive word which sums up their appearance.

A more extensive compilation – the New General Catalogue of Nebulae and Clusters of Stars followed in 1888. The original catalogue, published by J.L.E. Dreyer listed 7840 objects and was based on lists compiled by the Herschel family. So, for instance, the Andromeda Galaxy is also known as NGC224 in this designation. Dreyer added a further 5386 objects in his first and second Index Catalogues a few years later, giving rise to ‘IC’ designations for a few thousand more galaxies.

With the advent particularly of radio surveys of the sky from the 1960s onwards, many radio galaxies and quasars were discovered. The first to be detected were the brightest objects, and designated in a straightforward way. For instance, Virgo A is the brightest radio source in the constellation of Virgo, and identical with the giant elliptical galaxy M87. The third Cambridge survey (3C catalogue) is perhaps the most famous of these radio surveys, giving rise to names such as 3C273 for the brightest quasar in the sky. Another nomenclature commonly encountered is that resulting from the catalogue of active galaxies listed by B. E. Markarian in the 1970s. These are referred to by the designation ‘Mrk’, thus Mrk335 is an example of a type of active galaxy known as a Seyfert galaxy.

In recent years, the number of galaxies detected in various surveys, in different parts of the electromagnetic spectrum, has increased dramatically. It is no longer always practical to refer to galaxies simply by a number in a catalogue. Instead, galaxies are commonly referred to by a numerical name which encodes their position on the sky in terms of right ascension and declination. Thus PKS2155–304 is an active galaxy (a blazar) discovered by the Parkes radio survey and located approximately at right ascension: 21 h 55 min, declination: -30.4° .

Clearly, most galaxies will have a variety of different names in different catalogues. As an example, the following list shows some of the names currently associated with the radio galaxy Virgo A mentioned above:

M 87	NGC 4486	UGC 7654
2A 1228+125	3A 1228+125	APG 152
3C 274	4C 12.45	CTA 54
DA 325	DB 85	DGW65 57
DML87 747	DSB94 254	2E 2744
1ES 1228+12.6	EUVE J1230+12.3	2EUVE J1230+12.3
H 1228+127	H 1227+12	1H 1226+128
IRAS 12282+1240	1Jy 1228+12	1Jy 1228+126
MCG+02-32-105	Mills 12+1A	MRC 1228+126
NRAO 401	NRL 8	NVSS B122817+124004
PKS 1228+127	PKS J1230+1223	PKS 1228+12
PT56 24	RORF 1228+126	RX J1230.8+1223
3U 1228+12	4U 1228+12	VCC 1316
VPC 771	J123049.5+122328	J123049.5+122328
X Vir XR-1	Z 70 – 139	Z 1228.3+1240
1A 1228+12	3C 274.0	Cul 1228+12
DLB87 V12	2E 1228.2+1240	GIN 800
IRAS F12282+1240	1M 1228+127	NAME VIR A
PGC 41361	PKS 1228+126	RX J1230.1+1223
VDD93 163	X Vir X-1	

3.2.2 Origin and evolution of galaxies

Galaxies are thought to have formed as a result of tiny density fluctuations in the expanding cosmic gas produced by the Big Bang. These fluctuations grew in strength as over-dense regions attracted more matter due to their enhanced gravitational pull. Spiral galaxies formed as gas cooled and settled onto regions of higher density, with a disk forming due to the angular momentum of the material. Elliptical galaxies formed either by mergers of smaller spiral galaxies, or by gas cooling in regions of low angular momentum.

Using deep exposures of apparently ‘empty’ regions of the sky, such as the Hubble Deep Fields, astronomers can detect very faint galaxies which are very distant and therefore seen soon after their formation in the early Universe. Indeed, some images reveal sub-galactic objects which may be in the process of merging to form larger structures such as those we see today in our local neighbourhood. There is considerable evidence that the interaction of galaxies (including mergers and collisions) can be of great importance in shaping the galaxies we see. It almost certainly accounts for many of the distorted peculiar galaxies that are observed, and may be of more general importance in explaining the prevalence of ellipticals in many clusters of galaxies.

3.2.3 Measuring galaxy properties

Since galaxies are, in general, extended objects rather than point sources, it is usually most convenient to measure their **surface brightness**, or flux density per unit angular area. The surface density of a galaxy may therefore be expressed in units of $\text{W m}^{-2} \text{arcsec}^{-2}$ and will vary from point to point across its image.

Continuous lines passing through points of equal brightness are called **isophotes**. As it is often difficult to determine the edge of a galaxy, observations are often confined to the region within some specified isophote. Empirical relations between surface brightness and luminosity obtained from observations of nearby galaxies are then used to estimate the luminosity of distant galaxies based on their flux within a given isophote.

Galactic masses are generally hard to measure. For spiral galaxies, **rotation curves** may be used. These are a plot of the circular orbital speed of stars and gas in the galaxy (measured by Doppler shifts) against radial distance from the galactic centre. To determine the mass, the observed rotation curve is compared with a theoretical one, predicted by a model of the galaxy in which mass is distributed in a plausible way. The theoretical mass distribution is adjusted until a good fit between model and data is obtained, and the total mass of the galaxy is then obtained.

Rotation is relatively unimportant in elliptical galaxies, so to determine their mass, a method based on **velocity dispersion** is used instead. The velocity dispersion is a statistical quantity that characterizes the behaviour of a group of stars. Roughly speaking it provides a measure of the range of speeds of stars along a given line of sight. For elliptical galaxies, the velocity dispersion is expected to be proportional to $(M/R)^{1/2}$, where M is the mass of the galaxy and R is a scale length related to its radius. The result relies on the stars in the elliptical galaxy moving in random orbits while bound by each other's gravity – the generally well-ordered motion in spiral galaxies means that their masses cannot be estimated in this way.

3.3 The distances to other galaxies

Determining the distances of galaxies is of great importance in astronomy. Distance information can be crucial to the determination of other galactic properties; it plays a vital part in investigations of the large-scale distribution of galaxies; and it may provide the key to understanding the fate of the Universe as a whole. There are many different methods of distance determination. Those applicable to galaxies include: geometrical methods, such as those based on the diameters of material illuminated by supernovae; **standard candle** methods, such as those based on Cepheid variables, type Ia supernovae or the apparent magnitude of brightest cluster galaxies; and methods involving galactic properties such as the widths of 21 cm emission lines. An example of one of these methods is presented below.

Worked Example 3.1

Cepheids are a particular class of variable stars with the remarkable property that their period of pulsation is directly proportional to their absolute magnitude (or luminosity). This means that they can act as a kind of 'standard candle' and may be used to determine distances to other galaxies.

Figure 3.3 shows the period–luminosity relationship for Cepheid variable stars. A certain Cepheid variable star in the Andromeda galaxy is observed to have a period of 30 days and a peak apparent visual magnitude of

Essential skill:
Distance measurement using
Cepheid variables

$m_V = 20.0$. Assuming that the extinction to the Andromeda galaxy is $A_V = 0.2$, what is the distance to the Andromeda galaxy?

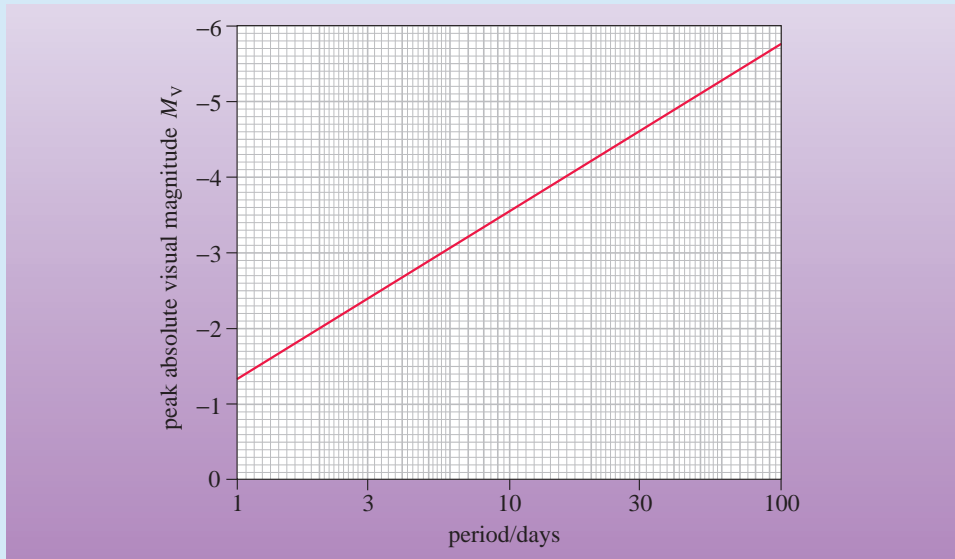


Figure 3.3 The period–luminosity relationship for Cepheid variables.

Solution

From Figure 3.3 the absolute magnitude of a Cepheid variable star with a period of 30 days is $M_V = -4.6$. So using Equation 2.10, the distance to the star can be found as

$$\log_{10} d = (m_V - M_V + 5 - A_V)/5 = (20.0 + 4.6 + 5 - 0.2)/5 = 5.88$$

So, $d = 7.6 \times 10^5$ pc. Since the star is within the Andromeda galaxy, which itself is small (a few tens of kpc) in relation to this distance, the distance to the Andromeda galaxy as a whole is about 760 kpc.

The distances to the most remote galaxies may be determined using the **Hubble law** which relates distance to redshift. The **redshift** of a spectrum of a galaxy is defined as

$$z = \frac{\Delta\lambda}{\lambda} \quad (3.1)$$

where $\Delta\lambda$ is the shift in wavelength of a particular feature observed in a spectrum and λ is the wavelength of that feature as it was emitted by the galaxy (or as observed in an Earth-bound laboratory). In the late 1920s Edwin Hubble demonstrated that there is a simple relationship between the redshifts of distant galaxies and their distance from us. This may be expressed as

$$z = \frac{H_0 d}{c} \quad (3.2)$$

where d is the distance to the galaxy, c is the speed of light and H_0 is a quantity now known as the **Hubble constant**. The relationship is illustrated in Figure 3.4. The best measurement of the Hubble constant currently available places its value

at $67.74 \text{ km s}^{-1} \text{ Mpc}^{-1}$ with an uncertainty of about $\pm 0.46 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Often, you will see the Hubble constant written as a pure number h defined as $h = H_0 / (100 \text{ km s}^{-1} \text{ Mpc}^{-1})$.

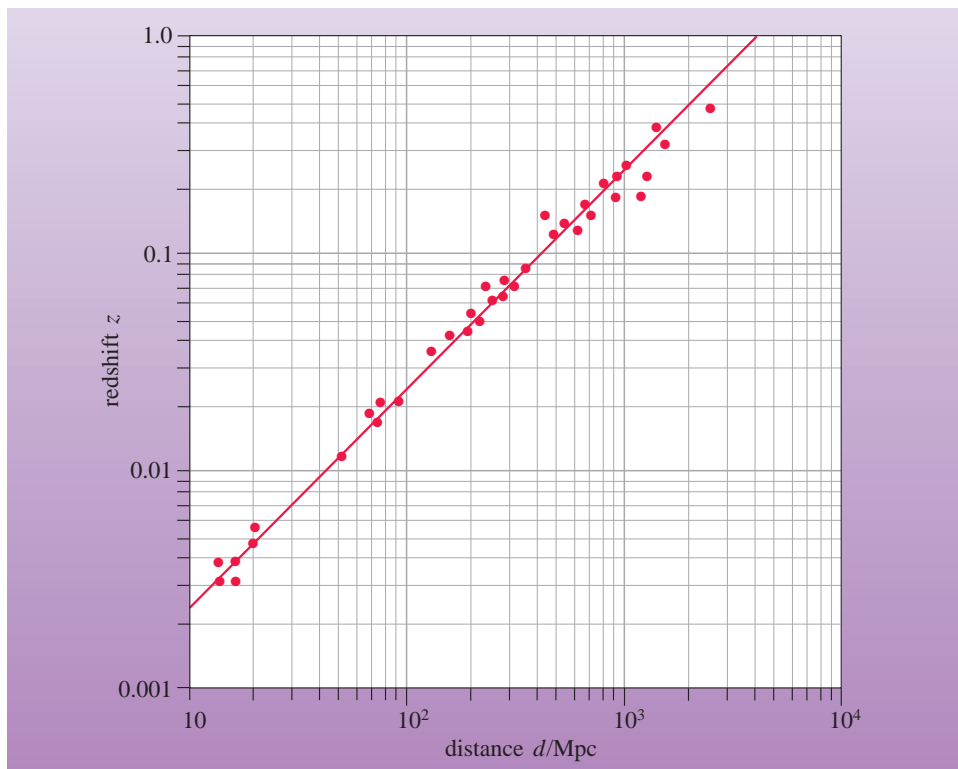


Figure 3.4 A plot of redshift against distance for some galaxies.

The observed redshift is a consequence of the overall expansion of the Universe and, for speeds less than about 10% of the speed of light, is related to the apparent speed of recession of a galaxy by

$$z = \frac{v}{c} \quad (3.3)$$

Essential skill:
Using the Hubble law

Worked Example 3.2

The $H\beta$ line in the spectrum of a distant galaxy is observed to have a wavelength of 5007 \AA instead of its rest wavelength of 4861 \AA . What is the apparent speed of recession of the galaxy and what is its distance?

Solution

The redshift of the galaxy is calculated using Equation 3.1 as

$$z = \frac{5007 - 4861}{4861} = 0.0300$$

Then using Equation 3.3, the apparent speed of recession is $v = cz = 3.00 \times 10^5 \text{ km s}^{-1} \times 0.0300 = 9000 \text{ km s}^{-1}$. Hence, using Equation 3.2 the distance to the galaxy is $d = cz/H_0 = (3.00 \times 10^5 \text{ km s}^{-1} \times 0.0300) / (67.74 \text{ km s}^{-1} \text{ Mpc}^{-1}) = 133 \text{ Mpc}$.

The various methods for distance determination, taken together, form a galactic distance ladder (Figure 3.5). Each rung in this ladder requires a calibration process, the accuracy of which is itself dependent on lower rungs of the ladder. This means that uncertainties tend to increase as the ladder is climbed.

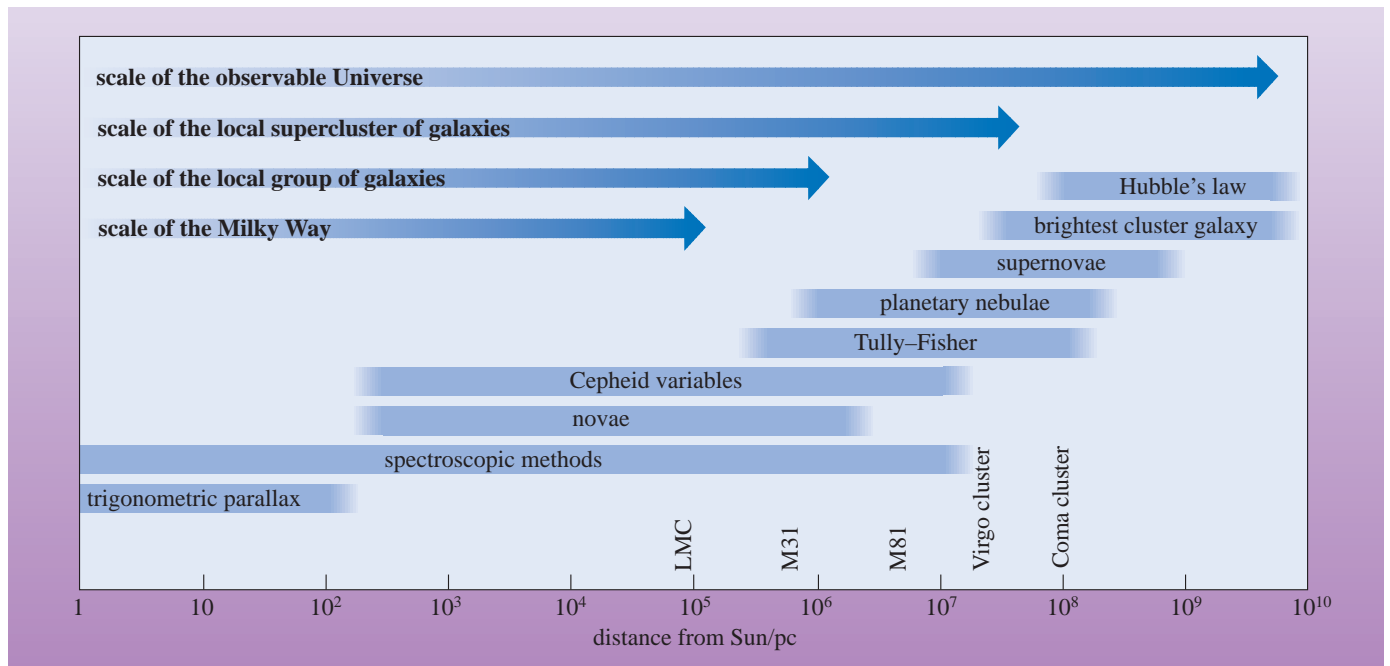


Figure 3.5 Methods of measuring astronomical distances. (Some of the techniques indicated are not discussed here, but their names are included for completeness.)

3.4 Active galaxies

There is an important class of galaxies known as active galaxies that exhibit rather more extreme properties than their quiescent relatives.

3.4.1 The spectra of active galaxies

A normal elliptical galaxy is composed mainly of stars, and has an optical spectrum that looks rather like the optical spectrum of a star, but with fainter absorption lines. Similarly, a normal spiral galaxy has an optical spectrum that is the composite of its stars (which show absorption lines) and its HII regions (which show rather weak emission lines) as shown in Figure 3.6.

An **active galaxy** has an optical spectrum that is the composite of the spectrum of a normal galaxy and powerful additional radiation which is often characterized by broad emission lines in the optical (Figure 3.7). In addition, active galaxies may emit powerfully in other wavebands, and it is typically their radio or X-ray emission that reveals the presence of the active galactic nucleus lying at their centres.

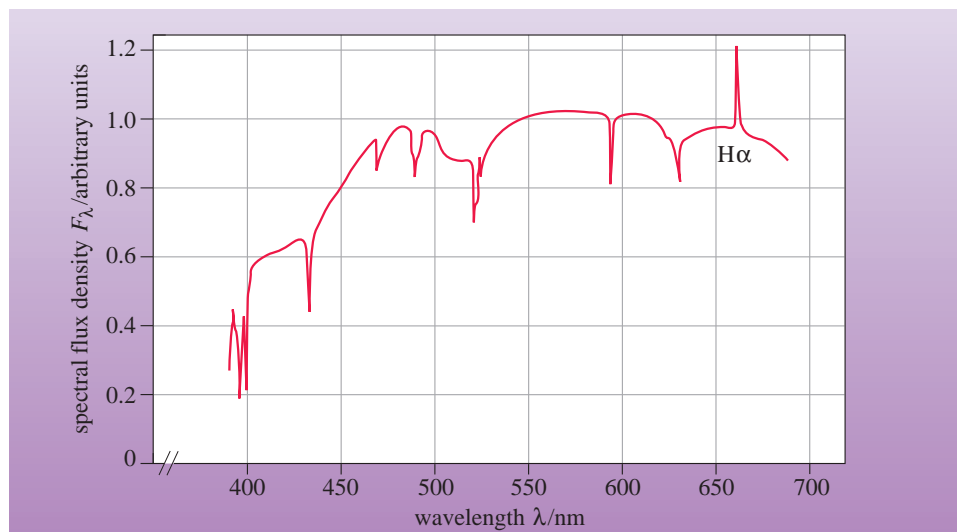


Figure 3.6 The optical spectrum of a normal spiral galaxy shown schematically.

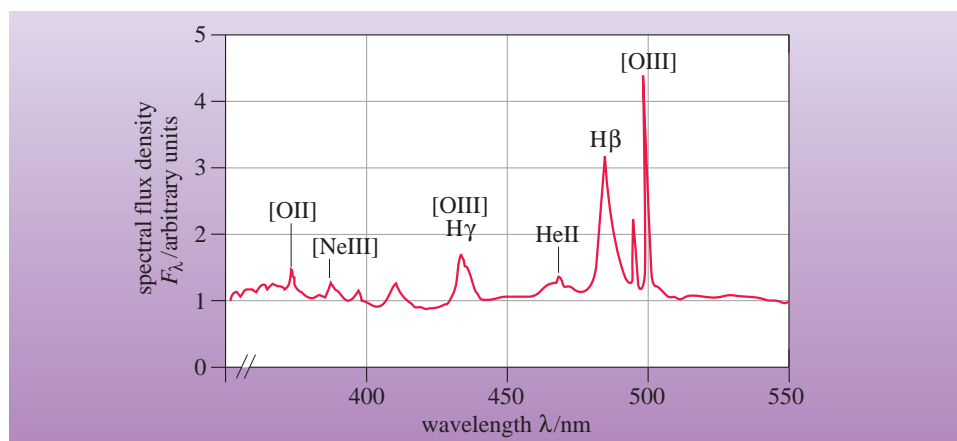


Figure 3.7 The schematic optical spectrum of an active galaxy.

To judge the overall emission from a galaxy it is useful to look at the overall, or broadband, spectrum, that is extending beyond the visible into the other regions of the electromagnetic spectrum. To plot such a spectrum it is necessary to use logarithmic axes, because both the brightness and wavelength vary by many powers of ten. Furthermore, the vertical axis of the spectrum is most commonly plotted as a **spectral flux density**. At any given wavelength, the spectral flux density in terms of wavelength F_λ can be determined by the following procedure:

1. Using an appropriate detector of unit area, pointed directly towards the source, measure the rate at which energy from the source is delivered to the detector by electromagnetic waves with wavelengths in a fixed narrow range $\Delta\lambda$ centred on λ .
2. Divide the measured rate of energy detection by the wavelength range $\Delta\lambda$ to obtain the detected power per unit area per unit wavelength range. This is the value of F_λ at wavelength λ . It is typically measured in the SI units of $\text{W m}^{-2} \mu\text{m}^{-1}$ (or $\text{J s}^{-1} \text{m}^{-2} \mu\text{m}^{-1}$), or in the cgs units of $\text{erg s}^{-1} \text{cm}^{-2} \text{\AA}^{-1}$.

A related spectral flux density, in terms of frequency instead of wavelength, is also used and denoted by F_ν . It is defined as follows:

1. Using an appropriate detector of unit area, pointed directly towards the source, measure the rate at which energy from the source is delivered to the detector by electromagnetic waves with frequencies in a fixed narrow range $\Delta\nu$ centred on ν .
2. Divide the measured rate of energy detection by the frequency range $\Delta\nu$ to obtain the detected power per unit area per unit frequency range. This is the value of F_ν at frequency ν . It is typically measured in the SI units of $\text{W m}^{-2} \text{Hz}^{-1}$ (or $\text{J s}^{-1} \text{m}^{-2} \text{Hz}^{-1}$), or in the cgs units of $\text{erg s}^{-1} \text{cm}^{-2} \text{Hz}^{-1}$.

In fact there is a unit of spectral flux density F_ν commonly used in astrophysics which is rather simpler. The unit is the **jansky** (symbol Jy) named in honour of the radio astronomer Karl Jansky. The conversion is $1 \text{ Jy} = 10^{-26} \text{ W m}^{-2} \text{Hz}^{-1} = 10^{-23} \text{ erg s}^{-1} \text{cm}^{-2} \text{Hz}^{-1}$ or $1 \text{ W m}^{-2} \text{Hz}^{-1} = 10^{26} \text{ Jy}$.

The schematic broadband spectrum of a normal galaxy, plotted as F_λ versus λ , is shown in Figure 3.8a. The peak occurs in the visible part of the spectrum and the spectrum falls away either side to the X-ray and radio wave regions. However, such a graph does not present the full picture. A better way to present the information is as follows.

The power F received per unit area of a telescope, over a wavelength range of width $\delta\lambda$ is given by

$$F = F_\lambda \delta\lambda \quad (3.4)$$

Suppose we wish to compare the value of F at X-ray wavelengths with the value of F at radio wavelengths, keeping $\delta\lambda$ fixed. In this case $\delta\lambda$ would be a far smaller fraction of the radio wavelength range than of the X-ray wavelength range, and therefore a plot of F versus λ would under-represent the radio range. We can compensate for this by multiplying $\delta\lambda$ by λ , to boost the longer wavelength ranges. Thus in place of Equation 3.4, we have

$$\lambda F = F_\lambda (\lambda \delta\lambda) = \lambda F_\lambda \times \delta\lambda \quad (3.5)$$

The product **lambda eff lambda**, λF_λ , is thus a useful quantity when we are comparing widely separated parts of a broad spectrum. Such a spectrum of a normal spiral galaxy is shown in Figure 3.8b, and is referred to as a **spectral energy distribution**. Now, the highest points of λF_λ will indicate the wavelength regions of maximum power received from the source. You will also see broadband spectra plotted as νF_ν and referred to as **nu eff nu** spectra, which is simply the frequency equivalent of the above representation. From the way the quantities are defined, it is always the case that

$$\lambda F_\lambda = \nu F_\nu \quad (3.6)$$

and graphs of both are referred to as spectral energy distributions.

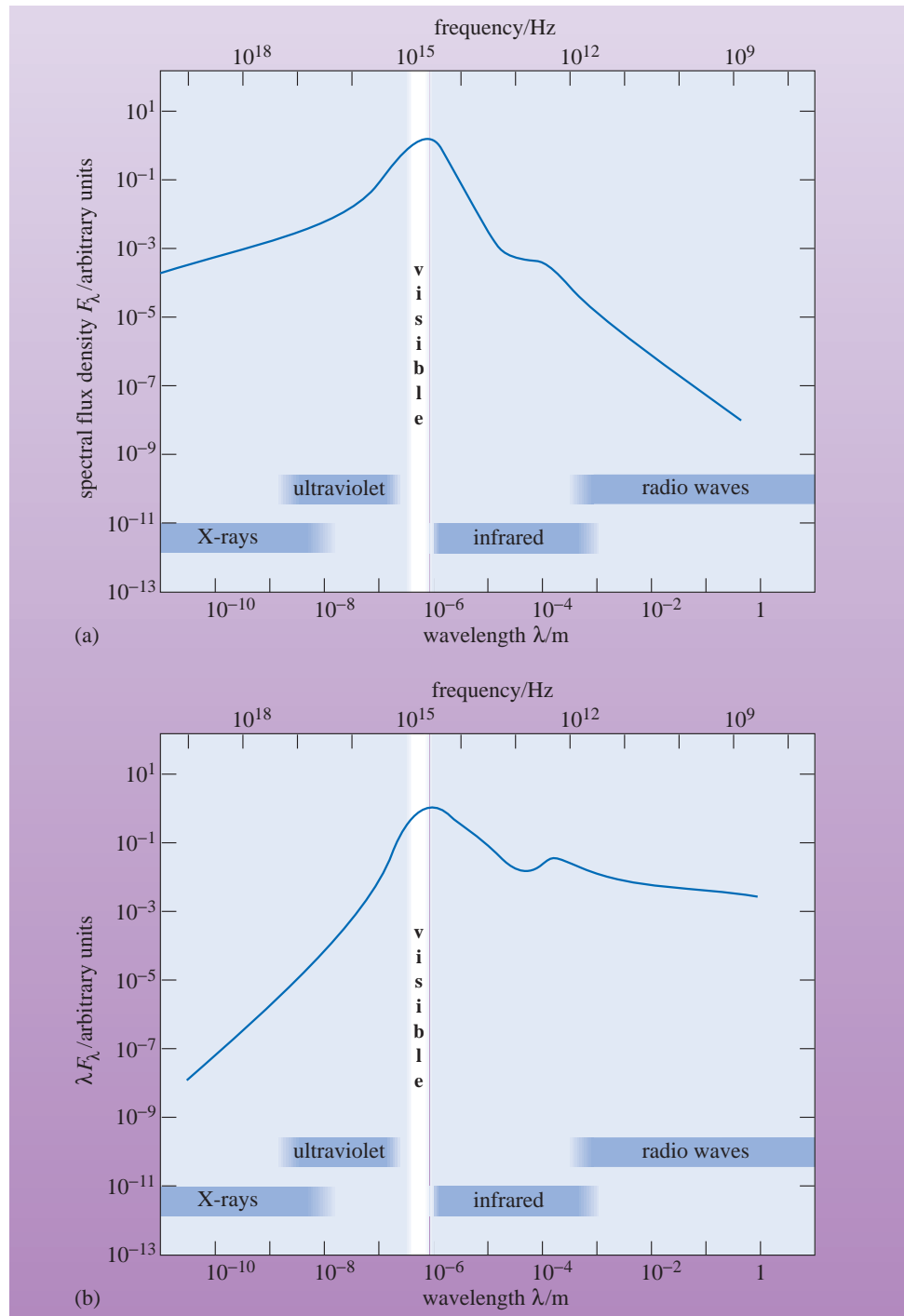


Figure 3.8 The schematic broadband spectrum of a normal spiral galaxy. (a) In terms of an F_λ versus λ plot, and (b) in terms of a λF_λ versus λ plot, i.e. a spectral energy distribution.

The broadband spectra of normal galaxies peak in the optical (Figure 3.8b); and the broadband spectra of active galaxies generally reach a maximum in the X-ray or ultraviolet (Figure 3.9), but occasionally in other parts of the spectrum. The term **spectral excess** is used to refer loosely to the prominence of certain other wavelength regions in the broadband spectra of active galaxies.

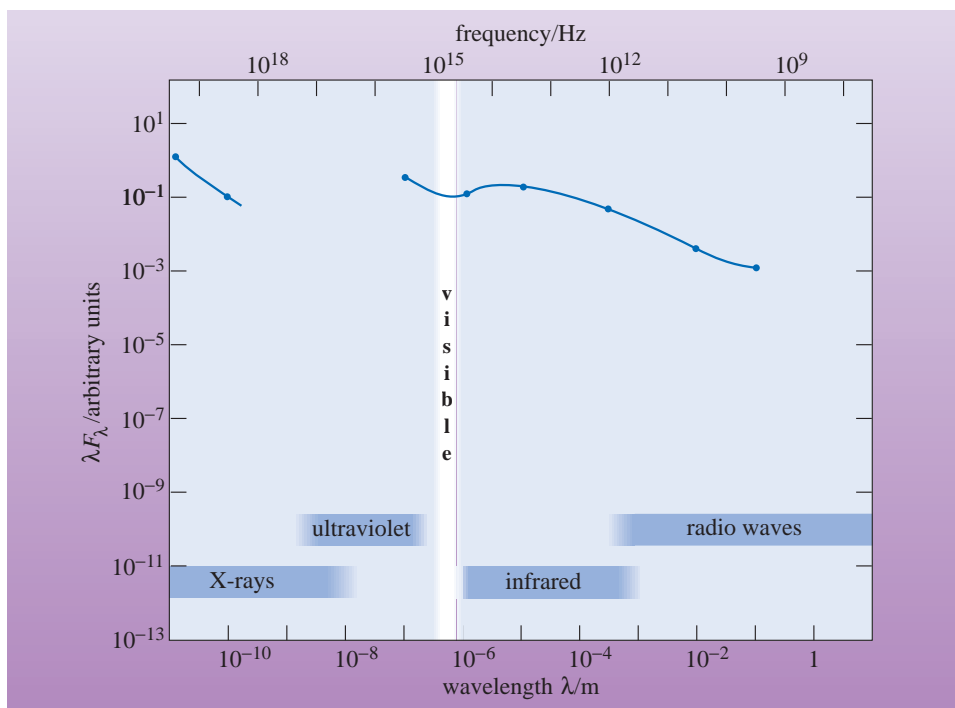


Figure 3.9 The spectral energy distribution of an active galaxy, the quasar 3C 273.

Conversions between the various ways of expressing F_λ and F_ν are not difficult, they are just rather tedious as the example below demonstrates!

Worked Example 3.3

If the spectral flux density of a galaxy is $F_\lambda = 10^{-14} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ \AA}^{-1}$ at a wavelength of $\lambda = 4500 \text{ \AA}$, what is its spectral flux density F_ν in units of mJy (millijansky)?

Solution

The key thing here is to keep track of units throughout the calculation.

The relationship between wavelength and frequency for electromagnetic radiation is $c = \lambda\nu$, so in this case, the frequency is

$$\nu = c/\lambda = (3.00 \times 10^8 \text{ m s}^{-1})/(4500 \times 10^{-10} \text{ m}) = 6.67 \times 10^{14} \text{ Hz}$$

Now, using Equation 3.6,

$$F_\nu = \lambda F_\lambda / \nu = (4500 \text{ \AA}) \times (10^{-14} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ \AA}^{-1}) / (6.67 \times 10^{14} \text{ Hz})$$

$$\text{so } F_\nu = 6.75 \times 10^{-26} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ Hz}^{-1}$$

Now, to convert this into jansky, we need the spectral flux density in the units of $\text{W m}^{-2} \text{ Hz}^{-1}$. First, since $1 \text{ erg s}^{-1} = 10^{-7} \text{ J s}^{-1} = 10^{-7} \text{ W}$, the spectral flux density may be written

$$F_\nu = 6.75 \times 10^{-26} \times 10^{-7} \text{ W cm}^{-2} \text{ Hz}^{-1} = 6.75 \times 10^{-33} \text{ W cm}^{-2} \text{ Hz}^{-1}$$

and since $1 \text{ cm}^{-2} = 10^4 \text{ m}^{-2}$, the spectral flux density may also be written

$$F_\nu = 6.75 \times 10^{-33} \times 10^4 \text{ W m}^{-2} \text{ Hz}^{-1} = 6.75 \times 10^{-29} \text{ W m}^{-2} \text{ Hz}^{-1}$$

Finally, since $1 \text{ W m}^{-2} \text{ Hz}^{-1} = 10^{26} \text{ Jy}$, the answer is

$$F_\nu = 6.75 \times 10^{-29} \times 10^{26} \text{ Jy} = 6.75 \times 10^{-3} \text{ Jy} = 6.75 \text{ mJy}$$

Essential skill:

Converting between different measures of flux density

3.4.2 Types of active galaxy

It is clear that all active galaxies have a compact energetic nucleus – an AGN (Active Galactic Nucleus) – and that the broadband spectrum of an active galaxy gives one indication of the underlying energetic processes which power the emission and which set such galaxies apart from those without an active nucleus. However, further clues to the nature of active galaxies are obtained from resolved images, particularly images obtained in the radio waveband. It is here that extended structures including jets and lobes may be seen, emanating from the galaxy core. Amongst the various subclasses of active galaxy are:

- **Seyfert galaxies** are spiral galaxies with bright, point-like nuclei which vary in brightness, although in general they are relatively low luminosity AGN. They show excesses at far-infrared and other wavelengths, and have strong emission lines. The spectra of Seyfert 1 galaxies show narrow emission lines (widths of a few hundred km s^{-1}) as well as broad lines (widths of a few thousand km s^{-1}); whereas the spectra of Seyfert 2 galaxies contain only the narrow emission lines.
- **Quasars** are the most luminous AGN, and are very variable in brightness. About 10% of quasars are strong radio sources, thought to be powered by jets of material moving at speeds close to the speed of light.
- **Radio galaxies** are distinguished by having giant radio lobes fed by one or two jets. They are usually identified with giant elliptical galaxies or with quasars
- **Blazars** exhibit a continuous spectrum across a wide range of wavelengths and emission lines, when present, are broad and weak. They are variable on very rapid timescales.

The central engine of a typical active galaxy is believed to contain a supermassive black hole of mass $\sim 10^8 M_{\odot}$ contained within a region that is less than 2 AU ($\sim 3 \times 10^{11}$ m) in radius. The small size is in part deduced from the fact that an object which fluctuates in brightness on a timescale Δt can have a radius no greater than

$$R_{\text{max}} \sim c \Delta t \quad (3.7)$$

Infalling material forms an accretion disc around the black hole converting gravitational energy into thermal energy and radiation. Jets are emitted, in some systems, perpendicular to the accretion disc. A typical AGN luminosity of 10^{38} W can be produced by an accretion rate of around $0.2 M_{\odot}$ per year. The maximum luminosity of an accreting black hole is given by the Eddington limit, at which the gravitational force on the infalling material is balanced by the radiation pressure of the emitted radiation.

Exercise 3.1 The AGN shown in Figure 3.10 is seen to double its X-ray brightness during the observation shown here. What is the maximum size of the region that could produce this radiation?



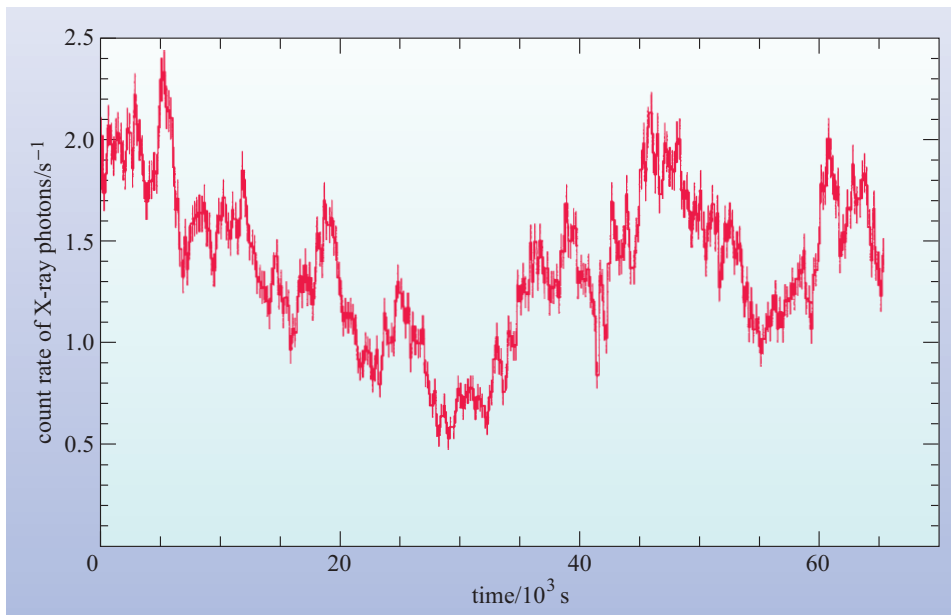


Figure 3.10 An example of X-ray variability shown by the Seyfert galaxy MCG-6-30-15 during an observation made by the Chandra X-ray observatory. The fastest fluctuations are spurious noise, but the variability over a few thousand seconds is a property of the AGN. (Lee et al. 2002)

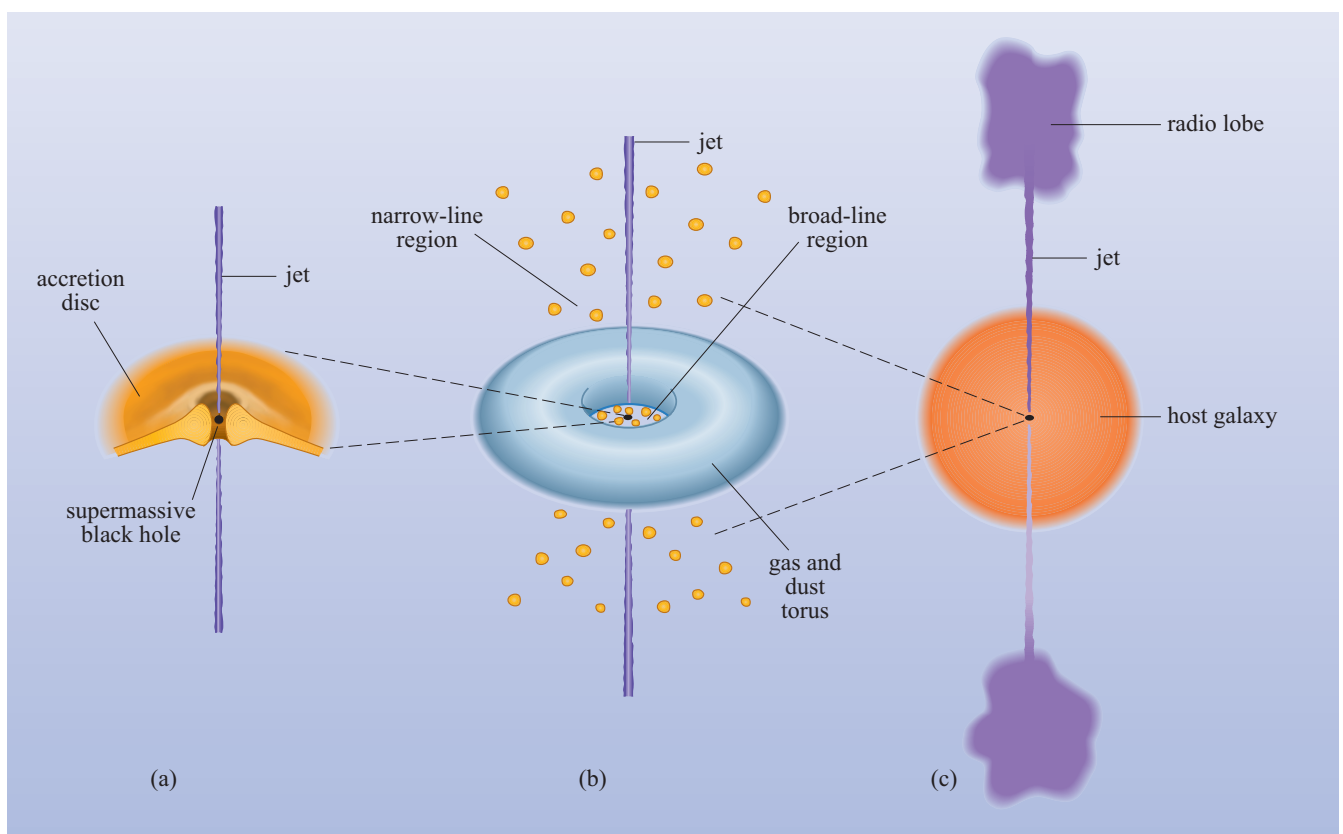


Figure 3.11 A generic model for an active galaxy. (a) A supermassive black hole is surrounded by an accretion disc; jets emerge perpendicular to it. (b) An obscuring torus of gas and dust encloses the broad line region (a few light-days across) with the narrow line region (a few hundreds of parsecs across) lying further out. (c) The entire AGN appears as a bright nucleus in an otherwise normal galaxy, whilst jets (hundreds of kiloparsec in length) terminate in radio lobes.

The standard model of an AGN (Figure 3.11) consists of a supermassive black hole (the engine) accreting from a hot accretion disc. The disc is the source of the ultraviolet and X-ray emission from the AGN. Surrounding this is a broad-line region, contained within a torus of infrared emitting dust, and a narrow-line region lies further out. The broad-line region contains clouds that are moving with speeds of thousands of km s^{-1} , so giving rise to broad emission lines as a result of Doppler broadening. The larger narrow-line region contains clouds that are moving with speeds of a few hundred km s^{-1} and so gives rise to narrower emission lines.

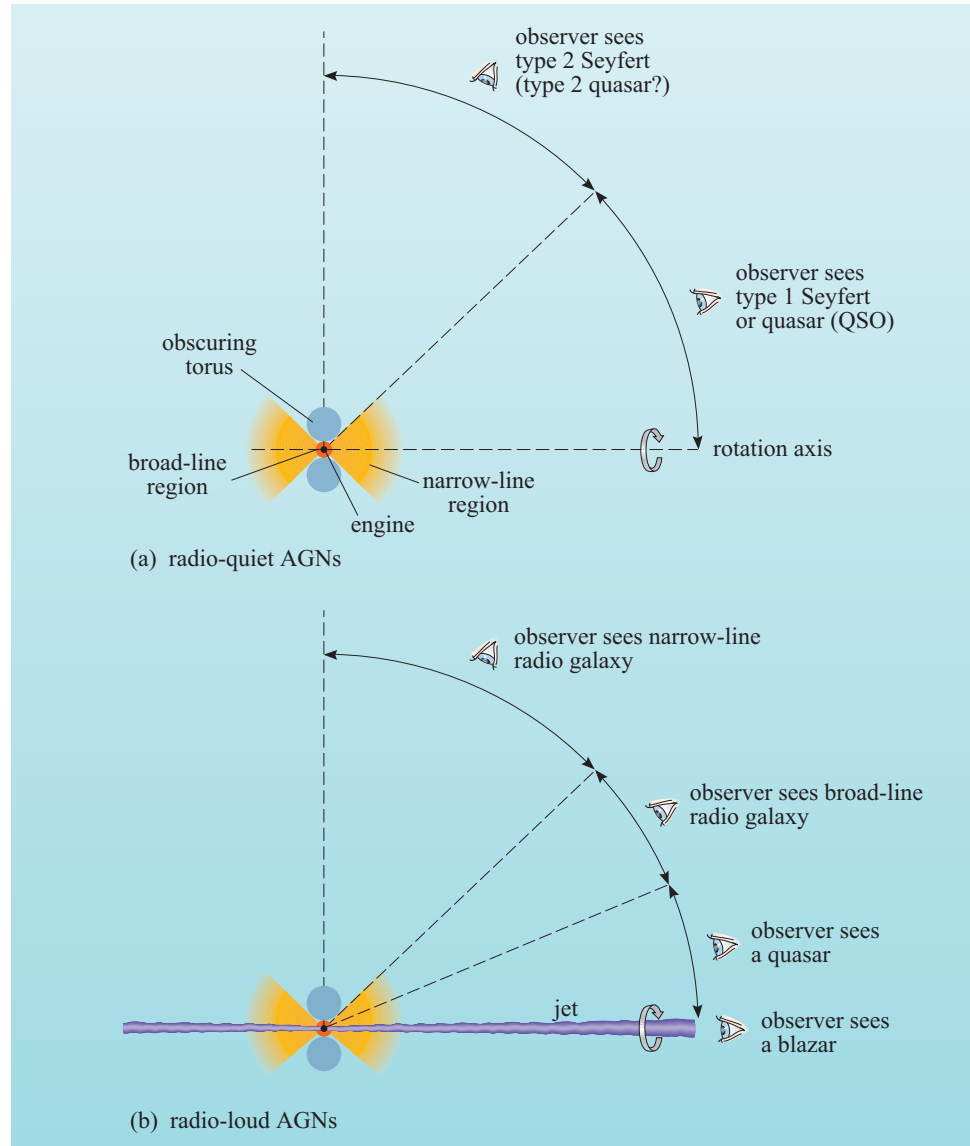


Figure 3.12 Two possible unified models for AGN. Note that these diagrams are rotated by 90° with respect to Figure 3.11 – the plane of the torus is vertical here and the jets emerge horizontally. (a) Radio-quiet AGN. (b) Radio-loud AGN.

Unified models of AGN (Figure 3.12) attempt to explain the range of AGN on the assumption that they differ only in luminosity and the angle at which they are viewed. Amongst the radio-quiet AGN, type 1 Seyfert galaxies and type 2 Seyfert galaxies differ only in the angle at which they are viewed. Radio-quiet quasars (QSOs) are similar to Seyferts but much more powerful.

The radio-loud AGN are those which produce radio jets. The idea here is that the observer sees a radio galaxy, a quasar, or a blazar as the viewing angle moves from side-on to the jets to end-on to the jets. The difference between radio-loud and radio-quiet AGN may lie in how fast their black holes are spinning. The faster-spinning ones may have arisen from mergers of black holes resulting from the collision of their host galaxies.

It is also important to realise that quasars evolve with time. There are very few to be found in the nearby Universe, but many to be seen in the distant Universe where we are effectively looking back in time as we look further away. Consequently, it is clear that quasars were much more common in the past (i.e. when the Universe was young) than they are today. Despite this, it is apparent that dormant black holes reside in the centres of most massive galaxies (including our own Milky Way and the Andromeda galaxy). It has therefore been suggested that quasar activity might trace an early stage in the evolution of all galaxies.

3.5 The spatial distribution of galaxies

Galaxies are gravitationally clustered into **groups** (containing up to about 50 galaxies) and **clusters** (which contain from about 50 to over 1000 galaxies). Our Galaxy belongs to the Local Group of galaxies consisting of about 30 members, but dominated by the Milky Way and Andromeda galaxy (M31), see Figure 3.13.

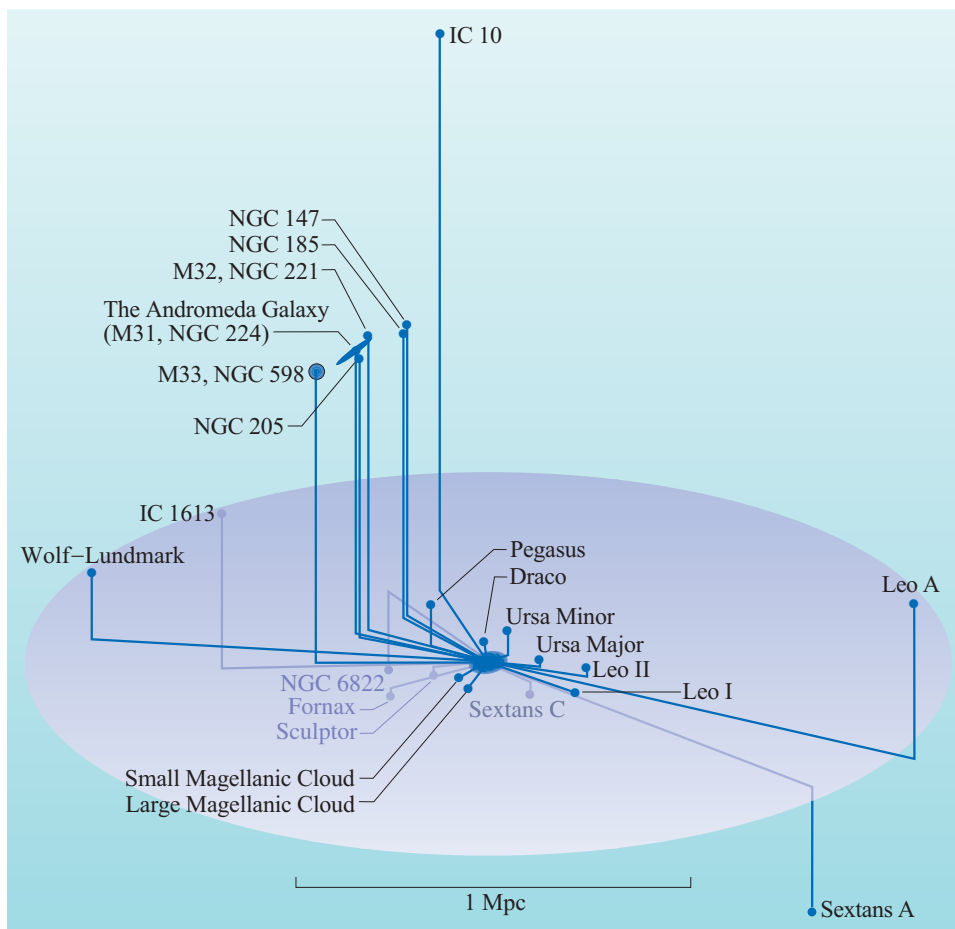


Figure 3.13 The main members of the Local Group of galaxies; the Milky Way is located at the centre.

Medium-scale three-dimensional surveys confirm the existence of **superclusters** – loose collections of clusters which are about 30 – 50 Mpc in extent. Our Local Group is at the outer edge of the Local Supercluster which is centred on, and dominated by, the Virgo cluster.

On the very largest scales (hundreds of megaparsecs), galaxy redshift surveys can be used to reveal the three-dimensional **large scale structure** of the visible Universe. Recent surveys show that superclusters are not themselves organised into ever larger clusters of superclusters. Instead they are distributed in a vast network consisting of high-density regions connected by filaments and sheets wrapped around (relatively) empty voids.

Most clusters of galaxies have a radius of about 2 Mpc (known as the Abell radius, R_A) and the mass of a typical cluster is of the order of 10^{14} to $10^{15} M_\odot$. Cluster masses can be estimated by three main methods – velocity dispersion, X-ray emission and gravitational lensing:

- A cluster is said to be virialized if it is a gravitationally bound system and is in dynamical equilibrium. The mass of such a cluster can be obtained from the dispersion of the line of sight velocities (Δv) using

$$M \approx R_A (\Delta v)^2 / G \quad (3.8)$$

- Rich clusters are strong X-ray emitters due to the presence of hot intracluster gas. X-ray observations can be used to estimate the total mass of the cluster and the mass of X-ray emitting gas.
- A cluster can act as a gravitational lens of a distant galaxy, producing distorted multiple images. This is a means of detecting distant objects and can also be used to estimate the mass of the intervening cluster. The effect is illustrated in Figure 3.14.

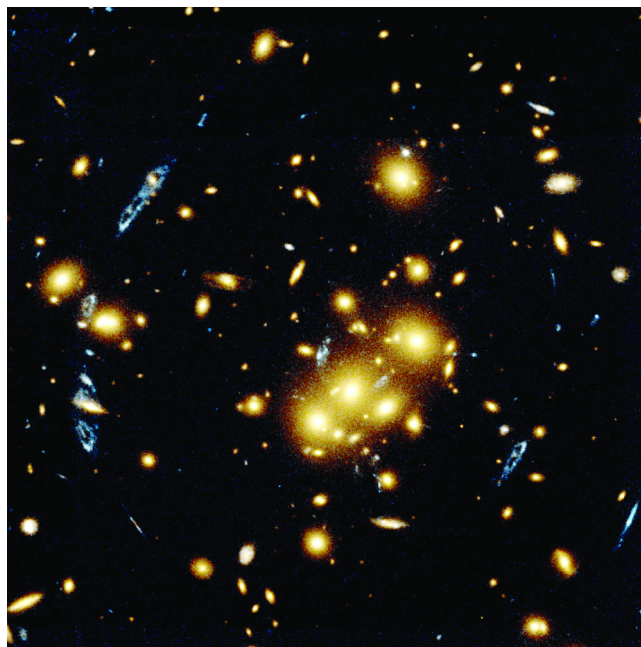


Figure 3.14 Lensing by CL0024+1624. Several distorted images of a distant blue galaxy can be seen encircling the yellower galaxies within the cluster. (W.N. Colley and E. Turner (Princeton University), J.A. Tyson (AT & T Bell Labs, Lucent Technologies) and NASA)

The masses obtained by the three methods mentioned above typically agree within a factor of two or three, but the mass estimates indicate that there is far

more matter in clusters than can be accounted for by the mass of material which is observed to be emitting electromagnetic radiation. This suggests the presence of dark matter. Within clusters, there are indications that the dark matter is distributed more smoothly than the matter that is present in the form of galaxies.

X-ray spectra from clusters show that the hot intracluster medium contains an unexpectedly high metal content. This enrichment of the ICM is the result of supernova explosions in energetic young star-forming galaxies. Temperature maps of clusters indicate that many clusters are not in a state of hydrostatic equilibrium. This can give us information about the formation of clusters, and suggests that clusters may grow from the merging of smaller subclusters.

Some quasar spectra contain multiple absorption lines indicating the presence of gas (mainly neutral hydrogen) at different distances along the line of sight. This is called the **Lyman α forest** and can be used to detect the presence of neutral gas in the intergalactic medium, as illustrated in Figure 3.15. Although the effect is most often studied in terms of the Ly α line, similar structures are also seen from heavier elements, such as magnesium and carbon, in galaxies along the line of sight.

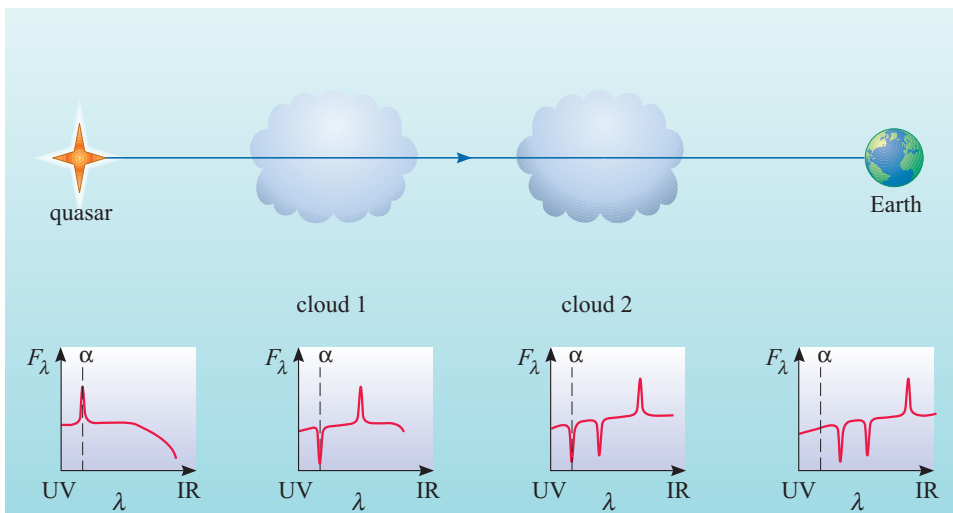


Figure 3.15 Intervening clouds in the line of sight from a quasar. As seen from Earth, the Lyman α emission line from the distant quasar is redshifted into the infrared part of the spectrum. The intervening clouds each have a somewhat smaller redshift and therefore result in a series of Lyman α absorption lines, with the same set of smaller redshifts.

3.6 The structure of the Universe

Clearly, the Universe contains matter in the form of planets, stars and galaxies, not to mention various clouds of gas and dust. However, this (largely) luminous matter comprises only a tiny fraction of the matter content of the Universe. The vast majority of the matter content of the Universe is in the form of **dark matter**. This dark matter is probably comprised of a relatively small amount of **baryonic dark matter** (consisting mainly of hydrogen and helium) and a dominant contribution from **non-baryonic dark matter** (i.e. not composed of the familiar protons, neutrons and electrons). The nature of dark matter is presently unknown. The Universe also contains electromagnetic radiation. Much of it is visible light, but the major part of the energy is contained in the **cosmic microwave background (CMB)** radiation.

The Universe is observed to be very uniform. That is to say, all regions that are sufficiently large to be representative have the same average density and pressure,

wherever they are located. This claim is consistent with the observed distributions of matter and radiation.

As noted earlier, distant galaxies all exhibit redshifts, which increase with distance away from the Earth. The interpretation of this is that the Universe is expanding and this is described by the Hubble law, $z = (H_0/c)d$, where the Hubble constant H_0 provides a measure of the rate of cosmic expansion at the present time.

According to Einstein’s theory of general relativity the geometric properties of spacetime are related to the distribution of energy and momentum within that spacetime. The precise relationship is described by the field equations of **general relativity**, which provide the basis for Einstein’s theory of gravity and for relativistic cosmology. The geometric properties of spacetime include **curvature**, which can be quantified by a parameter k , as illustrated in Figure 3.16. In a curved space, geometric results can take on unfamiliar forms. The interior angles of a triangle may have a sum that is different from 180° and pairs of straight lines that are initially parallel may converge or diverge. The value of k determines whether space is finite ($k = +1$) or infinite ($k = 0$ or -1).

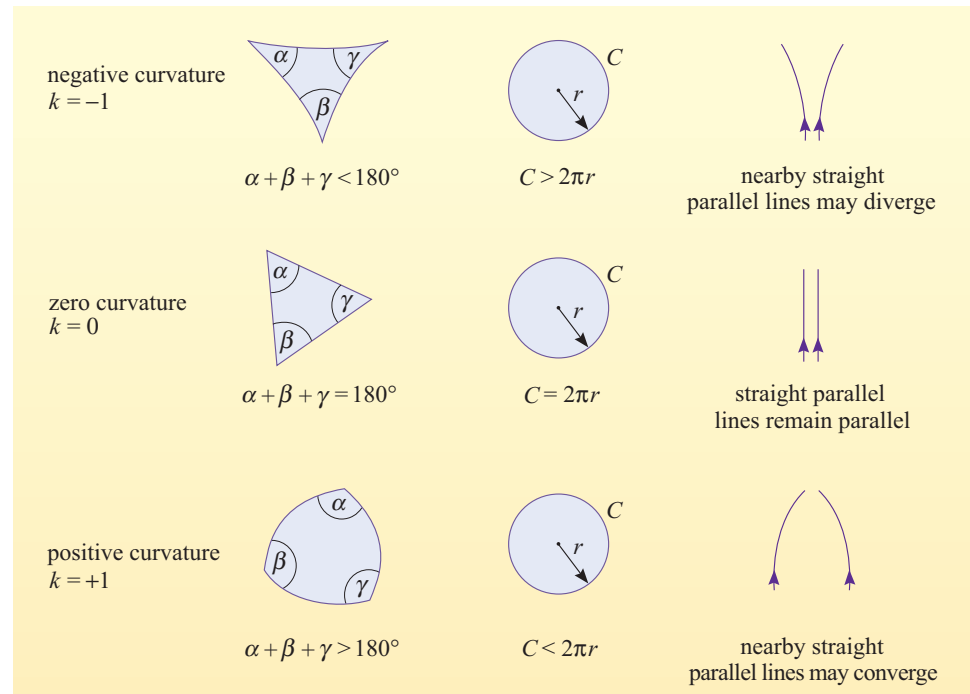


Figure 3.16 The effect of the curvature parameter k in determining the large-scale geometry of a cosmological model.

The geometric properties of any particular spacetime can be summarized by writing down an appropriate four-dimensional generalization of Pythagoras’s theorem. In the case of a static (i.e. non-expanding), flat (i.e. zero curvature) spacetime, this takes the form

$$(ds)^2 = (dx)^2 + (dy)^2 + (dz)^2 - c^2(dt)^2 \tag{3.9}$$

The distribution of energy and momentum throughout spacetime is believed to be uniform on the large scale. This assertion is given precise form by the **cosmological principle** according to which, on sufficiently large size scales, the

Universe is **homogeneous** and **isotropic**. Simple cosmological models that are consistent with this principle assume that a gas uniformly fills the Universe. Describing the state of this gas involves specifying its density and pressure, $\rho(t)$ and $p(t)$, both of which are expected to change with time due to the expansion or contraction of the Universe.

In applying general relativity to cosmology, Einstein introduced a cosmological constant Λ . Thanks to this he was able to formulate a relativistic cosmological model that is neither expanding nor contracting, and in which space has a uniform positive curvature. Later, the work of Friedmann, Robertson and Walker resulted in the specification of the class of cosmological models that are consistent with general relativity and with the cosmological principle. These models involve a curvature parameter k that characterizes the geometry of space, and a **scale factor** $R(t)$ that describes the expansion or contraction of space. The Friedmann equation which describes the way the scale factor varies with time may be written

$$\dot{R}^2(t) = \frac{8\pi G R^2(t)}{3} \left(\rho(t) + \frac{\Lambda c^2}{8\pi G} \right) - kc^2 \tag{3.10}$$

The full range of FRW models, obtained by solving this equation, includes cases that are closed, critical, open and accelerating. Some examples of FRW models are shown in Figure 3.17.

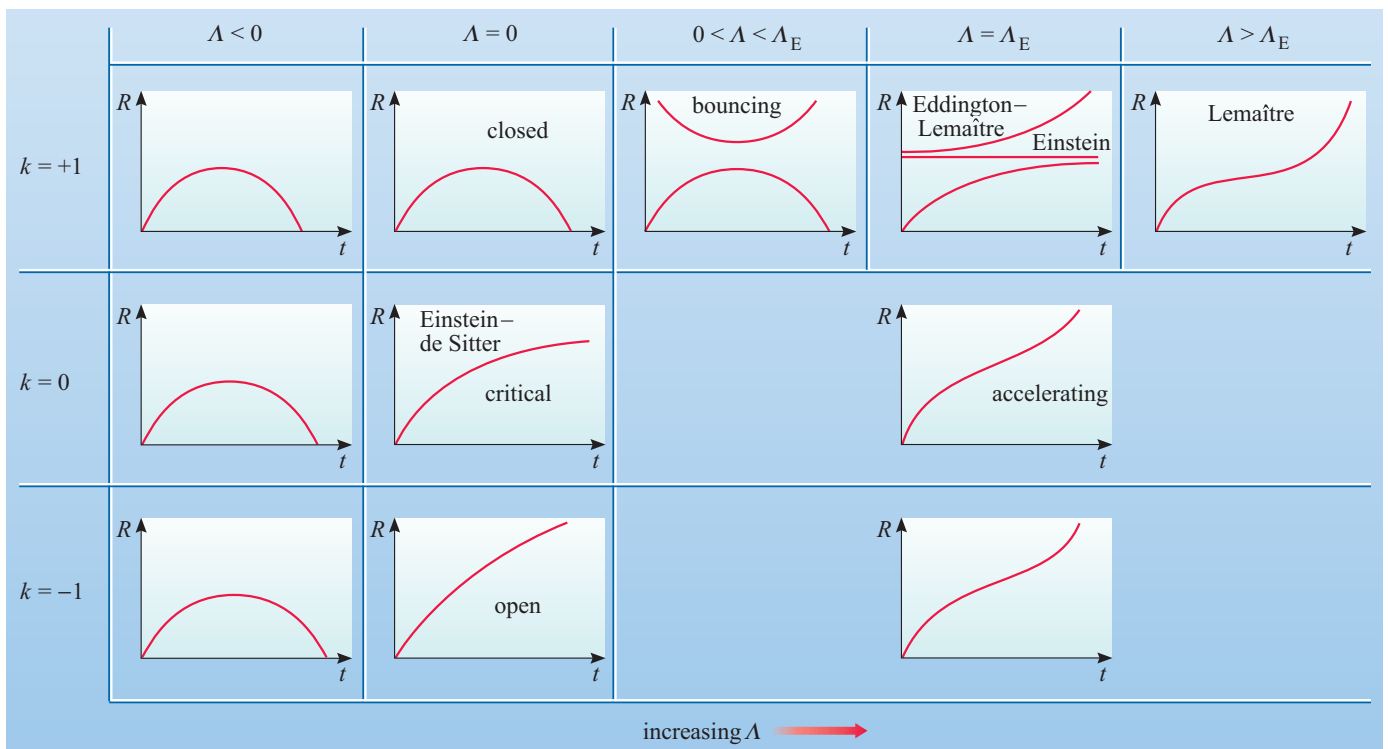


Figure 3.17 The Friedmann–Robertson–Walker models, classified according to the values of Λ and k . In each case the model is represented by a small graph of R against t , which encapsulates the history of spatial expansion and/or contraction implied by the model. Λ_E represents the value of the cosmological constant in the Einstein model.

Models which allow the scale factor to be zero at some time in the past are consistent with the idea that the Universe began with a **big bang**, in which spacetime was created. Models which predict a scale factor of zero at some time in the future imply that the Universe will end with a **big crunch**. The **critical universe** has $k = 0$ and $\Lambda = 0$; it is sometimes known as the Einstein-de Sitter model, and corresponds to a universe which is spatially flat, and has no cosmological constant. Such a universe expands forever, but the expansion always decelerates. Amongst the other models with $\Lambda = 0$, that with a positive curvature is known as a **closed universe** model and that with a negative curvature is known as an **open universe** model.

Exercise 3.2 In the context of the FRW models shown in Figure 3.17, which values or ranges of the parameters k and Λ correspond to universes with the following characteristics?

- The universe is neither homogeneous nor isotropic.
- There is no possibility of a big bang.
- A big bang is possible, but there is at least one other possibility.
- The particular point in space where the big bang happened can still be determined long after the event.
- At any time, the large-scale geometrical properties of space are identical to those of a three-dimensional space with a flat geometry.
- Space has a finite volume, and straight lines that are initially parallel may eventually meet.
- There is a big bang, but the volume of space is infinite from the earliest times.

The FRW models provide a natural interpretation of the redshifts of distant galaxies as cosmological redshifts caused by the stretching of light waves while they move through an expanding space. This allows us to rewrite the expression for redshift (Equation 3.1) as follows. From

$$z = \frac{\Delta\lambda}{\lambda} = \frac{\lambda_{\text{obs}} - \lambda_{\text{em}}}{\lambda_{\text{em}}}$$

we have

$$z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{em}}} - 1$$

The ratio of the two wavelengths must be equal to the ratio of the two scale factors corresponding to the time when the light is observed and the time when it is emitted, hence

$$z = \frac{R(t_{\text{obs}})}{R(t_{\text{em}})} - 1 \quad (3.11)$$

The Hubble parameter $H(t)$ provides a measure of the rate of expansion of space in any FRW model. It is defined by

$$H(t) = \dot{R}/R \quad (3.12)$$

where \dot{R} denotes the rate of change of R with time. Observations of distant

galaxies are predicted to show that, to a first approximation, $d = cz/H_0$, where H_0 represents the value of the Hubble parameter at the time of observation.

The density parameters, Ω_m and Ω_Λ , provide a useful means of representing the cosmic matter density and the density associated with the cosmological constant at any time. The parameters are defined by

$$\Omega_m = \rho/\rho_{\text{crit}} \tag{3.13}$$

and

$$\Omega_\Lambda = \rho_\Lambda/\rho_{\text{crit}} \tag{3.14}$$

respectively, where ρ is the matter density at the time of observation,

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi G} \tag{3.15}$$

is a ‘density’ associated with the cosmological constant, and

$$\rho_{\text{crit}} = \frac{3H^2(t)}{8\pi G} \tag{3.16}$$

is the density that the critical universe would have at the time of observation. The quantity $\rho_\Lambda c^2$ can be thought of as the density of dark energy. In a universe with a flat space (i.e. $k = 0$), the Friedmann equation implies that

$$\Omega_m + \Omega_\Lambda = 1 \tag{3.17}$$

at all times. Various possible behaviours for the Universe depending on the values of Ω_m and Ω_Λ are shown in Figure 3.18.

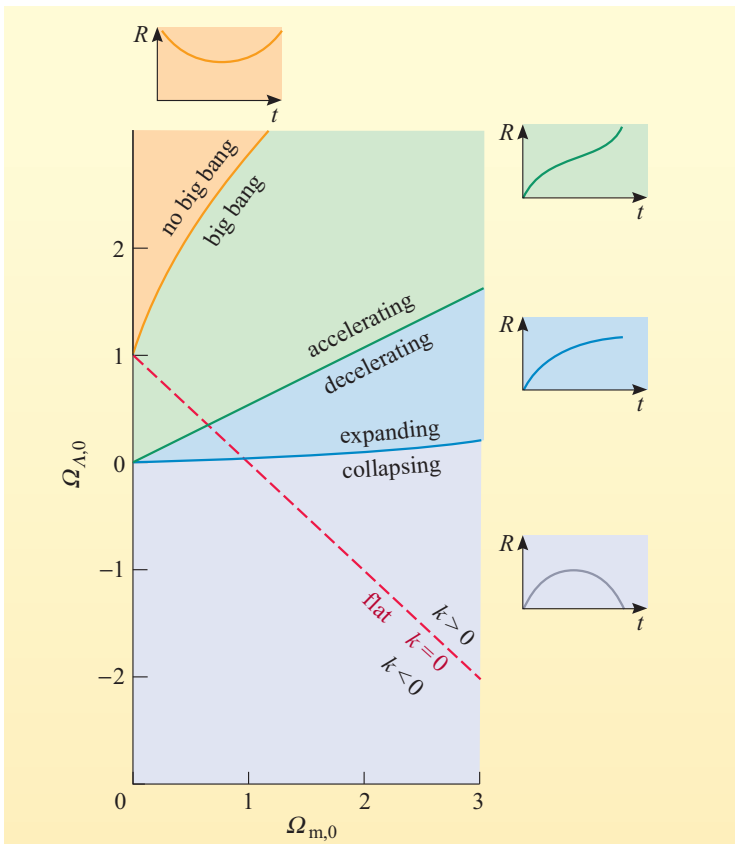


Figure 3.18 A plot of $\Omega_{\Lambda,0}$ (the density associated with the cosmological constant today) versus $\Omega_{m,0}$ (the cosmic matter density today). The values of these two quantities determine important characteristics of an FRW cosmological model, such as the sign of the curvature parameter k , whether the Universe will expand forever or eventually collapse, whether the expansion will accelerate or decelerate, and whether or not there was a big bang.

The age of the Universe t_0 may be conveniently expressed in terms of the Hubble time $1/H_0$ in any FRW model. In the case of the critical model, $t_0 = 2/3H_0$. In other models t_0 may be a different fraction of the Hubble time, depending on the values of Ω_m and Ω_Λ . Increasing the value of Ω_Λ increases the age of the universe for a given value of the Hubble constant.

3.7 The evolution of the Universe

The Universe began with a hot big bang, at which instant spacetime was created. The relic of this event is the **cosmic background radiation** which pervades the entire Universe. It is observed to have a near perfect black-body spectrum and its current temperature is about 2.73 K (see Figure 3.19). In fact, the temperature T varies with the scale factor $R(t)$ according to

$$T \propto 1/R(t) \quad (3.18)$$

At times when the scale factor of the Universe was much smaller than at present, the temperature of the cosmic background radiation would therefore have been much higher.

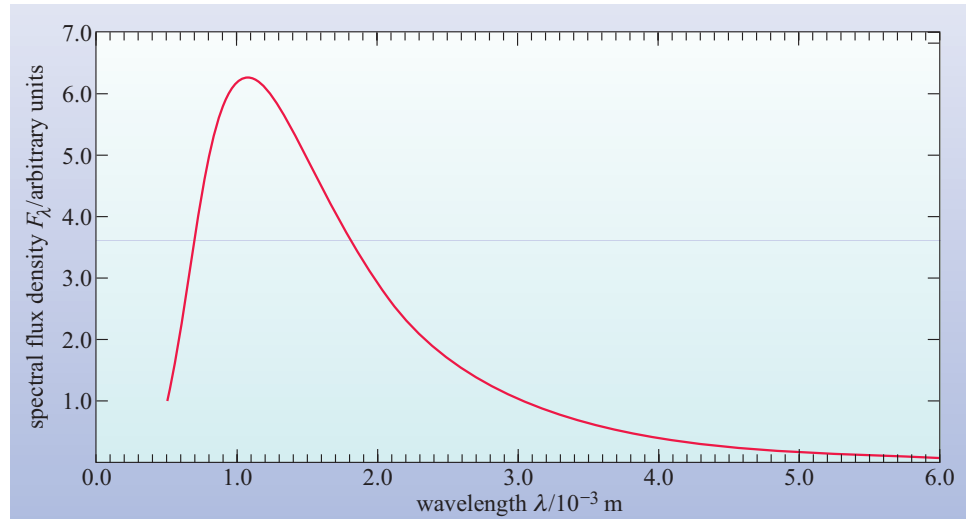


Figure 3.19 The spectrum of the cosmic microwave background. The peak of the spectrum, at a wavelength of around 1 mm, corresponds to a black-body temperature of about 3 K.

At early times the dominant contribution to the energy density of the Universe was that due to radiation. At such times, the temperature is related to time by

$$T/K \approx 1.5 \times 10^{10} \times (t/s)^{-1/2} \quad (3.19)$$

Later on in the evolution of the Universe, the dominant contribution to the energy density was provided by matter. Still more recently, at the present time the dominant energy density is believed to be due to dark energy, as illustrated in Figure 3.20.

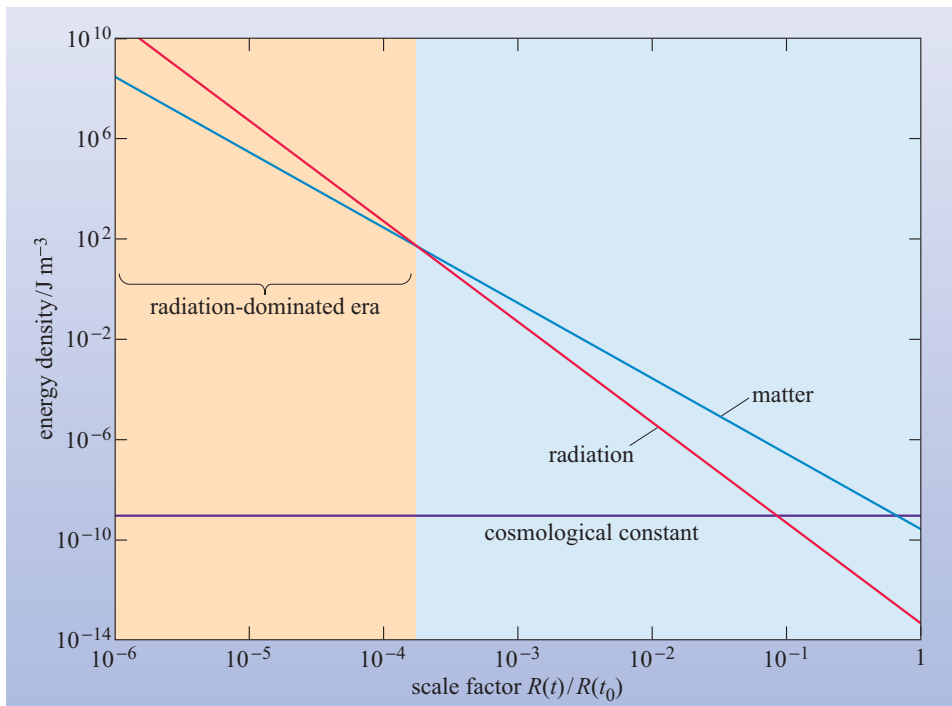


Figure 3.20 The energy densities of matter (blue line) and radiation (red line) as a function of scale factor. At a time when $R(t)/R(t_0) \approx 10^{-4}$ the energy densities of matter and radiation were equal. Prior to this time, the energy density of radiation exceeded that of matter – during this era the dynamical evolution of the Universe was determined by its radiation content. After this time, the energy density of matter was greater, so it was the matter in the Universe that controlled its dynamical evolution. The behaviour of the energy density due to the cosmological constant is also shown (purple line) – this does not vary with redshift and is exceeded by the energy densities of matter and radiation at early times.

Current physical theory breaks down in describing events that took place at, or before, the **Planck time** ($t \sim 10^{-43}$ s after the big bang). It is speculated that major physical effects could have arisen when grand unification ended (at $t \sim 10^{-36}$ s after the big bang). At this time, the strong and electroweak interactions became distinct. One such effect may be the process of **inflation**, which resulted in the scale factor increasing very rapidly for a short period of time.

At early times, the content of the Universe would have been all types of quark and lepton and their antiparticles. There were also particles present that mediate the fundamental interactions (such as the photon), as well as dark-matter particles. There was also a slight excess of matter over antimatter, although the cause for this imbalance is not currently known. At $t \sim 10^{-5}$ s after the big bang, free quarks became bound into hadrons. Most of these hadrons either decayed or annihilated with their antiparticles, leaving only protons and neutrons. For every 10^{10} or so annihilation events that occurred, there would have been one proton or neutron left over.

At $t \sim 0.7$ s after the big bang, neutrinos had their last significant interaction with other particles (apart from the effects of gravity). Shortly after this, when the Universe was about 10 s old, electron–positron pairs annihilated, leaving only a residual number of electrons whose summed electric charges exactly balance the

charge on the protons.

In the first few hundred seconds of the history of the Universe, the physical conditions were such that nuclear fusion reactions could occur. Such reactions led to the formation of deuterium, helium and lithium. The first step in the production of helium is the formation of deuterium. This nuclide is unstable to photodisintegration at temperatures above 10^9 K. The formation of helium did not start until $t \approx 225$ s. During this time, some neutrons decayed to protons, and this had an effect on the mass fraction of helium that was produced by **primordial nucleosynthesis**. The mass fraction of helium that is predicted by primordial nucleosynthesis is about 24%. This is in good agreement with measurements of the helium abundance in interstellar gas and stars, and provides very strong evidence to support the hot big bang model. (NB. Although stars process hydrogen into helium and expel this into the interstellar medium via planetary nebulae and supernovae, there is far more helium in the Universe than can be produced by this route.)

The cosmic microwave background that is observed at the present time appears to originate from a particular last-scattering surface. The scattering of background radiation photons stopped when the number density of free electrons became very low, and this occurred because of the recombination of electrons and nuclei to form neutral atoms. The age of the Universe at this point was about $t = 300,000$ years after the big bang and the temperature of the Universe when this occurred was around 3000 K.

Exercise 3.3 Calculate the redshift at which the last scattering occurred. (*Hint:* Start by using Equation 3.18 and Equation 3.11 to determine the change in scale factor.)

Exercise 3.4 Suppose we received a message from (hypothetical) astronomers in a galaxy that has a current redshift of $z = 2.5$. What would they say they found as the temperature of the cosmic microwave background at the time of their transmission?



The observed high degree of uniformity of the cosmic microwave background leads to the horizon problem – which is that regions of the last-scattering surface that are more than about 2° apart could not have come into thermal equilibrium by the time that last scattering occurred. The cosmic microwave background shows intrinsic anisotropies in temperature at a level of a few parts in 10^5 . These anisotropies result from density variations in the early Universe.

The formation of structure in the Universe would have proceeded by gravitational collapse from density fluctuations in the early Universe. Prior to recombination, the high degree of scattering between photons and electrons prevented density fluctuations in baryonic matter from growing substantially. If all matter was baryonic in form, then the level of fluctuation that is observed on the last scattering surface is too small to explain the structure that we observe at the present time. The observed level of structure in the present-day Universe can be explained, however, if density fluctuations in non-baryonic matter had begun to grow prior to recombination, and baryons were subsequently drawn into those collapsing clouds of dark matter.

3.8 Observational cosmology

So far, we have mostly described the Universe from the viewpoint of theory about its structure and evolution. In this section we present some information about the Universe from the viewpoint of observations or measurement.

As you know, the Hubble constant H_0 measures the current rate of expansion of the Universe. H_0 is traditionally determined by means of the Hubble diagram: a plot of redshift against distance for distant galaxies. When making such a plot, the redshift and the distance must be determined independently and the Hubble constant is obtained from the gradient of the plotted line. Other methods of determining H_0 include those based on gravitational lensing (via time delays between fluctuations in different images of the same lensed galaxy) and those based on observations of anisotropies in the cosmic microwave background radiation (CMB).

As noted earlier, the current values of the **density parameters**, $\Omega_{\Lambda,0}$, $\Omega_{m,0}$ and $\Omega_{b,0}$, measure the densities associated with the cosmological constant (dark energy), matter of all kinds, and baryonic matter, relative to the critical density

$$\rho_{\text{crit}} = \frac{3H_0^2}{8\pi G} \quad (\text{Eqn 3.16})$$

Results based on observations of anisotropies in the CMB strongly favour $\Omega_{\Lambda,0} + \Omega_{m,0} = 1$ and $k = 0$. When these are combined with the results of observations of Type Ia supernovae, the values of $\Omega_{\Lambda,0}$ and $\Omega_{m,0}$ are determined to be $\Omega_{\Lambda,0} = 0.691 \pm 0.006$ and $\Omega_{m,0} = 0.309 \pm 0.006$, as illustrated in Figure 3.21.

The current value of the density of baryonic matter can be determined in a number of ways. This quantity is constrained by primordial nucleosynthesis calculations and may also be directly measured based on baryon inventories. Taking everything together, the current best estimate is $\Omega_{b,0} = 0.049 \pm 0.001$. The remaining matter density is accounted for by dark matter which has a density parameter of $\Omega_{c,0} = 0.259 \pm 0.006$.

Although highly isotropic, the CMB exhibits anisotropies in intensity at the level of a few parts in 100 000 over a range of angular scales. These can be mapped, and are usually shown as variations in the temperature of the CMB, as shown in Figure 3.22. The angular power spectrum of an anisotropy map shows the level of variation that is present on any specified angular scale, as shown in Figure 3.23. Here the angular power is shown as a function of the multipole number l , where $l = 180^\circ/\theta$ and θ represents the angle between two particular directions on the sky. Values of cosmological parameters may be extracted from anisotropy measurements by comparing the observed angular power spectrum with that predicted by big bang cosmology.

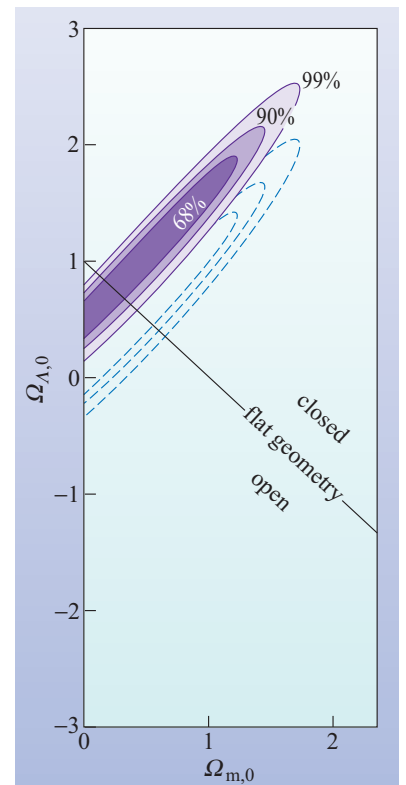


Figure 3.21 Results of the Supernova Cosmology Project, plotted as constraints (indicated by confidence level) on the current values of Ω_{Λ} and Ω_m . The results effectively rule out the kind of Universe in which $\Omega_{\Lambda,0} = 0$ and $\Omega_{m,0} = 1$ that was favoured by many cosmologists prior to the publication of the supernova data. (Adapted from Schwarzschild, 1998, based on the work of S. Perlmutter et al.)

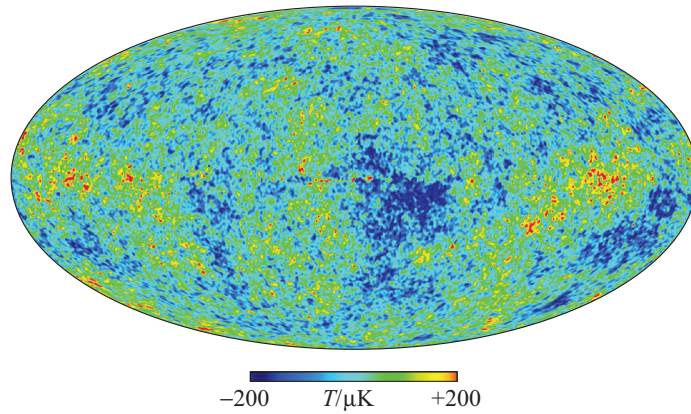


Figure 3.22 An all-sky CMB anisotropy map, based on data obtained by the WMAP space probe. The angular resolution of this map is about 0.1° . (Bennett et al., 2003)

WMAP measurements also indicate that the age of the Universe is $t_0 = (13.80 \pm 0.02) \times 10^9$ years. The results may indicate that we are now entering an era of precision cosmology in which cosmological speculations will be tightly constrained by measurements, and quantities that were previously very uncertain will become accurately known.

Exercise 3.5 Using the most accurate value for the Hubble constant, what is the best estimate for the current value of the critical density? (Use $1 \text{ Mpc} = 3.09 \times 10^{19} \text{ km}$; $G = 6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$.)

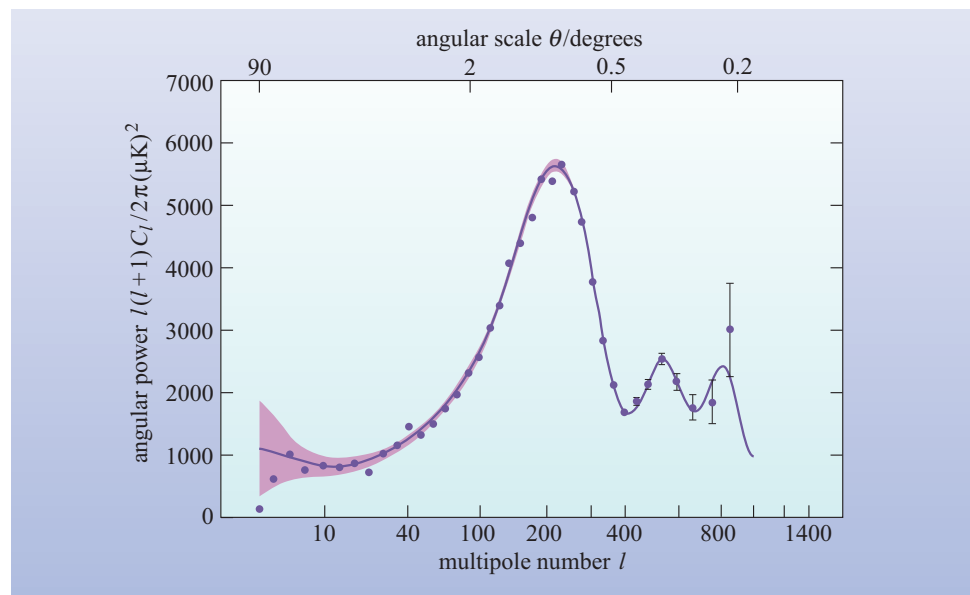


Figure 3.23 The angular power spectrum of the CMB as determined by WMAP. (Bennett et al., 2003)

3.9 Cosmological questions

The models devised by cosmologists are simplified representations of the Universe. Like all models, they are only partial analogies to reality and break down outside their limits of validity. The big bang model is successful as far as it goes, but there are several problems it cannot answer.

- Problem 1: *What is the dark matter?* Non-baryonic dark matter makes up about 26% of the energy density of the Universe. Its nature is largely unknown. The best candidate is the neutralino, a form of WIMP (Weakly Interacting Massive Particle), which may soon be discovered in laboratory experiments.
- Problem 2: *What is the dark energy?* Dark energy is a source of negative pressure that fills the Universe and drives the accelerating expansion. It should not be confused with dark matter. Its nature is still a mystery, but the leading contenders are Einstein's cosmological constant (a source of 'repulsive' gravity arising from general relativity), quantum vacuum energy (a consequence of Heisenberg's uncertainty principle) or 'quintessence' (an exotic form of matter).
- Problem 3: *Why is the Universe so uniform?* This is the horizon problem, which asks why widely separated regions have the same temperature and density, even though each has been beyond the horizon of the other throughout the history of the Universe. Inflation provides a possible answer. A small region of the Universe that had become homogeneous might have expanded so rapidly and by such an enormous factor that the whole of the currently observable part of the Universe (and perhaps more) is contained within the inflated homogeneous region.
- Problem 4: *Why does the Universe have a flat ($k = 0$) geometry?* Again, inflation may make it so. During the inflationary period large amounts of matter and energy were released into the Universe from the vacuum energy, leaving its density very close to the critical density, which corresponds to a flat geometry. Equivalently, whatever curvature the early Universe may have had would have been smoothed out by inflation leaving the spatial geometry of the observable Universe indistinguishable from that of a 'flat' space.
- Problem 5: *Where did the structure come from?* Clusters of galaxies were formed from density fluctuations in the early Universe which have left their imprint on the cosmic background radiation. Those fluctuations in turn may have arisen from tiny quantum fluctuations which were stretched by inflation from the microscopic scale up to and beyond the size of the then-observable Universe. At that point they would have become 'frozen in' as large-scale primordial fluctuations from which galaxies could condense.
- Problem 6: *Why is there more matter than antimatter?* Although one might expect equal numbers of particles and antiparticles to have been created in the early Universe, grand unified theories of physics allow a slight imbalance of matter over antimatter of 1 part in 10^{10} . The matter now in the Universe is that left over when the bulk of the matter and antimatter annihilated.
- Problem 7: *What happened at $t = 0$?* It's still anyone's guess. General relativity breaks down at the Planck time of 10^{-43} s, and to progress to earlier times requires a theory of quantum gravity that unifies general relativity with quantum physics. Inflation, too, remains without a firm grounding until this very early era is better understood. Limited progress has been made with quantum cosmology

but the new all-encompassing M-theory offers several intriguing lines of enquiry.

Summary of Chapter 3

1. The Milky Way is a typical spiral galaxy comprising a disc, a halo and a nuclear bulge. Young, high metallicity, population I stars are found mainly in the spiral arms whilst older, low metallicity, population II stars are found in the halo (including globular clusters) and the nuclear bulge. HII regions (ionized hydrogen) and open stars clusters are found in spiral arms and are associated with star formation.
2. Other galaxies are classified according to their shape as elliptical, lenticular, spiral or irregular. The most abundant galaxies are dwarf galaxies.
3. Galaxy masses are determined by various techniques, including the study of rotation curves (for spirals) and velocity dispersion (for ellipticals). Distances to other galaxies are determined by a range of techniques, including the use of standard candles such as Cepheid variable stars.
4. For nearby galaxies, the Hubble law relates the redshift of a galaxy to its distance via

$$z = \frac{\Delta\lambda}{\lambda} = \frac{H_0 d}{c}$$

The Hubble law is due to the overall expansion of the Universe and implies that the further away a galaxy is the faster it is receding from us; H_0 is about $70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. For relatively low speeds ($v < 0.1c$) the relationship $z = v/c$ holds true and therefore $v = H_0 d$.

5. A generic model for an active galaxy supposes that the central engine is a supermassive black hole surrounded by an accretion disc with jets emerging perpendicular to the accretion disc. The engine is surrounded by an obscuring torus of gas and dust. The broad line region occupies the hole in the middle of the torus and the narrow line region lies further out. The entire AGN appears as a bright nucleus in an otherwise normal galaxy.
6. Unified models of AGN attempt to explain the range of AGN on the assumption that they differ only in luminosity and the angle at which they are viewed.
7. A convenient way of comparing the energy output of an astronomical object over a broad range of wavelength (or frequency) is to plot the spectrum as a graph of λF_λ versus λ (or νF_ν versus ν). The spectral flux density F_λ (or F_ν) is defined as the power per unit area per unit wavelength range (or per unit frequency range) received from an astronomical object.
8. Galaxies are gravitationally clustered into groups and clusters. The mass of clusters of galaxies indicate that there is far more matter in clusters than the sum of the individual galaxy masses. This suggests the presence of dark matter in clusters.
9. The behaviour of spacetime in terms of the scale factor $R(t)$ may be described by the Friedmann equation, which depends on the values of the

curvature parameter k , the cosmological constant Λ and the density of the Universe ρ at some particular time:

$$\dot{R}^2(t) = \frac{8\pi G R^2(t)}{3} \left(\rho(t) + \frac{\Lambda c^2}{8\pi G} \right) - kc^2$$

The Friedmann–Robertson–Walker models are a set of cosmological models that are consistent with general relativity and the cosmological principle. They allow a variety of cases that are closed, critical, open, decelerating or accelerating.

10. The Hubble parameter is defined in terms of the scale parameter as

$$H(t) = \frac{\dot{R}}{R}$$

The critical density of the Universe (corresponding to a model with $\Lambda = 0$ and $k = 0$) is given by

$$\rho_{\text{crit}} = \frac{3H^2(t)}{8\pi G}$$

11. Recent results from the WMAP space probe and measurements of distant type Ia supernovae indicate $k = 0$, $\Omega_{\Lambda,0} = 0.691 \pm 0.006$, $\Omega_{\text{m},0} = 0.309 \pm 0.006$ and $\Omega_{\text{b},0} = 0.049 \pm 0.001$, implying a flat Universe dominated by dark energy, and in which most of the matter is non-baryonic dark matter. These measurements also indicate that $H_0 = (67.74 \pm 0.46) \text{ km s}^{-1} \text{ Mpc}^{-1}$, and that the age of the Universe is $t_0 = (13.80 \pm 0.02) \times 10^9$ years.
12. The Universe was created at the instant of the big bang. As it has aged, the Universe has cooled and distances within it have increased. At the earliest times, the four fundamental interactions were unified, but as the temperature of the Universe decreased, so these interactions became distinct. The earliest time about which anything can be said is the Planck time. Early in its history, the Universe is presumed to have undergone an extremely rapid period of expansion, known as inflation. One effect of this was to smooth out any irregularities, leading to today's remarkably uniform observable Universe. The early Universe contained almost equal numbers of particles and antiparticles, however, there was an asymmetry of a few parts in ten billion in favour of matter. The matter and antimatter underwent mutual annihilation and the result of this is that there are now about ten billion photons for every matter particle in the Universe. Equal numbers of protons and neutrons were initially produced in the Universe, however, free neutrons decay, and this reduced their number. All free neutrons were soon bound up within nuclei of deuterium, helium and lithium. The approximate distribution of mass in the Universe is about 25% helium-4 to 75% hydrogen, with small traces of other nuclei. Neutrinos ceased to interact with the rest of the Universe soon after protons and neutrons were formed, just before electrons and positrons annihilated. At around 300 000 years after the big bang, when the temperature was about 3000 K, photons produced from the matter–antimatter annihilations had their last interaction with the matter of the Universe. These photons, redshifted by a factor of a thousand by the subsequent expansion of the Universe, form the cosmic microwave background that is observed today.

13. Despite being in the era of 'precision cosmology', there are a number of unanswered questions. These include: What is the dark matter? What is the dark energy? Why is the Universe so uniform? Why does the Universe have a flat geometry? Where did the structure come from? Why is there more matter than antimatter? What happened at $t = 0$?

Chapter 4 Calculus

Introduction

In this second mathematical chapter we discuss one of the most powerful and useful techniques of mathematics, namely differential and integral calculus. Because calculus is a key *tool* in astrophysics and cosmology, there are a lot of worked examples in this section, many of which draw on astrophysical phenomena. This is *not* a mathematics course, so we are not interested in rigorous proofs of the various mathematical techniques. Rather, you must be comfortable using these tools and applying the techniques to real astrophysical and cosmological situations.

If you have not previously studied a Level 2 maths module (such as MST224 or MST210), you will need to study this chapter closely indeed. You must make sure you can answer all the questions it contains, and are able to reproduce the steps in each of the examples, before proceeding. If you have previously studied a Level 2 physics module (such as S217), then some of the basic concepts and techniques below should already be familiar to you from that material, but you will still have to pay close attention to the new ones that are introduced here. Do not be tempted to miss out the exercises in this chapter, it is only by repeated practice that you will become comfortable with applying and understanding these ideas and methods which are essential for studying astrophysics and cosmology.

4.1 Differentiation and curved graphs

Differentiation is a means of finding how one quantity changes as a result of changes in another. Where two quantities y and x are plotted on a graph, the rate of change of y in response to changes of x is simply the gradient (slope) of the graph. For quantities following a linear relation, the gradient can be calculated easily from the data values, as shown in Section 1.10. In this section, we consider the situation where the graph is curved, and in the next section we consider when the relationship between x and y is a known equation.

The *tangent* at a point on a curve is a straight line whose slope matches the tilt of the curve at that point. The gradient of a curved graph is simply the gradient of the tangent to the curve at the point of interest. Considering Figure 4.1, the gradient is very small for small values of x and y , becomes much higher for intermediate values of x and y , and falls to small values again for large values of x and y . So, whereas the gradient of a straight-line graph was the same for all values of x and y , for a curved graph the gradient changes depending on which part of the curve is considered.

It would be impractical to draw a tangent for every point on a curve to calculate the gradient, but some simple mathematical tools come to our aid. With reference to Figure 4.1, the gradient of the tangent at some point P can be approximated by the gradient of a straight line joining two points on either side of P, shown in the figure as $\Delta y/\Delta x$. This approximation is rather coarse if the chosen side points are too far away from the central point of interest, but it becomes more accurate the closer the points come to P.

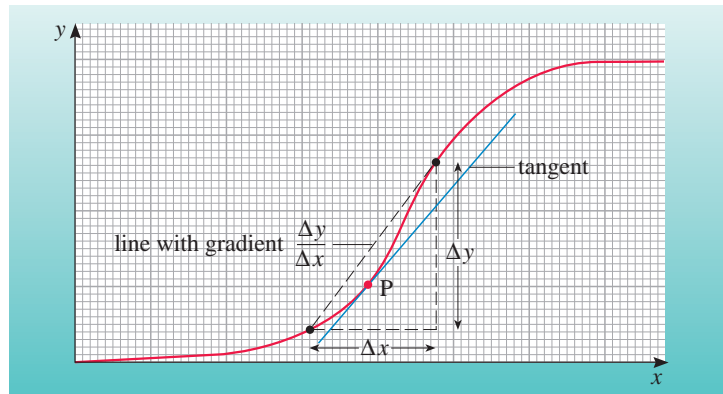


Figure 4.1 The gradient at a point P on a curved line is equal to the gradient of the tangent at that point.

In the limit where the points are so close to P as to be indistinguishable, the intervals Δy and Δx become infinitesimal, and the approximation becomes *exact*. The gradient in this limit is written mathematically not as $\Delta y/\Delta x$ but as dy/dx , the lowercase 'd's signifying the limit of infinitesimal intervals. That is, the gradient of a graph of y versus x at some point is described as

$$\frac{dy}{dx} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} \quad (4.1)$$

Thus dy/dx represents the rate of change of y with x , and is known mathematically as 'the derivative of y with respect to x '. The value of dy/dx usually changes from one point to the next, unless y is a linear function of x . Generally, if y is a function of x , then the derivative dy/dx will also be a function of x .

- What is the significance of a region of a graph where $dy/dx = 0$?
- This indicates a particular point on the graph where the gradient is zero, i.e. the graph is (momentarily) horizontal. The point(s) where dy/dx is zero therefore include locations where the graph is momentarily at a maximum or minimum.

When dy/dx (the gradient of a graph of y versus x) changes with x , it is often useful to ask *how rapidly* the *gradient* changes with x . We can determine this by applying the same process described above, not to y but to dy/dx , to find the gradient of dy/dx versus x . We would then be finding the derivative of the first derivative. Such a quantity is referred to as 'the second derivative of y with respect to x ', and extending the notation above it is written d^2y/dx^2 . (Of course, there is no reason to stop with second derivatives, since even this may vary with x , and it may be useful on rare occasions to calculate third or higher derivatives.)

In some cases it is useful to use shorthand notations for first and second derivatives, and you should be aware of them because you will encounter them in astrophysics and cosmology. Sometimes dy/dx and d^2y/dx^2 are denoted by y' and y'' (called 'y-prime' and 'y-double-prime'). Although you should note that primes are often used also for completely different topics having nothing to do with derivatives; the context of the problem will usually prevent confusion.

Another notation, \dot{y} and \ddot{y} (called 'y-dot' and 'y-double-dot') is often used for derivatives with respect to time, dy/dt and d^2y/dt^2 . This latter notation is particularly prevalent in astrophysics and cosmology. Note also that *any* symbol may be used for functions in this notation, not just y .

- If M is used to represent the mass of a star, what is meant by the notation \dot{M} ?
- \dot{M} ('em-dot') implies the rate of change of M with time (dM/dt). (The mass of a star may increase if it accretes material from a companion or decrease if it loses mass via a stellar wind or by transfer to a companion.)

4.2 Differentiation of known functions

A most valuable feature of derivatives is that, if we know the *analytic* or *functional form* (the equation) of the relationship between two variables, then derivatives can be obtained without having to draw graphs (though graphs may help visualize the relationship). In this document, rather than show the proofs of differentiation, we tabulate the derivatives of some general functions that are likely to be encountered in astrophysics and cosmology.

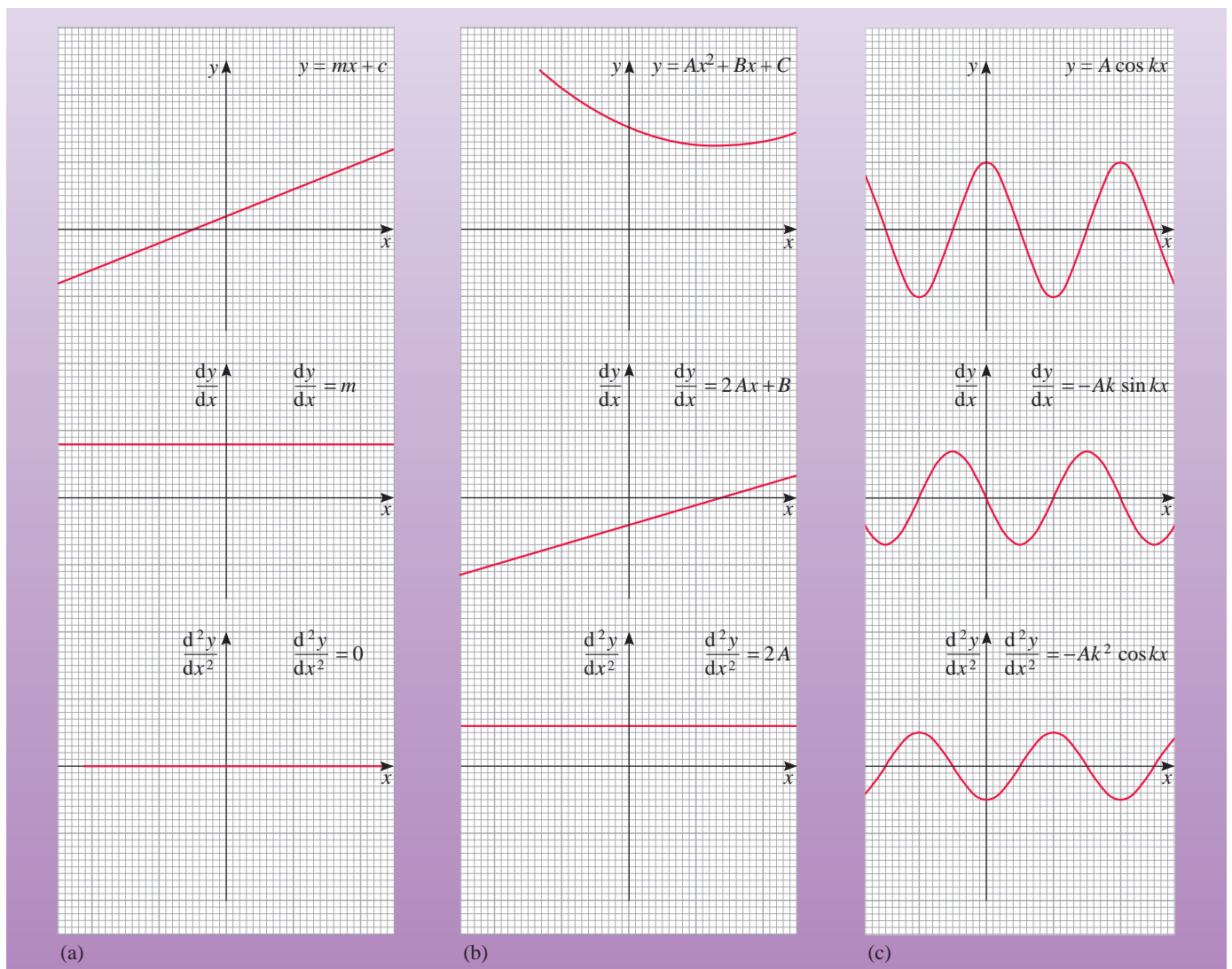


Figure 4.2 Three functions of x (a straight line, a parabola and a sinusoid) and their first and second derivatives.

Table 4.1 Some functions and their derivatives

Function	Derivative
$f(t)$	df/dt or f' or \dot{f}
a	0
at	a
at^n	nat^{n-1}
$a \sin(\omega t)$	$a\omega \cos(\omega t)$
$a \cos(\omega t)$	$-a\omega \sin(\omega t)$
$a \exp(kt)$	$ak \exp(kt)$
$\log_e t$	$1/t$
$a \log_e t$	a/t

In the following tabulation, derivatives are given with respect to some variable t . The symbols ω , a , n and k are constants (that is, independent of t), while u and v do depend on t , i.e. they are functions of t . Some simple functions (of an arbitrary variable x) and their first and second derivatives are shown in Figure 4.2. Convince yourself that you can work out the first and second derivatives using the results in Table 4.1.

In general, when faced with a function to differentiate, the first aim is to reduce the formula to something which may be differentiated using one of the rules in Table 4.1. To this end, it is often possible to use either the **sum rule** and **product rule**, namely

$$\frac{d(u+v)}{dt} = \frac{du}{dt} + \frac{dv}{dt} \quad (4.2)$$

$$\frac{d(uv)}{dt} = u \frac{dv}{dt} + v \frac{du}{dt} \quad (4.3)$$

where u and v are themselves functions of t in this case. Note that the variables u , v and t are merely used as examples, and the two rules can of course be used for any combination of variables. A couple of examples drawn from astrophysics should make the process clear.

Essential skill:

Using the sum rule for differentiation

Worked Example 4.1

The probability for an atomic nucleus with energy E in the core of a star to undergo nuclear fusion depends on two terms: $p(E) = -(E/kT) - (E_G/E)^{1/2}$, where k , T and E_G are constants in this case. Use the sum rule to evaluate the rate of change of this probability expression with respect to energy, dp/dE .

Solution

We use the sum rule, with $u = (E/kT)$ and $v = (E_G/E)^{1/2}$. Using the rule from line 3 of Table 4.1,

$$\frac{du}{dE} = \frac{1}{kT}$$

We now note that $(E_G/E)^{1/2}$ can be written $E_G^{1/2} \times E^{-1/2}$, so using the rule from line 4 of Table 4.1,

$$\frac{dv}{dE} = E_G^{1/2} \times \left(-\frac{1}{2} E^{-3/2} \right) = -\frac{E_G^{1/2}}{2E^{3/2}}$$

Then applying the sum rule,

$$\frac{d(u+v)}{dE} = \frac{du}{dE} + \frac{dv}{dE}$$

$$\text{so } \frac{dp}{dE} = -\frac{1}{kT} - \left(-\frac{E_G^{1/2}}{2E^{3/2}} \right)$$

$$\text{or } \frac{dp}{dE} = \left(\frac{E_G}{4E^3} \right)^{1/2} - \frac{1}{kT}$$

Worked Example 4.2

Suppose the pressure at a location inside an interstellar gas cloud varies according to the following equation as a sound wave travels through it:

$$P(t) = (A/t^2) \times \sin(\omega t)$$

where A and ω are constants. Use the product rule to determine the rate of change of pressure with respect to time, dP/dt .

Solution

We first write $u = At^{-2}$ and $v = \sin(\omega t)$, so that $P(t) = uv$. Using the various rules from Table 4.1 we then write $du/dt = (-2At^{-3})$ and $dv/dt = \omega \cos(\omega t)$. Then we use the product rule,

$$\frac{d(uv)}{dt} = u \frac{dv}{dt} + v \frac{du}{dt}$$

which in this case becomes

$$\frac{d(uv)}{dt} = \frac{A}{t^2} \omega \cos(\omega t) + \sin(\omega t) \times \left(-\frac{2A}{t^3} \right)$$

and this can be simplified a little to become

$$\frac{dP}{dt} = \left(\frac{A}{t^2} \right) \times \left[\omega \cos(\omega t) - \left(\frac{2}{t} \right) \sin(\omega t) \right]$$

Essential skill:

Using the product rule for differentiation

Exercise 4.1 (a) If $F(r) = k/r$, what is dF/dr ? (b) If $G(x) = ax^2 \sin(bx)$ what is dG/dx ? (You may assume that k , a and b are constants.)

4.3 The exponential function

From Table 4.1, if $y = a \exp(kt)$, then $dy/dt = ak \exp(kt)$. Note, however, that $ak \exp(kt)$ is just ky . That is,

$$\text{if } y = a \exp(kt) \text{ then } dy/dt = ky \quad (4.4)$$

In words, if $y = a \exp(kt)$, then the derivative of y (the gradient of the graph of y versus t) is directly proportional to y itself. This is the special property of the number e that was hinted at in the earlier discussion of natural logarithms (Section 1.9).

You will often see the exponential function ($\exp x$) represented by ' e^x '. So be aware that

$$\exp x \equiv e^x \quad (4.5)$$

A process in which the relevant quantity varies according to the relation $dy/dt = ky$ is referred to as an **exponential** process. Although these are frequently functions of time, as in the examples here, exponential processes are

also encountered as functions of distance, or some other variable. Examples of exponential curves are shown in Figure 4.3 for different values of k and Z_0 . In each case the two quantities Z and t are related according to the equation

$$Z(t) = Z_0 \exp(kt) \quad (4.6)$$

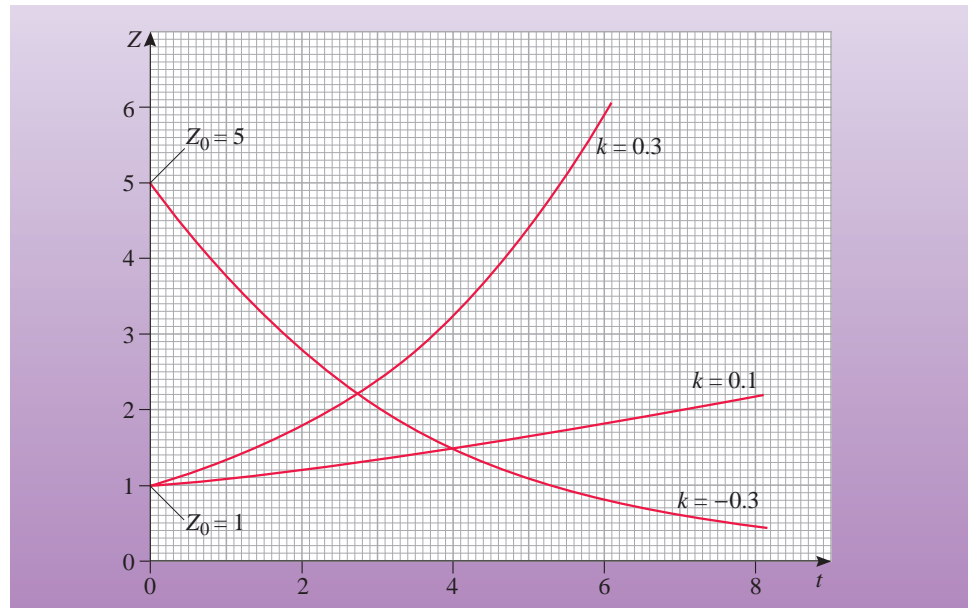


Figure 4.3 Exponential curves, $Z = Z_0 \exp(kt)$, for various values of Z_0 and k .

It is a general property of exponential growth curves (i.e. $k > 0$) that they approach a value of ∞ as the variable (in this case t) approaches ∞ and they tend to zero as the variable approaches $-\infty$. If the constant k is negative, then the process is an exponential decay process, rather than an exponential growth process. Exponential decay curves (i.e. $k < 0$) tend to zero as the variable (t) approaches ∞ , and approach ∞ as the variable approaches $-\infty$. Notice also that at $t = 0$, $\exp(kt) = 1$ and so $Z(0) = Z_0$ irrespective of the value or sign of k .

If an exponential process is a function of time, then the constant $\tau = 1/|k|$ is referred to as the ‘time constant’, ‘lifetime’, or ‘e-folding time’, since it is the time over which y increases or decreases by a factor of e (approximately 2.718). (τ is the Greek letter *tau* pronounced to rhyme with ‘cow’.) So an alternative version of the expression for an exponential process is

$$Z(t) = Z_0 \exp(t/\tau) \quad (4.7)$$

Exercise 4.2 As noted above, it is possible to describe an exponential decay by its e-folding time τ . However, one of the best known exponential functions, radioactive decay, is often characterized by its half-life $\tau_{1/2}$, the time it takes for the quantity (and decay rate) to halve. Calculate the ratio of these characteristic times, $\tau_{1/2}/\tau$. Confirm your calculation using Figure 4.4. (*Hint*: The exponential decay obeys the equation $y(t) = \exp(-t/\tau)$. When one half-life $\tau_{1/2}$ has elapsed, the new value of y is given by $y(t + \tau_{1/2}) = \exp[-(t + \tau_{1/2})/\tau]$, but this must be equal to *half* the original value, i.e. $y(t + \tau_{1/2}) = 0.5 \exp(-t/\tau)$.)

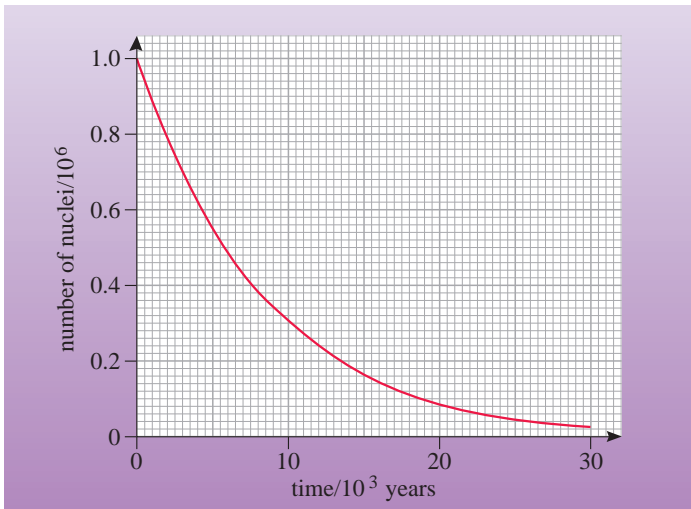


Figure 4.4 An example of an exponential decay, showing the number of radioactive carbon nuclei in a sample plotted against time. The time constant of this process is 8100 years (2.56×10^{11} s) and the initial number of nuclei is $N_0 = 1.0 \times 10^6$.

4.4 The chain rule

It is sometimes the case that we need to differentiate a function which is itself a function. A simple example is the expression $y(x) = \exp(x^3)$, where the function x^3 is itself the argument of an exponential function. Bearing in mind that the way forward is to reduce the formula in question to something that can be differentiated using one or more of the rules from Table 4.1, the way around this is to substitute one variable for another and apply the **chain rule**:

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx} \quad (4.8)$$

Of course, the variables y , x and u above are merely for illustration, and any three variables could be used. In the example introduced above, we would set $u = x^3$, so that we have a simple function $y(u) = \exp u$ to differentiate. Then $dy/du = \exp u$ and $du/dx = 3x^2$. So applying the chain rule, the required answer is $dy/dx = 3x^2 \times \exp u = 3x^2 \exp(x^3)$.

Notice that Equation 4.8 implies that differentials can be treated in some ways like common fractions. Although this is not strictly true, you will generally find that you *can* treat differentials in this way without causing any problems. The following examples drawn from astrophysics should help make the technique clear.

Worked Example 4.3

The pressure P inside a star, as a function of radius r from the centre, can be expressed as

$$P(r) = 2\pi G \rho_c^2 a^2 / 3 [\exp(-r^2/a^2) - \exp(-R^2/a^2)]$$

The quantities represented by G , ρ_c , a and R are all constants. What is the rate of change of pressure with respect to radius, dP/dr ?

Essential skill:

Using the chain rule to solve astrophysical problems

Solution

The first thing we can do is to simplify the expression somewhat by replacing the constant terms with some arbitrary constants, by writing $P(r) = A[\exp(-r^2/a^2) - B]$.

Now, the way to solve this problem is to substitute another variable for the function inside the bracket of the exponential. Suppose we write $u = -r^2/a^2$, then the original function becomes $P(u) = A[\exp(u) - B]$ which is clearly a lot simpler than the one we started with. Now $dP/du = A \exp u$. However, we require dP/dr not dP/du , so we use the chain rule

$$\frac{dP}{dr} = \frac{dP}{du} \times \frac{du}{dr}$$

where, in this case, $du/dr = -2r/a^2$. So we have

$$\frac{dP}{dr} = (A \exp u) \times \left(-\frac{2r}{a^2}\right)$$

Now, replacing the original substitution for u the answer is

$$\begin{aligned} \frac{dP}{dr} &= A \exp\left(-\frac{r^2}{a^2}\right) \times \left(-\frac{2r}{a^2}\right) \\ \text{or} \quad \frac{dP}{dr} &= \left(-\frac{2rA}{a^2}\right) \exp\left(-\frac{r^2}{a^2}\right) \end{aligned}$$

and finally replacing the terms for the constant A we have

$$\begin{aligned} \frac{dP}{dr} &= \left(-\frac{2r}{a^2} \times \frac{2\pi G \rho_c^2 a^2}{3}\right) \exp\left(-\frac{r^2}{a^2}\right) \\ \text{or} \quad \frac{dP}{dr} &= \left(-\frac{4\pi G \rho_c^2 r}{3}\right) \exp\left(-\frac{r^2}{a^2}\right) \end{aligned}$$

which describes the rate of change of pressure inside a star with respect to the distance from the centre.

Essential skill:
Using the chain rule to solve astrophysical problems

Worked Example 4.4

The probability for fusion of an atomic nucleus in the core of a star is given by

$$P = A \exp \left[-\frac{E}{kT} - \left(\frac{E_G}{E} \right)^{1/2} \right]$$

where A , k and E_G are constants, E is the kinetic energy of the nucleus and T the temperature. What is the rate of change of the probability P with respect to energy E ?

Solution

We need to calculate dP/dE . The way to solve this is to substitute another variable for the embedded function. Suppose we write

$u = -(E/kT) - (E_G/E)^{1/2}$, then the original function becomes simply $P(u) = A \exp u$, and it is a simple matter to evaluate $dP/du = A \exp u$.

However, we require dP/dE not dP/du , so we use the chain rule

$$\frac{dP}{dE} = \frac{dP}{du} \times \frac{du}{dE}$$

and the task is now to calculate du/dE .

Using the sum rule and standard derivatives from Table 4.1 (as in Worked Example 4.1),

$$\frac{du}{dE} = -\frac{1}{kT} - \left(-\frac{1}{2} \times \frac{E_G^{1/2}}{E^{3/2}} \right) = \left(\frac{E_G}{4E^3} \right)^{1/2} - \frac{1}{kT}$$

Applying the chain rule therefore,

$$\frac{dP}{dE} = A \exp u \times \left[\left(\frac{E_G}{4E^3} \right)^{1/2} - \frac{1}{kT} \right]$$

and replacing the original substitution the final answer is

$$\frac{dP}{dE} = \left[\left(\frac{E_G}{4E^3} \right)^{1/2} - \frac{1}{kT} \right] A \exp \left[-\frac{E}{kT} - \left(\frac{E_G}{E} \right)^{1/2} \right]$$

This is a fearsome looking expression (!) but it has a useful consequence which is not difficult to work out. A graph of the original expression reaches a maximum value at some particular value of E – the energy at which the probability of fusion occurring is the highest. What is that value of E ? It is where the rate of change of P with respect to E is zero, i.e. where the slope of the graph is zero (horizontal). It is easy to determine when dP/dE is zero from the expression we have just calculated, it will occur when the term in front of the exponential is zero, namely when

$$\left(\frac{E_G}{4E^3} \right)^{1/2} - \frac{1}{kT} = 0$$

which can be rearranged to give

$$\frac{E_G}{4E^3} = \left(\frac{1}{kT} \right)^2$$

or $E^3 = \frac{E_G(kT)^2}{4}$

and therefore $E = [E_G(kT)^2/4]^{1/3}$. So the maximum probability for fusion to occur is when a nucleus has just this amount of energy.

Exercise 4.3 When a quasar ejects a cloud of plasma at high speed, the transverse velocity of the cloud measured by an observer on Earth is given by

$$V = \frac{\beta \sin \theta}{1 - \beta \cos \theta} \times c$$

where c is the speed of light, β is the speed of ejection divided by the speed of light and θ is the angle between the line of sight and direction of ejection.

(a) Use the product rule for differentiation by setting $p = c\beta \sin \theta$ and $q = (1 - \beta \cos \theta)^{-1}$ to calculate $dV/d\theta$. (*Hint*: You will need to use the chain rule in order to work out $dq/d\theta$. You will also need to use the identity $\sin^2 \theta + \cos^2 \theta = 1$ in order to simplify things during this calculation.)

(b) Use this result to determine the angle at which the maximum value of V is observed. (*Hint*: This will occur when the rate of change of V with respect to θ is zero, i.e. when $dV/d\theta = 0$.)



4.5 Logarithmic differentiation

A technique you will often come across in astrophysics and cosmology is that of **logarithmic differentiation**. The process as usual is best considered in terms of a couple of examples. Example 4.5 illustrates the derivative with respect to time of a logarithmic function and Example 4.6 is a specific example drawn from astrophysics.

Worked Example 4.5

What is the derivative with respect to time of the quantity $\log_e x$?

Solution

First, we use the chain rule to note that

$$\frac{d(\log_e x)}{dt} = \frac{d(\log_e x)}{dx} \times \frac{dx}{dt}$$

From the rule given at the bottom of Table 4.1,

$$\frac{d(\log_e x)}{dx} = \frac{1}{x}$$

So
$$\frac{d(\log_e x)}{dt} = \frac{1}{x} \times \frac{dx}{dt}$$

and, as noted in Section 4.1, dx/dt can more compactly be written as \dot{x} (pronounced ‘ex-dot’).

So the derivative with respect to time of $\log_e x$ is simply

$$\frac{d(\log_e x)}{dt} = \frac{\dot{x}}{x} \tag{4.9}$$

This is a very powerful result. Stated in words, the derivative with respect to time of the natural logarithm of a variable x is equal to the rate of change of x with respect to time divided by the variable x .

Essential skill:
Using logarithmic differentiation

Worked Example 4.6

In a binary star system, mass is transferred from one star to the other. The magnitude of the angular momentum (see Section 5.5) of such a system at a given instant is

$$J = M_1 M_2 \left(\frac{Ga}{M_1 + M_2} \right)^{1/2}$$

where M_1 and M_2 are the masses of the two stars, a is their separation and G is the gravitational constant. Carry out the logarithmic differentiation of this equation with respect to time in order to express how the separation of the stars varies in terms of the angular momentum and masses of the stars. (You may assume that all the mass lost from one star is accreted by the other, so that the total mass of the system, $M_1 + M_2$, is constant.)

Solution

First we take natural logarithms of each side of the equation (refer back to Section 1.9 if necessary):

$$\log_e J = \log_e(M_1 M_2) + \frac{1}{2} \log_e(Ga) - \frac{1}{2} \log_e(M_1 + M_2)$$

For simplicity we replace $M_1 + M_2$ by the total mass of the system M , and rearrange further to give

$$\log_e J = \log_e M_1 + \log_e M_2 + \frac{1}{2} \log_e G + \frac{1}{2} \log_e a - \frac{1}{2} \log_e M$$

Then we differentiate each part of this expression with respect to time, noting the general result from above that $d(\log_e x)/dt = \dot{x}/x$. So we have

$$\frac{\dot{J}}{J} = \frac{\dot{M}_1}{M_1} + \frac{\dot{M}_2}{M_2} + \frac{\dot{G}}{2G} + \frac{\dot{a}}{2a} - \frac{\dot{M}}{2M}$$

However, G and M are constants, so their rates of change with respect to time are zero in each case, i.e. $\dot{G} = \dot{M} = 0$, so this becomes

$$\frac{\dot{J}}{J} = \frac{\dot{M}_1}{M_1} + \frac{\dot{M}_2}{M_2} + \frac{\dot{a}}{2a}$$

which may be rearranged as

$$\frac{\dot{a}}{a} = \frac{2\dot{J}}{J} - \frac{2\dot{M}_1}{M_1} - \frac{2\dot{M}_2}{M_2}$$

This is an important equation governing the evolution of binary stars.

Essential skill:

Using logarithmic differentiation

4.6 Expansions

One of the most useful mathematical tools involving calculus is known as Taylor's theorem. It is named after Brook Taylor, an English mathematician who published

the result in 1715. It provides a way of rewriting a (possibly) complex function in terms of its successive derivatives, such that it can be evaluated more easily. Taylor's theorem can be stated as follows: a function $f(x)$ can be 'expanded' about the point $x = a$ by writing the **Taylor series**:

$$f(x) = f(a) + (x-a)f'(a) + (x-a)^2 f''(a)/2! + (x-a)^3 f'''(a)/3! \dots \quad (4.10)$$

where the '!' indicates the mathematical operation of a *factorial*, such that $2! = 2 \times 1$, $3! = 3 \times 2 \times 1$, etc. and the 'prime' notation has been used to indicate derivatives such that $f' = df/dx$, $f'' = d^2 f/dx^2$, etc.

The special case where the point about which the expansion is made is $a = 0$ had been derived earlier by the Scottish mathematician Colin Maclaurin. The **Maclaurin series** may be written:

$$f(x) = f(0) + x f'(0) + x^2 f''(0)/2! + x^3 f'''(0)/3! \dots \text{etc.} \quad (4.11)$$

As noted above, the use of either the Taylor series, or the related Maclaurin series, can enable a complex function to be evaluated in an approximate manner. As usual, a couple of examples will help to demonstrate this. The first is a general technique which crops up in many areas of physical science, the second is a specific example of importance to electromagnetic spectra and astrophysics.

Worked Example 4.7

What is the Maclaurin series expansion of the function $f(x) = (1+x)^n$?

Solution

We work out the first, second and third derivatives of the function as

$$f'(x) = n(1+x)^{n-1}$$

$$f''(x) = n(n-1)(1+x)^{n-2}$$

$$f'''(x) = n(n-1)(n-2)(1+x)^{n-3}$$

So the Maclaurin series expansion is

$$(1+0)^n + xn(1+0)^{n-1} + \frac{x^2 n(n-1)(1+0)^{n-2}}{2} + \frac{x^3 n(n-1)(n-2)(1+0)^{n-3}}{6} + \dots$$

Since 1 raised to any power is still 1, this becomes

$$1 + xn + x^2 n(n-1)/2 + x^3 n(n-1)(n-2)/6 + \dots$$

The usefulness of this particular expansion becomes apparent if we restrict ourselves to values of x that are between -1 and $+1$, in other words if the magnitude of x is a small number: $|x| < 1$. In that case x^2 will be smaller than $|x|$ and x^3 smaller still. So we may make the following approximation:

For $|x| < 1$, the first-order expansion is

$$(1+x)^n \approx 1 + nx \quad (4.12)$$

Essential skill:
Using expansions to solve problems

- Calculate $(1 + 0.02)^{-1.3}$ using your calculator, and then work out the first order expansion using the equation above.
- The accurate value is $(1 + 0.02)^{-1.3} = 0.974\,585\,120\,98$
The first order expansion is $(1 + 0.02)^{-1.3} \approx 1 + (-1.3 \times 0.02) = 0.974$.

Worked Example 4.8

The Planck function (see Section 5.9) describes the power per unit area per unit frequency per unit solid angle emitted by a so-called black-body source of electromagnetic radiation. It may be written as the rather fearsome formula:

$$B_\nu(T) = \left(\frac{2h\nu^3}{c^2} \right) \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} \text{ W m}^{-2} \text{ Hz}^{-1} \text{ sr}^{-1}$$

where ν is the frequency of radiation, T is the temperature of the black-body, and h , c and k are all constants. In the low-frequency limit, i.e. $h\nu \ll kT$, the Planck function reduces to a much simpler expression known as the Rayleigh–Jeans formula.

Use a first-order Maclaurin series expansion of the exponential function to derive the Rayleigh–Jeans formula from the Planck function.

Solution

We take the exponential function $f(\nu) = \exp(h\nu/kT)$ and calculate its first derivative as $f' = (h/kT) \exp(h\nu/kT)$. So the first-order expansion of $f(\nu)$ is

$$\begin{aligned} f(\nu) &= f(0) + \nu f'(0) \\ f(\nu) &= \exp 0 + (h\nu/kT) \exp 0 \end{aligned}$$

Since $\exp 0 = 1$, this becomes simply

$$f(\nu) = 1 + (h\nu/kT)$$

So substituting this back into the Planck function, we have

$$\begin{aligned} B_\nu(T) &= \frac{2h\nu^3}{c^2} \frac{1}{1 + (h\nu/kT) - 1} \\ B_\nu(T) &= \frac{2h\nu^3}{c^2} \frac{kT}{h\nu} \\ B_\nu(T) &= \frac{2kT\nu^2}{c^2} \end{aligned}$$

which is the low-frequency approximation to the Planck function known as the Rayleigh–Jeans formula.

Essential skill:

Using expansions to solve problems

- Exercise 4.4** (a) Use a first-order Maclaurin series expansion to verify the small-angle approximation: $\sin \theta \approx \theta$, where the angle θ is small and in radians.
(b) What would a slightly more accurate expansion be?

4.7 Partial differentiation

It is often the case that a function will depend on two (or more) variables, rather than a single variable. For instance, the function $h(x, y)$ may represent the height of the landscape as a function of two position coordinates x and y . The rate of change of h with respect to (say) x only is written as $\partial h(x, y)/\partial x$ and read as ‘partial dee h by dee x’. The symbol ∂ indicates that a derivative with respect to only one of the variables (that on the bottom line) is to be considered. For the purposes of the differentiation, the other variables are treated as constants. So in the example above, the partial derivative of h with respect to x represents the slope of the land in the x -direction (Figure 4.5).

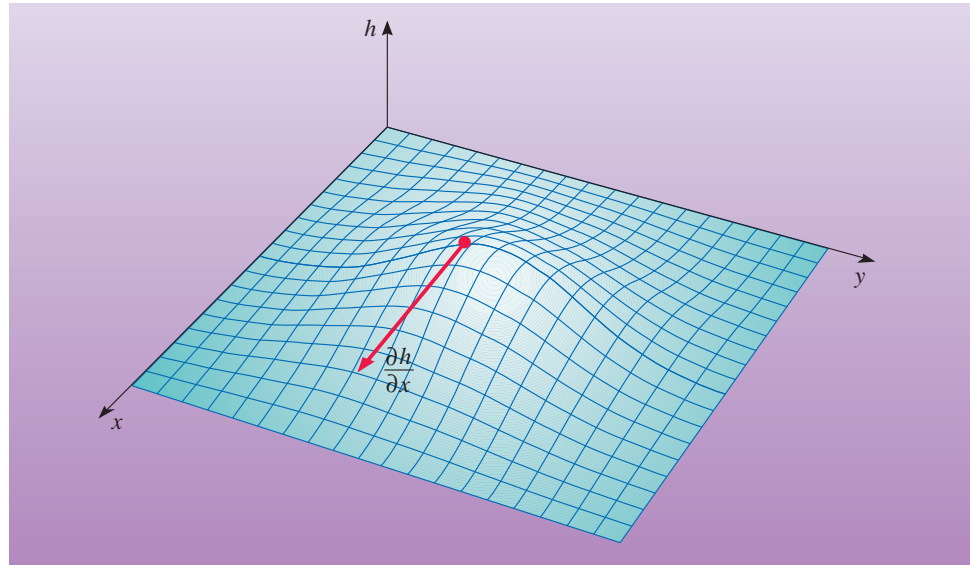


Figure 4.5 The partial derivative of the function describing the height of the landscape is the slope of the land in a particular direction.

Essential skill:
Using partial derivatives

Worked Example 4.9

The height of a particular region obeys the equation $h(x, y) = x^2 + 3xy + 4y^2$. What is the slope of the land in the y -direction?

Solution

The slope of the land in the y -direction is found as the partial derivative of h with respect to y . For the purposes of this, we treat x as just another constant, so,

$$\frac{\partial h(x, y)}{\partial y} = 3x + 8y$$

- Exercise 4.5** (a) What is the partial derivative with respect to x of the function $y(x, t) = A \sin(kx + \omega t)$, where k and ω are constants?
- (b) What is the partial derivative of $y(x, t)$ with respect to t ?
- (c) What are the *second* partial derivatives of $y(x, t)$ with respect to x and t , namely $\partial^2 y/\partial x^2$ and $\partial^2 y/\partial t^2$?

4.8 Differentiation and vectors

Differentiation is not restricted to scalars; it can be done to vectors too.

$$\text{If } \mathbf{p} = (p_x, p_y, p_z) \quad \text{then} \quad \frac{d\mathbf{p}}{dt} = \left(\frac{dp_x}{dt}, \frac{dp_y}{dt}, \frac{dp_z}{dt} \right)$$

Note furthermore that the derivative of a vector is also a vector, so has direction as well as magnitude. It is possible to imagine cases where $p = \text{constant}$, but $d\mathbf{p}/dt \neq 0$. That is, where the magnitude of the vector is constant, but the direction changes.

In order to consider the differentiation of a *field*, it is useful to define the vector differential operator **nabla** (also sometimes called *del*, or the *grad* operator) in terms of its three components as

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (4.13)$$

where the ∂ symbol represents a partial derivative. Nabla is called a *vector differential operator* because it is a vector that can be used to perform a differential operation on other quantities, namely fields. Nabla is itself a three-dimensional vector with the three components indicated. It can be applied to either scalar or vector fields in several ways, as described below.

The **gradient of a scalar field**, such as $T(x, y, z)$, is defined as ∇T , pronounced ‘grad T’. Since ∇ is a vector and T is a scalar field (i.e. a scalar function defined at each point in space), the expression ∇T is a vector field. So we can write the components of ∇T as

$$\nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z} \right) = \text{a vector field} \quad (4.14)$$

Suppose $T(x, y, z)$ is a scalar field representing the temperature inside a star. The gradient of the scalar field is simply the way in which the temperature changes with distance in each direction – the temperature gradient in kelvin per metre if you like. The vector ∇T points in the direction of the steepest slope.

Worked Example 4.10

Suppose T is a scalar field represented by $T = 3xy^2 + 4yz^2 + 5zx^2$. What is the gradient of this scalar field at any point?

Solution

The x -component of the gradient is $\partial T/\partial x = 3y^2 + 10zx$; the y -component of the gradient is $\partial T/\partial y = 6xy + 4z^2$; and the z -component of the gradient is $\partial T/\partial z = 8yz + 5x^2$. So the gradient of the scalar field T is the vector field with components $(3y^2 + 10zx, 6xy + 4z^2, 8yz + 5x^2)$.

Exercise 4.6 If a certain scalar field $h(x, y)$ represents the altitude of the landscape as a function of position coordinates x and y , what does the gradient of this scalar field ∇h represent? ■

The vector differential operator ∇ can also operate on vector fields, as well as on scalar fields. However, as you know from Section 1.13, there are two ways of multiplying vectors together. So, as you might guess, there are two ways of using the vector differential operator when operating on vector fields. The equivalent of the scalar (or dot) product is the operation known as **divergence** or *div*, whilst the equivalent of the vector (or cross) product is the operation known as **curl**.

Consider a vector field \mathbf{A} with components A_x, A_y, A_z at each point in space. The divergence of this vector field is given by

$$\operatorname{div} \mathbf{A} = \nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} = \text{a scalar field} \quad (4.15)$$

The curl of a vector field \mathbf{A} is given by

$$\begin{aligned} \operatorname{curl} \mathbf{A} &= \nabla \times \mathbf{A} & (4.16) \\ &= \left(\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \right), \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \right), \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) \right) \\ &= \text{a vector field} \end{aligned}$$

The new vector field which is the curl of the vector field \mathbf{A} , describes the rotation of the original vector field. For instance, if the vector field \mathbf{A} represents the flow velocity of a moving fluid, then the curl is the circulation density of the fluid.

4.9 Differential equations

In Chapter 1, you saw how to rearrange equations so that the variable of interest is isolated on the left-hand side. However, it is not always possible to solve equations in such a straightforward manner. One such form of equation that defies a simple algebraic solution is when a quantity and its derivative both appear together. Such an equation is called a **differential equation**.

In astrophysics and cosmology, several common differential equations are encountered. One is where the decay rate of a radioactive sample is proportional to the amount present: $dy/dt = -ky$. A second example is simple harmonic motion, in which the force on an object is proportional to its displacement from an equilibrium position: $md^2x/dt^2 = -kx$.

Many differential equations, including most that you will encounter, have *general solutions* that can be applied when needed. General solutions are functions (usually, but not always, expressed as an equation) that always satisfy the differential equation, but which necessarily contain extra constants whose values can only be determined by reference to the specific problem at hand. If a differential equation contains only first derivatives (a first-order differential equation), then the general solution will contain just one extra constant compared to the original equation. The general solution for a differential equation involving second derivatives (a second-order differential equation) will require two additional constants.

For example, the general solution for the radioactive decay problem has already been met in Section 4.3. You can verify that $y = a \exp(-kt)$ is a solution by substituting this into the differential equation $dy/dt = -ky$, recalling from

Table 4.1 that the derivative of $a \exp(-kt)$ with respect to t is $-ak \exp(-kt)$. Note that the solution has one additional parameter, a , compared to the original differential equation, so is the general solution.

The differential equation

$$\frac{dy}{dt} = -ky \quad (4.17)$$

has a solution

$$y = a \exp(-kt) \quad (4.18)$$

Additional information is required to determine the values of the extra constants in the general solution and thus to convert the general solution into a *particular solution* for a particular problem. For the radioactive decay problem, this would involve finding the value of a . The additional information often comes in the form of *initial conditions* or *boundary conditions* that describe the state of the physical system at some particular time or place. For the decay problem, knowledge of the decay rate at one time would provide enough information to determine the value of a in terms of k (which is specified as part of the problem). If the *value* of k is not already specified in the problem, then measurement of a second decay rate at a second time will complete the information to measure both a and k .

Another important type of differential equation is that describing simple harmonic motion: $m d^2x/dt^2 = -kx$. This describes the acceleration of a particle which is subject to a force whose magnitude is proportional to its displacement and whose direction is such as to always oppose the motion. The general solution for simple harmonic motion is $x = A \sin(\omega t + \phi)$.

The differential equation

$$m \frac{d^2x}{dt^2} = -kx \quad (4.19)$$

has a solution

$$x = A \sin(\omega t + \phi) \quad (4.20)$$

$$\text{where } \omega = \sqrt{k/m} \quad (4.21)$$

Note the appearance of two additional constants in the general solution, A and ϕ , as expected for a differential equation containing second derivatives. Finding a particular solution for x in this case would require knowledge of the parameters k and m (specified as part of the problem) to give ω , plus two initial and/or boundary conditions capable of giving the amplitude A of the oscillation and the initial phase ϕ .

Exercise 4.7 Use the standard rules in Table 4.1 to verify that $x = A \sin(\omega t + \phi)$ is a solution to the differential equation $m d^2x/dt^2 = -kx$, where $\omega = \sqrt{k/m}$.



4.10 Integration and curved graphs

Previous sections have described how the gradient of a graph of y versus x has a special physical significance: it tells us the rate of change of the y -variable as the x -variable changes. The gradient can be computed by graphical methods or by differentiation.

The *area under a graph* of y versus x , meaning the area enclosed by (i) the plotted curve, (ii) the x -axis, and (iii) two boundaries formed by vertical lines at two particular values of x , also often has physical significance. For example, if we have a graph of the speed of an object as a function of time, the area under the curve indicates the distance travelled by the object between two specified values of time. This result probably isn't clear to you yet – it will be explained below – but it will give you an idea of why it is useful in physics and astrophysics to study the area under a graph.

Begin by considering a body moving with constant speed v . The speed is given by the distance Δs travelled in some time interval Δt , divided by the time interval: $v = \Delta s / \Delta t$. If we know the speed, but not the distance travelled, we can nevertheless calculate the distance travelled from the rearranged equation: $\Delta s = v \Delta t$.

Now consider a body moving with non-uniform speed, whose speed versus time graph is shown in Figure 4.6a. If we consider a brief time interval Δt_1 from $t = t_A$ to $t = t_A + \Delta t_1$ during which the speed changes very little, then we can *approximate* the distance travelled as $v_1 \Delta t_1$. The subscript '1' on the speed indicates that we want the speed during the time interval Δt_1 . The distance Δs_1 travelled during this time interval is $\Delta s_1 \approx v_1 \Delta t_1$.

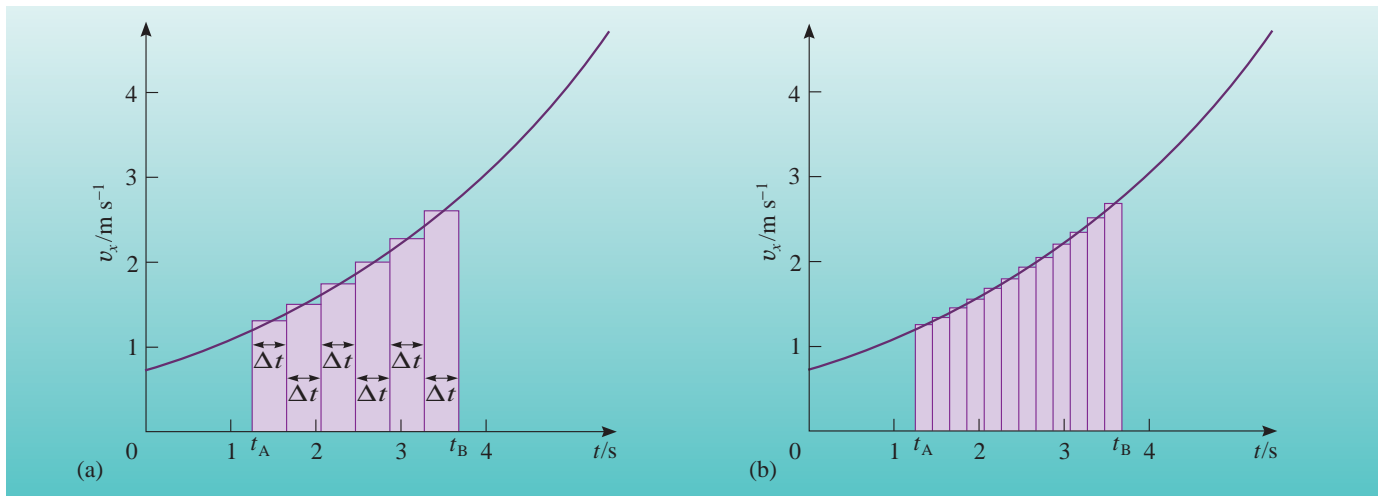


Figure 4.6 A graph of speed versus time, with the interval between $t = t_A$ and $t = t_B$ divided into (a) six time intervals of uniform width Δt and (b) into twelve finer time intervals.

Likewise, the distance Δs_2 travelled during the time interval Δt_2 from $t = t_A + \Delta t_1$ to $t = t_A + \Delta t_1 + \Delta t_2$ is just $\Delta s_2 \approx v_2 \Delta t_2$. To calculate the total distance travelled between $t = t_A$ and $t = t_B$, we simply add the distance travelled during each brief time interval; Figure 4.6a divides this into six intervals,

so

$$\begin{aligned}\Delta s &= \Delta s_1 + \Delta s_2 + \Delta s_3 + \Delta s_4 + \Delta s_5 + \Delta s_6 \\ &\approx v_1 \Delta t_1 + v_2 \Delta t_2 + v_3 \Delta t_3 + v_4 \Delta t_4 + v_5 \Delta t_5 + v_6 \Delta t_6\end{aligned}$$

Using *summation notation*, we could write this as

$$\Delta s \approx \sum_{i=1}^6 v_i \Delta t_i \quad (4.22)$$

Note, however, that this is just the area under the curve, i.e. the area enclosed by (i) the velocity curve, (ii) the x -axis, and (iii) the vertical boundaries at $t = t_A$ and $t = t_B$. That is, the area under the speed versus time curve, calculated over the time interval from t_A to t_B , gives the distance travelled by the body over that time.

The calculation of the area given above is not perfect because the area $v \Delta t$ of each rectangle of height v and width Δt in Figure 4.6a is only an approximation to the area under the curve. It would be exact if the speed increased linearly with time, but, because the graph is curved, the approximation is imperfect. However, the approximation can be made better by choosing smaller time intervals, as in Figure 4.6b. In fact, as still smaller time intervals are chosen, the approximation becomes even better, and it becomes exact in the limit where Δt shrinks to zero. Just as for *differentiation*, where the notation is changed in the limit where intervals shrink to zero, we can replace the summation with a new form:

$$\Delta s \approx \sum_{i=1}^6 v_i \Delta t_i \text{ becomes } \Delta s = \int_{t_A}^{t_B} v(t) dt \quad (4.23)$$

where

$$\int_{t_A}^{t_B} v(t) dt = \lim_{\Delta t \rightarrow 0} \sum_i v_i \Delta t_i \quad (4.24)$$

where $v(t)$ indicates that the speed is a *function* of time, i.e. not constant, dt indicates that we are considering the limit of infinitesimal time intervals, and $\int_{t_A}^{t_B}$ indicates that we sum over the time interval from $t = t_A$ to $t = t_B$. This expression is referred to as ‘the *definite integral* of v with respect to t from t_A to t_B .’

4.1 | Integration of known functions

The next thing to note is that **integration**, the technique of forming *integrals*, can be performed directly on equations without needing to consider graphs. Furthermore, integration can be regarded *almost* as the inverse of differentiation; if the derivative of f with respect to t is f' , then the integral of f' with respect to t is $f + C$, where C is a constant. It is because of the need for the constant C that integration and differentiation are not *exact* inverses.

Finally note that we can distinguish between a *definite integral*, which is an integration between two specified limits (e.g. $t = t_A$ and $t = t_B$ as above), and an

Table 4.2 Some functions and their indefinite integrals.

Function	Integral
$f(t)$	$\int f(t) dt$
a	$at + C$
at	$\frac{1}{2}at^2 + C$
at^n ($n \neq -1$)	$\frac{at^{n+1}}{n+1} + C$
a/t	$a \log_e t + C$
$a \sin(\omega t)$	$-\frac{a \cos(\omega t)}{\omega} + C$
$a \cos(\omega t)$	$\frac{a \sin(\omega t)}{\omega} + C$
$a \exp kt$	$\frac{a \exp kt}{k} + C$

indefinite integral in which the *equation* of the integral is found without reference to specified limits. As an example, recall from Section 4.2 that the derivative of x^2 with respect to x is $2x$. The *indefinite* integral of $2x$ with respect to x is therefore $x^2 + C$. The *definite* integral of $2x$ with respect to x over the interval from $x = 2$ to $x = 6$, say, can be found by *evaluating* the indefinite integral at both limits, and subtracting the first from the second. That is,

$$\int_{x=2}^{x=6} 2x dx = [x^2 + C]_{x=2}^{x=6}$$

where the square brackets signify that the integral is to be evaluated between the two limits shown. Hence

$$\int_{x=2}^{x=6} 2x dx = (6^2 + C) - (2^2 + C) = 32$$

Some other indefinite integrals are listed in Table 4.2, where integrals are given with respect to some variable, t . The symbols ω , a , n and k are constants (that is, independent of t) and C is an unknown constant introduced in each case.

The equivalent of the sum rule for differentiation is the following simple expression, known as the sum rule for integration:

$$\int (u + v) dt = \int u dt + \int v dt \quad (4.25)$$

where u and v are both themselves functions of t in this case. The equivalent of the product rule for differentiation will be discussed in Section 4.13. The following examples illustrate some of the rules for integration in Table 4.2 using examples from astrophysics and cosmology.

Essential skill:

Using integration to solve astrophysical problems

Worked Example 4.11

The spectral flux density of a quasar in a particular region of its spectrum can be represented by the expression $F_\nu = K\nu^{-0.7}$, where K is a constant. Evaluate the indefinite integral $\int F_\nu d\nu$.

Solution

Using the rule from line 4 of Table 4.2.

$$\begin{aligned} \int K\nu^{-0.7} d\nu &= K \left[\frac{\nu^{-0.7+1}}{-0.7+1} \right] + C \\ &= (K\nu^{+0.3}/0.3) + C \end{aligned}$$

Essential skill:

Using integration to solve astrophysical problems

Worked Example 4.12

The spectral flux density of the same quasar in a different region of its spectrum can be represented by the expression $F_\nu = K\nu^{-1.0}$, where K is a constant. Evaluate the definite integral $\int_{\nu_1}^{\nu_2} F_\nu d\nu$ which represents the total power received per unit area in the spectral range ν_1 to ν_2 .

Solution

Since the power to which ν is raised is -1.0 , this time we need the rule from the fifth row of Table 4.2.

$$\begin{aligned}\int_{\nu_1}^{\nu_2} K\nu^{-1.0} d\nu &= K [\log_e \nu + C]_{\nu_1}^{\nu_2} \\ &= K(\log_e \nu_2 + C - \log_e \nu_1 - C) \\ &= K \log_e \left(\frac{\nu_2}{\nu_1} \right)\end{aligned}$$

Exercise 4.8 Evaluate the following indefinite integrals:

$$(a) \int \frac{GMm}{r^2} dr \quad (b) \int \left(b \exp x + \frac{1}{x} \right) dx$$



4.12 Integration by substitution

As with differentiation, the key to carrying out integration of a function is to try to reduce it to something that can be integrated by following one of the standard rules in Table 4.2. Just as it is sometimes necessary to substitute one variable for another in order to carry out a differentiation, it is also sometimes necessary to do this to carry out an integration. The technique of integration by substitution is also referred to as integration by change of variable. As before, some examples should make the process clear. The first example simply illustrates the technique, whilst the second one applies the technique to an astrophysical problem.

Worked Example 4.13

Evaluate the indefinite integral

$$\int \frac{x}{(x^2 + 1)^5} dx$$

Solution

We make the substitution $z = x^2 + 1$, and note that $dz/dx = 2x$. In order to carry out the integration with respect to this new variable, z , we must eliminate all occurrences of x in the original integral, and write things in terms of the new variable z only. We therefore write the bottom line of the integral as z^5 and replace dx by $dz/2x$, giving

$$\int \frac{x}{z^5} \frac{dz}{2x} = \int \frac{1}{2z^5} dz$$

Essential skill:

Using the technique of integration by substitution

Luckily the remaining x s cancelled out leaving an expression we can integrate using line 4 from Table 4.2, so

$$\int \frac{1}{2z^5} dz = \frac{1}{2} \times \frac{1}{-4z^4} + C = C - \frac{1}{8z^4}$$

Then reversing the original substitution gives

$$\int \frac{x}{(x^2 + 1)^5} dx = C - \frac{1}{8(x^2 + 1)^4}$$

Notice that, once again, we treated ‘ dz/dx ’ like a normal fraction when we replaced dx by $dz/2x$.

Essential skill:
Using the technique of
integration by substitution

Worked Example 4.14

A star essentially forms by the self-gravitational contraction of a cloud of gas. In order to estimate the timescale for this to happen it is useful to evaluate the so-called free-fall time t_{ff} for a spherical shell of initial radius r_0 enclosing mass m_0 . This is given by the integral

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \int_{x=0}^{x=1} \left[\frac{x}{1-x} \right]^{1/2} dx$$

where $x = r/r_0$, i.e. the radius as a fraction of the initial radius. Evaluate the free-fall time by using the substitution $x = \sin^2 \theta$. You will also need the trigonometric relationships: $\sin^2 \theta + \cos^2 \theta = 1$ and $2 \sin^2 \theta = 1 - \cos(2\theta)$.

Solution

As suggested, we make the substitution $x = \sin^2 \theta$, so the first thing to calculate is $dx/d\theta$ in order to substitute for dx in the original integral.

A simple way to calculate this is to make a further substitution, $\sin \theta = z$, so we now have $x = z^2$. Now, $dx/dz = 2z$ and $dz/d\theta = \cos \theta$, so by the chain rule, $dx/d\theta = dx/dz \times dz/d\theta = 2z \times \cos \theta = 2 \sin \theta \cos \theta$.

Substituting for x and dx , the original integral now becomes

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \int_{x=0}^{x=1} \left[\frac{\sin^2 \theta}{1 - \sin^2 \theta} \right]^{1/2} \times 2 \sin \theta \cos \theta d\theta$$

From Equation 1.37, $1 - \sin^2 \theta = \cos^2 \theta$, so we have

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \int_{x=0}^{x=1} \left[\frac{\sin^2 \theta}{\cos^2 \theta} \right]^{1/2} \times 2 \sin \theta \cos \theta d\theta$$

Simplifying the term in square brackets under the integral

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \int_{x=0}^{x=1} \frac{\sin \theta}{\cos \theta} \times 2 \sin \theta \cos \theta d\theta$$

and cancelling the $\cos \theta$ terms

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \int_{x=0}^{x=1} 2 \sin^2 \theta \, d\theta$$

In order to integrate this function, we now use the suggested trigonometric relationship $2 \sin^2 \theta = 1 - \cos(2\theta)$,

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \int_{x=0}^{x=1} [1 - \cos(2\theta)] \, d\theta$$

and finally we have a function we can integrate using the sum rule:

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \left[\theta + \frac{\sin(2\theta)}{2} \right]_{x=0}^{x=1}$$

To evaluate the integral between the limits, we need to convert the limits into their equivalent values of θ . Since $x = 0 \sin^2 \theta$, when $x = 0$, then $\theta = 0$ and when $x = 1$, then $\theta = \pi/2$ radians. So the equation becomes

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \left[\theta + \frac{\sin(2\theta)}{2} \right]_{\theta=0}^{\theta=\pi/2}$$

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \left[\frac{\pi}{2} + \frac{\sin \pi}{2} - 0 - \frac{\sin 0}{2} \right]$$

$\sin \pi$ and $\sin 0$ are both equal to zero, so the final answer for the free-fall time is

$$t_{\text{ff}} = \left[\frac{r_0^3}{2Gm_0} \right]^{1/2} \times \frac{\pi}{2} = \left(\frac{\pi^2 r_0^3}{8Gm_0} \right)^{1/2}$$

This was a laborious integration to carry out, but the (many!) individual steps are each straightforward. You will not be asked to carry out such lengthy integrals in assignments or examinations, but you *should* be able to follow the reasoning of an example such as this, and appreciate how the final result is arrived at.

The obvious question you may be asking is, how do I know what substitution is appropriate in a given case? Well, there is no simple answer to this question. In the examples you will meet, the substitution will either be given to you, or will be reasonably obvious, as in the two examples above.

Exercise 4.9 Evaluate the definite integral

$$\int_{x=0}^{x=a} \frac{1}{\sqrt{a^2 - x^2}} \, dx$$

by using the substitution $x = a \sin \theta$. (*Hint:* You will need to use the identity

$\sin^2 \theta + \cos^2 \theta = 1$ in order to simplify things during this calculation.)

4.13 Integration by parts

As you saw earlier, the product rule for differentiation is

$$\frac{d(uv)}{dt} = u \frac{dv}{dt} + v \frac{du}{dt}$$

where u and v are both functions of t in this case. If we now *integrate* the above expression with respect to t we obtain

$$\int \frac{d(uv)}{dt} dt = \int u \frac{dv}{dt} dt + \int v \frac{du}{dt} dt$$

remembering that $dt/dt = 1$, this simplifies to

$$uv = \int u dv + \int v du$$

Finally, rearranging this result we obtain

$$\int u dv = uv - \int v du \quad (4.26)$$

This expression describes a technique referred to as **integration by parts**. It allows us to integrate quite complicated products of functions. Success with the formula relies on choosing u and dv such that $\int v du$ is easier to calculate than $\int u dv$. Once again, an example should help to make things clear.

Worked Example 4.15

The so-called mean free path travelled by an atomic nucleus before it undergoes a nuclear reaction in the core of a star is given by

$$L = \int_{x=0}^{x=\infty} \sigma n x \exp(-\sigma n x) dx$$

where σ and n are constants representing the reaction cross-section and the number of target particles per unit volume respectively. Use the technique of integration by parts to evaluate this definite integral.

Solution

We first make the substitutions, $u = \sigma n x$ and $dv = \exp(-\sigma n x) dx$.

So we are now trying to evaluate $\int u dv$ and we know from Equation 4.26 that this is equal to $uv - \int v du$. In order to work this out, we need to calculate v and du .

Clearly, $du/dx = \sigma n$, so we can write $du = \sigma n dx$.

Also,

$$v = \int dv = \int \exp(-\sigma n x) dx = \frac{\exp(-\sigma n x)}{-\sigma n}$$

Essential skill:
Carrying out integration by parts

We neglect the integration constant here, since we're carrying out a definite integral which means it would cancel out eventually.

To summarize, the four substitutions we now have are:

$$\begin{aligned} u &= \sigma n x \\ du &= \sigma n dx \\ v &= \frac{\exp(-\sigma n x)}{-\sigma n} \\ \text{and } dv &= \exp(-\sigma n x) dx \end{aligned}$$

So now we can write

$$uv - \int v du = \left[\frac{\sigma n x \exp(-\sigma n x)}{-\sigma n} \right]_{x=0}^{x=\infty} - \int_{x=0}^{x=\infty} \frac{\sigma n \exp(-\sigma n x)}{-\sigma n} dx$$

The various σn terms cancel, leaving

$$L = [-x \exp(-\sigma n x)]_{x=0}^{x=\infty} + \int_{x=0}^{x=\infty} \exp(-\sigma n x) dx$$

We now integrate the right-hand term to give

$$L = [-x \exp(-\sigma n x)]_{x=0}^{x=\infty} + \left[\frac{\exp(-\sigma n x)}{-\sigma n} \right]_{x=0}^{x=\infty}$$

At this point we note that $\exp(-\infty) = 0$ and $\exp 0 = 1$, so we have

$$L = [0 - 0] + \left[\frac{0}{-\sigma n} - \frac{1}{-\sigma n} \right] = \frac{1}{\sigma n}$$

So the mean free path L is just the reciprocal of the reaction cross-section times the number of target particles per unit volume. This makes sense – if either of these quantities increase, the mean distance a nucleus travels before reacting is reduced.

Clearly, the choice of u and dv worked well. Had we chosen to put $u = \exp(-\sigma n x)$ and $dv = \sigma n x dx$, the integration would have been much more complicated than the one we started with.

Exercise 4.10 Use the technique of integration by parts to evaluate the indefinite integral $\int \log_e x dx$. You should use the substitutions $u = \log_e x$ and $dv = dx$.

4.14 Multiple integrals

It is frequently the case when solving physical problems that an integral has to be carried out in more than one dimension.

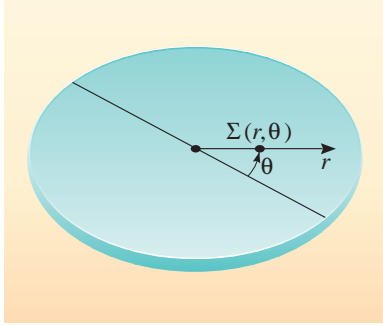


Figure 4.7 The surface density $\Sigma(r, \theta)$ of an accretion disc around a compact star may vary in both a radial coordinate r and an azimuthal coordinate θ .

Essential skill:

Carrying out a multiple integral

The first such example to consider is that of a **surface integral**. For instance, an accretion disc around a compact star will have a surface density, denoted by the Greek upper-case letter Σ (*sigma*), which may be a function of both distance from the centre (r) and the azimuthal angle around the disc (θ). This is effectively a system of plane polar coordinates. The surface density at a particular position (r, θ) , as illustrated in Figure 4.7, is defined as the total mass per unit area within a vertical column through the disc at that position. The *total* mass within the disc is therefore the integral of the surface density with respect to both r and θ , which may be written

$$M_{\text{disc}} = \int_{r=0}^{r=R} \int_{\theta=0}^{\theta=2\pi} \Sigma(r, \theta) r \, dr \, d\theta \quad (4.27)$$

Notice the extra factor of r introduced by converting from Cartesian coordinates (x, y) to plane polar coordinates (r, θ) .

The integrals are carried out in turn for one variable then the other, with ϕ ranging from 0 to 2π and r ranging from 0 to R , the outer radius of the disc. In practice, Σ may be a function of r only, so the integral over the angle ϕ is straightforward, and if Σ is a constant, the integration is simpler still, as the following example illustrates.

Worked Example 4.16

If the surface density of an accretion disc is constant throughout with a value Σ , what is the total mass of a disc of radius R ?

Solution

The mass may be found by first integrating over the angle ϕ to get

$$\begin{aligned} M_{\text{disc}} &= \int_{r=0}^{r=R} \int_{\theta=0}^{\theta=2\pi} \Sigma r \, dr \, d\theta \\ &= \int_{r=0}^{r=R} [\Sigma \theta r]_{\theta=0}^{\theta=2\pi} \, dr \\ &= \int_{r=0}^{r=R} \Sigma 2\pi r \, dr \end{aligned}$$

then by integrating over r

$$M_{\text{disc}} = \left[\frac{\Sigma 2\pi r^2}{2} \right]_{r=0}^{r=R} = \Sigma \pi R^2$$

That is to say, the mass of the disc is just the surface density multiplied by the area of the disc, as expected for a constant surface density.

Another commonly encountered situation is that of a **volume integral**. For instance, the density inside a star, denoted by the symbol ρ , may be a function of the x, y, z Cartesian coordinates within the star, i.e. $\rho(x, y, z)$, or equivalently it may be expressed more conveniently in terms of spherical coordinates with respect to the centre of the star, i.e. $\rho(r, \phi, \theta)$. As in the example above, the *total* mass of the star is then the integral of the density over all three coordinates,

written as

$$M_{\text{star}} = \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} \int_{\theta=0}^{\theta=\pi} \rho(r, \phi, \theta) r^2 \sin \theta \, dr \, d\phi \, d\theta \quad (4.28)$$

Notice the extra factor of $r^2 \sin \theta$ introduced by converting from Cartesian coordinates (x, y, z) to spherical coordinates (r, ϕ, θ) .

Worked Example 4.17

Use Equation 4.28 to determine the mass of a spherical gas cloud of radius R and of uniform density ρ_0 .

Solution

The mass of the cloud is given by

$$M = \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} \int_{\theta=0}^{\theta=\pi} \rho_0 r^2 \sin \theta \, dr \, d\phi \, d\theta$$

First integrating with respect to the angle θ :

$$M = \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} [-\rho_0 r^2 \cos \theta]_{\theta=0}^{\theta=\pi} \, dr \, d\phi$$

$$M = \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} [(-\rho_0 r^2 \times -1) - (-\rho_0 r^2 \times 1)] \, dr \, d\phi$$

$$M = \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} 2\rho_0 r^2 \, dr \, d\phi$$

then integrating with respect to the angle ϕ :

$$M = \int_{r=0}^{r=R} [2\rho_0 r^2 \phi]_{\phi=0}^{\phi=2\pi} \, dr$$

$$M = \int_{r=0}^{r=R} [(2\rho_0 r^2 \times 2\pi) - (2\rho_0 r^2 \times 0)] \, dr$$

$$M = \int_{r=0}^{r=R} 4\pi\rho_0 r^2 \, dr$$

and finally integrating with respect to r :

$$M = \left[\frac{4\pi\rho_0 r^3}{3} \right]_{r=0}^{r=R} = \frac{4}{3}\pi\rho_0 R^3$$

which is as expected, given that the volume of a sphere is $4\pi R^3/3$.

Essential skill:

Carrying out a multiple integral

Exercise 4.11 Suppose that the density within a star is given by $\rho(r, \phi, \theta) = R^2 \rho_0 / r^2$, where ρ_0 is a constant (the density at $r = R$) and R is the radius of the star. Calculate the total mass of the star.

Summary of Chapter 4

1. Differentiation is a means of finding how one quantity changes as a result of changes in another. The function dy/dt represents the rate of change of y with t , and is known mathematically as the derivative of y with respect to t .
2. Alternative ways of writing dy/dt include \dot{y} and y' . The second derivative of y with respect to t may be written d^2y/dt^2 or \ddot{y} or y'' .
3. Two particularly useful rules for differentiating functions are the sum rule and product rule, namely

$$\frac{d(u+v)}{dt} = \frac{du}{dt} + \frac{dv}{dt}$$

$$\frac{d(uv)}{dt} = u \frac{dv}{dt} + v \frac{du}{dt}$$

where u and v are themselves functions of t in this case.

4. The chain rule for differentiation states that

$$dy/dx = dy/du \times du/dx$$

5. There are many standard derivatives, but two of the most useful are

$$\begin{array}{lll} \text{if } y = at^n & \text{then } dy/dt = nat^{n-1} \\ \text{if } y = a \exp(kt) & \text{then } dy/dt = ak \exp(kt) \end{array}$$

6. Logarithmic differentiation involves taking the natural logarithm of an equation before evaluating the derivative. The derivative of a natural logarithm with respect to time is often written as

$$\frac{d(\log_e x)}{dt} = \frac{\dot{x}}{x}$$

7. The Maclaurin series expansion of a function $f(x)$ about the point $x = 0$ may be written

$$f(x) = f(0) + xf'(0) + x^2 f''(0)/2! + x^3 f'''(0)/3! + \dots \text{ etc.}$$

8. A partial derivative indicates the rate of change of a function with respect to only *one* of the variables on which it depends. If h is a function of both x and y then $\frac{\partial h(x,y)}{\partial x}$ represents the rate of change of h with respect to x only, with y held constant.

9. The vector differential operator nabla is defined as

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

and the gradient of a scalar field T is then given by

$$\text{grad } T = \nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z} \right) = \text{a vector field}$$

The divergence of a vector field \mathbf{A} with components A_x, A_y, A_z at each point in space is given by

$$\text{div } \mathbf{A} = \nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} = \text{a scalar field}$$

The curl of this vector field \mathbf{A} is given by

$$\begin{aligned}\text{curl } \mathbf{A} &= \nabla \times \mathbf{A} \\ &= \left(\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \right), \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \right), \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) \right) \\ &= \text{a vector field}\end{aligned}$$

10. The first-order differential equation $\frac{dy}{dt} = -ky$ has a solution $y = a \exp(-kt)$, where a is an arbitrary constant determined by the boundary conditions.
11. The second-order differential equation $m \frac{d^2x}{dt^2} = -kx$ has a solution $x = A \sin(\omega t + \phi)$, where $\omega = \sqrt{k/m}$, and both A and ϕ are arbitrary constants determined by the boundary conditions.
12. The expression $\int_{t_A}^{t_B} v(t) dt$ is referred to as the definite integral of v with respect to t from t_A to t_B . It can be envisaged as the area under a graph of v against t between the two limits specified.
13. The evaluation of indefinite integrals (no limits specified) always involves the introduction of a constant of integration.
14. There are many standard integrals, but two of the most useful are:

$$\text{if } y = at^n \text{ then } \int y dt = \frac{at^{n+1}}{n+1} + C \quad (\text{for } n \neq -1)$$

$$\text{if } y = a \exp(kt) \text{ then } \int y dt = \frac{a \exp(kt)}{k} + C$$

15. The sum rule for integration is $\int (u + v) dt = \int u dt + \int v dt$, where both u and v are functions of t in this case.
16. Integration by substitution (or change of variable) is a technique that allows difficult integrals to be reduced to simpler ones.
17. Integration by parts is described by the expression $\int u dv = uv - \int v du$; it relies on choosing u and dv such that $\int v du$ is easier to calculate than $\int u dv$.
18. Surface and volume integrals are carried out by integrating a function with respect to each coordinate in turn.

This page is intentionally left blank to ensure that subsequent chapters begin on an odd-numbered page.

Chapter 5 Physics

Introduction

This chapter will allow you to revise your knowledge of physics. If you have recently completed a Level 2 physics module (such as S217), most of this chapter will probably be familiar to you.

It is particularly true of this chapter that we simply *present* results to you rather than deriving them from first principles. As with the rest of this document, you should therefore think of this chapter as a resource which brings together principles of physics which may be referred to in subsequent parts of the module.

5.1 Describing motion

A description of the motion of bodies lies at the heart of much of physics and astrophysics, and provides a sensible place to start consolidating your knowledge of the physics required to study Level 3 astrophysics and cosmology.

5.1.1 Motion in one dimension

The movement of a particle along a line can be described graphically by plotting values of the particle's **position** x , against the corresponding times t , to produce a position–time graph (Figure 5.1a). Alternatively, by choosing an appropriate reference position x_{ref} and defining the **displacement** from that point by $s_x = x - x_{\text{ref}}$, the motion may be described by means of a displacement–time graph (Figure 5.1b).

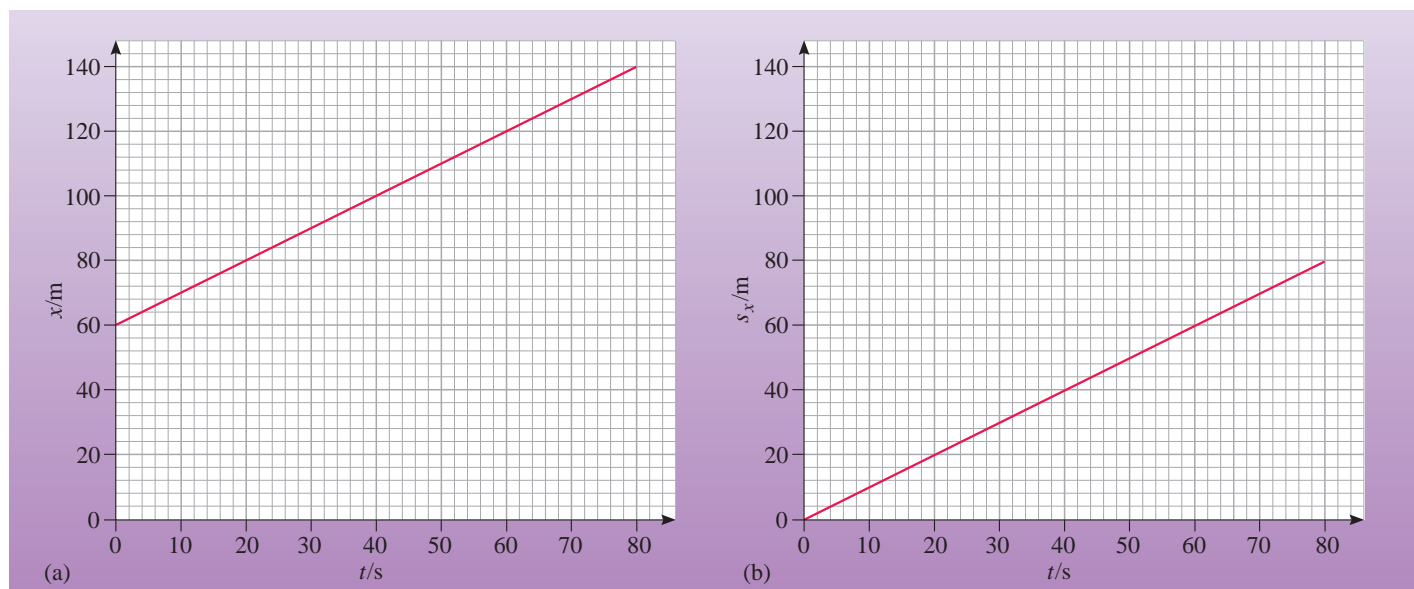


Figure 5.1 (a) A straight line position–time graph. (b) A straight line displacement–time graph.

Uniform motion along a line is characterized by a straight line position–time graph that may be described by the equation

$$x = v_x t + x_0$$

where v_x and x_0 are constants. Physically, v_x represents the particle's **velocity**, the rate of change of its position with respect to time, and is determined by the gradient of the position–time graph

$$v_x = \frac{\Delta x}{\Delta t} = \frac{x_2 - x_1}{t_2 - t_1}$$

x_0 represents the particle's initial position, its position at $t = 0$, and is determined by the intercept of the position–time graph, the value of x at which the plotted line crosses the axis labelled x , provided that axis has been drawn through $t = 0$.

Non-uniform motion along a line is characterized by a position–time graph that is not a straight line. In such circumstances the rate of change of position with respect to time may vary from moment to moment and defines the instantaneous velocity. Its value at any particular time is determined by the gradient of the tangent to the position–time graph at that time (Figure 5.2a).

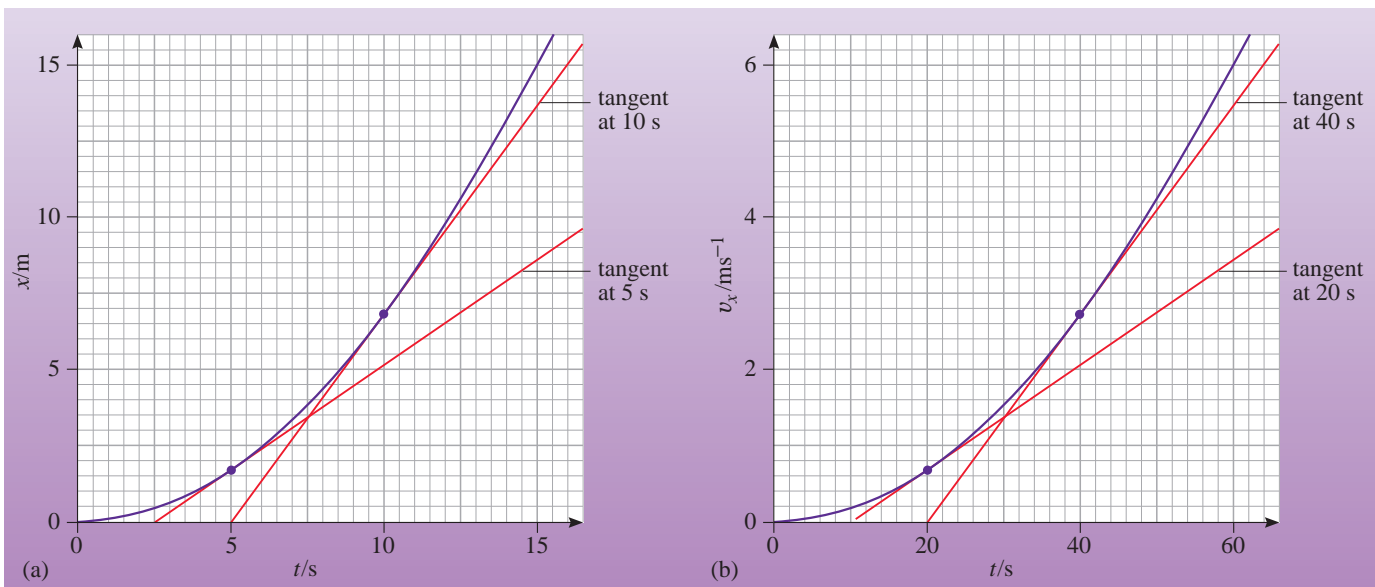


Figure 5.2 (a) A position–time graph for non-uniform motion. The instantaneous velocity at a particular time is determined by the gradient of the tangent to the graph at that time. (b) A velocity–time graph for non-uniform motion. The instantaneous acceleration at a particular time is determined by the gradient of the tangent to the graph at that time.

More generally, if the position of a particle varies with time in the way described by the function $x(t)$, then the way in which the (instantaneous) velocity varies with time will be described by the associated derivative

$$v_x(t) = \frac{dx(t)}{dt} \quad (5.1)$$

The instantaneous **acceleration** is the rate of change of the instantaneous velocity with respect to time. Its value at any time is determined by the gradient of the

tangent to the velocity–time graph at that time (Figure 5.2b). More generally, the way in which the (instantaneous) acceleration varies with time will be described by the derivative of the function that describes the instantaneous velocity, or, equivalently, the second derivative of the function that describes the position:

$$a_x(t) = \frac{dv_x(t)}{dt} = \frac{d^2x(t)}{dt^2} \quad (5.2)$$

The area under a velocity–time graph, between specified values of time, represents the change in position of the particle during that interval.

Uniformly accelerated motion is a special form of non-uniform motion characterized by a constant value of the acceleration ($a_x = \text{constant}$). In such circumstances the most widely used equations describing uniformly accelerated motion are

$$s_x = u_x t + \frac{1}{2} a_x t^2 \quad (5.3)$$

$$v_x = u_x + a_x t \quad (5.4)$$

$$v_x^2 = u_x^2 + 2a_x s_x \quad (5.5)$$

$$s_x = \left(\frac{v_x + u_x}{2} \right) t \quad (5.6)$$

Position, x , displacement, s_x , velocity, v_x , and acceleration, a_x , are all quantities that may be positive or negative, depending on the associated direction. The magnitude of each of these quantities is a positive quantity that is devoid of directional information. The magnitude of the displacement of one point from another, $s = |s_x|$, represents the **distance** between those two points, while the magnitude of a particle's velocity, $v = |v_x|$, represents the **speed** of the particle.

Exercise 5.1 A particle accelerates from rest at a rate of 5.0 m s^{-2} . (a) What distance has it covered after 10 s? (b) How fast is it travelling after this time? ■

5.1.2 Motion in two or three dimensions

In describing motion in one dimension, although we wrote everything in terms of scalar quantities, we were actually using vector components in the x -direction. (Remember, each component of a vector is itself a scalar.) In turning to describe motion in two or three dimensions, the simple extrapolation is to now use the full vector descriptions of quantities such as displacement, velocity and acceleration. (Refer back to Section 1.13 if necessary for a reminder about vectors.)

The displacement vector $\mathbf{s} = (s_x, s_y, s_z)$ from point P_1 with position vector $\mathbf{r}_1 = (x_1, y_1, z_1)$ to point P_2 with position vector $\mathbf{r}_2 = (x_2, y_2, z_2)$ is given by

$$\mathbf{s} = \mathbf{r}_2 - \mathbf{r}_1 = (x_2 - x_1, y_2 - y_1, z_2 - z_1) = (s_x, s_y, s_z) \quad (5.7)$$

The magnitude of this vector, $s = \sqrt{s_x^2 + s_y^2 + s_z^2}$, is the distance from P_1 to P_2 .

The velocity $\mathbf{v} = (v_x, v_y, v_z)$ of a particle is determined by the rate of change of the particle's position:

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right) \quad (5.8)$$

This vector's magnitude, $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$, is the speed v of the particle.

The acceleration $\mathbf{a} = (a_x, a_y, a_z)$ of a particle is determined by the rate of change of the particle's velocity:

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \left(\frac{dv_x}{dt}, \frac{dv_y}{dt}, \frac{dv_z}{dt} \right) \quad (5.9)$$

In two (or three) dimensions the motion of a particle may be regarded as the sum of two (or three) separate motions, mutually at right angles, that are independent apart from their common duration.

5.1.3 Periodic motion

A study of periodic motion is particularly useful in a topic such as astrophysics, where bodies are often in orbit around other bodies: planets around stars, stars around each other, and groups of stars around the centre of galaxies, for instance. By definition, periodic motion is repetitive, and the simplest such motion to consider is that of motion in a circle (Figure 5.3).

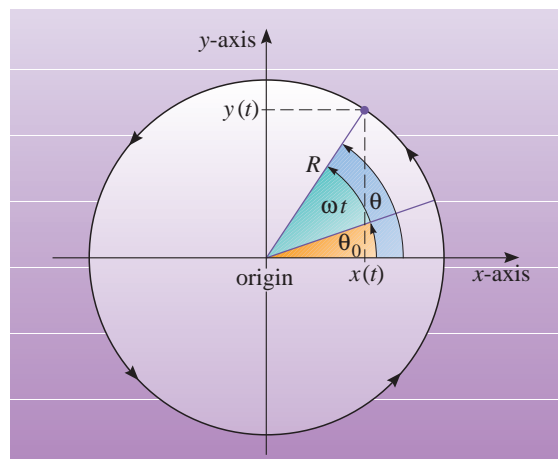


Figure 5.3 Uniform circular motion around a circle of radius R , centred on the origin, in the anticlockwise sense, with angular speed ω . The initial value (at time $t = 0$ s) of the angular coordinate θ is indicated by θ_0 .

The symbol ω (*omega*) is generally used to represent the (constant) **angular speed** of the motion, defined by $\omega = |d\theta/dt|$, where θ is the angle between the position vector of the particle and the x -axis. A particle in uniform circular motion completes one revolution (2π radians) in one **period** (usually represented by the symbol P or T). So in this case the angular speed is simply

$$\omega = 2\pi/P \quad (5.10)$$

The instantaneous velocity is tangential to the circle and has magnitude $v = r\omega$. The **centripetal acceleration** is directed towards the centre of the circle and has magnitude

$$a = v\omega = r\omega^2 = v^2/r \quad (5.11)$$

Exercise 5.2 A neutron star is in a circular orbit around its companion star, completing one orbit every 10 days. If the distance of the neutron star from the centre of the companion star is 3.7×10^{10} m, (a) what is the magnitude of the instantaneous velocity of the neutron star and (b) what is the magnitude of its centripetal acceleration?

More generally, astronomical bodies orbit each other following paths which are ellipses (Figure 5.4). An ellipse of semimajor axis a and semiminor axis b may be described by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Such an ellipse has eccentricity $e = \frac{1}{a}\sqrt{a^2 - b^2}$, where $0 \leq e < 1$, and contains two foci, located at the points $(ae, 0)$ and $(-ae, 0)$. The sum of the two distances from any point on the ellipse to the two foci is a constant, equal to $2a$.

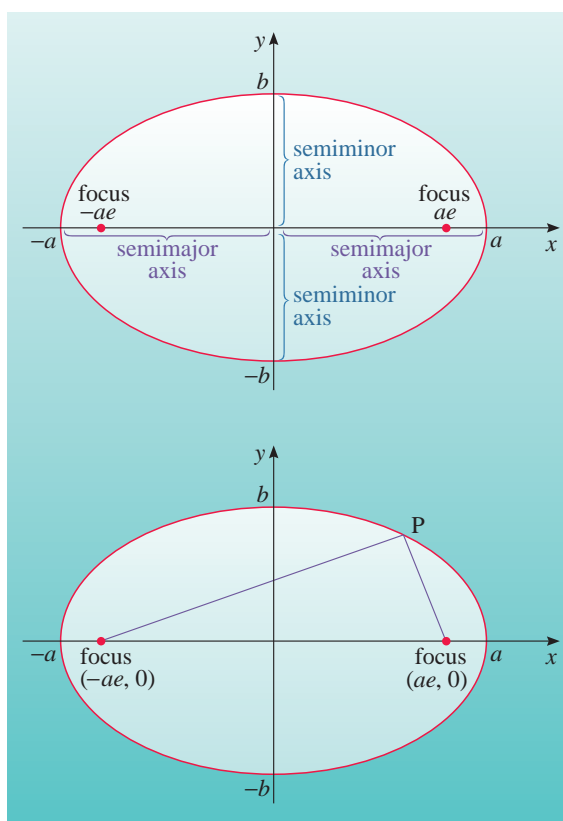


Figure 5.4 The properties of an ellipse.

5.2 Newton's laws

Newton's three laws of motion lie at the heart of predicting how bodies will move. In addition, his law of universal gravitation can be used to account for many of the types of motion that are encountered in astrophysics and cosmology. In this section we look at each of these laws in turn.

5.2.1 Newton's laws of motion

According to Newton's first law of motion:

A body remains at rest or in a state of uniform motion unless it is acted on by an unbalanced force.

The law therefore introduces **force** as a quantity that changes the motion of a body by causing it to accelerate.

The description of a body's motion depends on the frame of reference from which the motion is observed. (A frame of reference is a system for assigning position coordinates and times to events.) An **inertial frame** is a frame of reference in which Newton's first law holds true; it is a frame of reference that is not itself accelerating. Any frame that moves with constant velocity relative to an inertial frame, while maintaining a fixed orientation, will also be an inertial frame.

According to Newton's second law of motion: an unbalanced force acting on a body of fixed mass will cause that body to accelerate in the direction of the unbalanced force. The magnitude of the force is equal to the product of the mass and the magnitude of the acceleration. The law therefore enables force to be quantified and is usually expressed by the vector equation

$$\mathbf{F} = m\mathbf{a} \quad (5.12)$$

Force is measured in the SI unit of newtons where 1 newton (1 N) is equal to 1 kg m s^{-2} .

According to Newton's third law of motion: if body A exerts a force on body B, then body B exerts a force on body A. These two forces are equal in magnitude, but point in opposite directions. The law therefore indicates that:

To every *action* there is an oppositely directed *reaction* of equal magnitude.

Observers who attempt to apply Newton's laws in a non-inertial frame (i.e. one which is itself accelerating) will observe phenomena that indicate the existence of fictitious forces, such as centrifugal force and Coriolis force. These phenomena are real but the fictitious forces are not; they appear because of the acceleration of the observer's frame of reference relative to an inertial frame. By contrast, **centripetal force** is a real force in the sense that it arises in an inertial frame. From Equation 5.11:

$$F_{\text{cent}} = mr\omega^2 = mv^2/r \quad (5.13)$$

- A star of mass $1.5 \times 10^{30} \text{ kg}$ experiences a centripetal acceleration of magnitude 0.5 m s^{-2} . What must be the magnitude of the centripetal force acting upon it?
- From Newton's second law, the magnitude of the centripetal force is

$$F = ma = (1.5 \times 10^{30} \text{ kg}) \times (0.5 \text{ m s}^{-2}) = 7.5 \times 10^{29} \text{ N}$$

5.2.2 Newton's law of gravitation

According to **Newton's law of gravitation** (or as it is sometimes known Newton's law of universal gravitation): every particle of matter attracts every other particle of matter with a gravitational force, whose magnitude is directly proportional to the product of the masses of the particles, and inversely proportional to the square of the distance between them. The law therefore implies that the force on a particle of mass m_2 with position vector \mathbf{r} , due to a particle of mass m_1 at the origin will be

$$\mathbf{F}_{21} = -\frac{Gm_1m_2}{r^2}\hat{\mathbf{r}} \quad (5.14)$$

where G is Newton's (universal) gravitational constant ($6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$), and $\hat{\mathbf{r}} = \mathbf{r}/r$ is a unit vector pointing from the origin towards the particle of mass m_2 . The magnitude of the force on the first particle due to the second is equal to the magnitude of the force on the second particle due to the first, i.e.

$$|\mathbf{F}_{21}| = |\mathbf{F}_{12}| = \frac{Gm_1m_2}{r^2} \quad (5.15)$$

Another of Newton's contributions to this area is known as **Newton's theorem** which states that the gravitational effect of any spherically symmetric body, outside its own surface, is identical to that of a single particle, with the same mass as the body, located at the centre of the body. This means that stars and planets can, for the purposes of Newton's law of universal gravitation, be treated like point particles located at the centre of the original body.

Combining Newton's law of universal gravitation with ideas about periodic motion allows us to derive **Kepler's third law** for the relationship between the orbital periods of planets and their mean distance from the Sun, as shown by the following example.

Worked Example 5.1

A planet of mass m moves in a circular orbit of radius r with uniform angular speed ω about a star of mass M . What is the relationship between the planet's orbital period P and its orbital radius r ?

Solution

The magnitude of the centripetal acceleration of the planet is $a = r\omega^2$, so the planet must be subject to a centripetal force of magnitude $F = ma = mr\omega^2$. If this force is supplied by the gravitational attraction of the star, then by Newton's law of universal gravitation,

$$mr\omega^2 = \frac{GMm}{r^2}$$

For a planet with constant angular speed ω , the orbital period is $P = 2\pi/\omega$, so replacing ω in the equation above by $2\pi/P$ gives

$$\frac{(2\pi)^2mr}{P^2} = \frac{GMm}{r^2}$$

Essential skill:
Deriving Kepler's third law

Rearranging and cancelling the common terms, this yields

$$P^2 = \frac{(2\pi)^2 r^3}{GM} \quad (5.16)$$

This agrees with Kepler's third law which states that $P^2 \propto r^3$.

Although Kepler's laws were derived for the motion of planets around the Sun, they apply to any body moving under the gravitational force of another body. For instance they also apply to the motion of binary stars.

Since the period of an orbit is related to the angular speed of the orbiting body by $P = 2\pi/\omega$ (Equation 5.10), an alternative expression describing the motion of a body in a so-called Keplerian orbit is

$$\omega = \left(\frac{GM}{r^3} \right)^{1/2} \quad (5.17)$$

Furthermore, since the tangential speed is related to the angular speed by $v = r\omega$ (Equation 5.11), yet another expression describing such motion is

$$v = \left(\frac{GM}{r} \right)^{1/2} \quad (5.18)$$

5.3 Relativistic motion

The mostly intuitive ideas about describing motion embodied in Newton's laws and formulae, such as Equations 5.3 – 5.6, turn out *not* to be true when the speeds involved approach the speed of light. In that circumstance, those formulae turn out to be only approximations to the real situation. For a more accurate description of nature we must turn to the **theory of special relativity** derived by Albert Einstein and published in 1905. Einstein's theory is based on two postulates:

Einstein's postulates

Postulate I – The principle of relativity

The laws of physics can be written in the same form in all inertial frames.

Postulate II – The principle of the constancy of the speed of light

The speed of light in a vacuum has the same constant value in all inertial frames, $c = 3.00 \times 10^8 \text{ m s}^{-1}$.

These two simple postulates lead to many intriguing and counter-intuitive results. Amongst these are the fact that the duration of a time interval is a relative quantity. The rate at which a clock ticks depends on the frame of reference in which it is measured. This is often paraphrased as 'moving clocks run slow'. It may be expressed as

$$\Delta T = \frac{\Delta T_0}{\sqrt{1 - \frac{V^2}{c^2}}} \quad (5.19)$$

where ΔT is the time interval measured by an observer in one inertial frame of reference, and ΔT_0 is the time interval measured by an observer in another inertial frame of reference which is moving at speed V relative to the first.

Another result is that length is also a relative quantity. The length of a rod depends on the frame of reference in which it is measured. This is often paraphrased as ‘moving rods contract in the direction of motion’. It may be expressed as

$$L = L_0 \sqrt{1 - \frac{V^2}{c^2}} \quad (5.20)$$

where L is the length of the rod measured by an observer in one inertial frame of reference, and L_0 is the length measured by an observer in another inertial frame of reference which is moving at speed V relative to the first.

Both these results may be understood in terms of the **Lorentz transformation** which links the coordinates measured in two inertial frames of reference that are moving relative to each other.

Consider an inertial frame of reference B which is moving parallel to the x -axis of an inertial frame of reference A, with a speed V . These two frames of reference are said to be in ‘standard configuration’. The coordinates in space and time of a point in frame B (x', y', z', t') may be expressed in terms of the coordinates of the same point in frame A (x, y, z, t) as

$$x' = \frac{x - Vt}{\sqrt{1 - \frac{V^2}{c^2}}} \quad (5.21)$$

$$y' = y \quad (5.22)$$

$$z' = z \quad (5.23)$$

$$t' = \frac{t - Vx/c^2}{\sqrt{1 - \frac{V^2}{c^2}}} \quad (5.24)$$

The factor $1/\sqrt{1 - V^2/c^2}$ occurs so often in special relativity that it is often given the symbol γ (*gamma*) and is referred to as the Lorentz factor. Using this notation, the Lorentz transformations may be written

$$x' = \gamma(x - Vt) \quad (5.25)$$

$$y' = y \quad (5.26)$$

$$z' = z \quad (5.27)$$

$$t' = \gamma \left(t - \frac{Vx}{c^2} \right) \quad (5.28)$$

One of the most dramatic consequences of the relative nature of time is that the order in which two events occur can, in certain circumstances, depend on the frame of reference of the observer. However, the ‘cause’ of an event must always be observed to precede its ‘effect’. This causality will be preserved as long as the speed of light c is the maximum speed at which a signal can travel.

We can also write down the transformation between the velocity measured in one frame of reference and another. Using the same definition of inertial frames of reference A and B as above, it can be shown that

$$v'_x = \frac{v_x - V}{1 - \frac{Vv_x}{c^2}} \quad (5.29)$$

where v'_x is the x -component of the velocity measured in frame B, and v_x is the x -component of the velocity measured in frame A.

Exercise 5.3 An astronaut sitting in the observation tower of a space station sees two spaceships X and Y approaching her from opposite directions at high speed. She measures the approach speed of both X and Y to be $3c/4$. What is the speed of spaceship Y as measured by spaceship X? (*Hint:* The trick here lies in describing the situation in terms of standard frames of reference, as described above.)

5.4 Predicting motion

The concepts of work, energy, power and momentum are fundamental in many areas of physics, and are equally important when it comes to predicting how objects behave in an astrophysical or cosmological context.

5.4.1 Work, energy, power and momentum

The **energy** of a system is a measure of its capacity for doing work. The SI unit of work and of energy is the joule (J), where $1 \text{ J} = 1 \text{ kg m}^2 \text{ s}^{-2} = 1 \text{ N m}$. **Power** is defined as the rate at which work is done and energy transferred. The SI unit of power is the watt (W), where $1 \text{ W} = 1 \text{ J s}^{-1}$.

The **translational kinetic energy** of a body of mass m and speed v is

$$E_{\text{KE}} = \frac{1}{2}mv^2 \quad (5.30)$$

The **work** done on any body by a force is the energy transferred to or from that body by the force. When a non-zero resultant force acts on a body, the work done by that force is equal to the change in the body's translational kinetic energy:

$$W = \Delta E_{\text{KE}} = \frac{1}{2}mv^2 - \frac{1}{2}mu^2 \quad (5.31)$$

where v and u are the final and initial speeds of the body, respectively.

Exercise 5.4 The speed of a particle of mass 10^{-6} kg increases from 5 m s^{-1} to 10 m s^{-1} . How much work is done on the particle in this process?

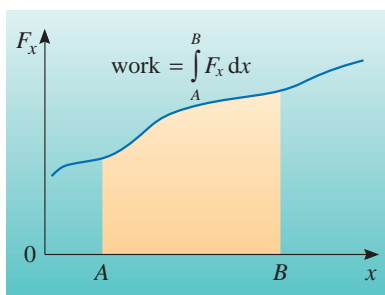


Figure 5.5 The work done by a force is equal to the area under a graph of F_x against x .

The work done by a force that varies in strength but always acts along the x -axis is defined by the following integral

$$W = \int_A^B F_x dx \quad (5.32)$$

which may be interpreted as the area under the graph of F_x against x between $x = A$ and $x = B$, as shown in Figure 5.5.

A **conservative force** is one where the total work done by the force is zero for any round trip, or, equivalently, where the work done by the force is independent of the path connecting the start- and end-points. Other forces, which do not satisfy this condition, are non-conservative forces.

A **potential energy** may be associated with each conservative force that acts on a body or between a system of bodies. The potential energy E_{POT} associated with any particular configuration is the work that would be done by the relevant conservative force in going from that configuration to an agreed 'reference' configuration that has been arbitrarily assigned zero potential energy. Because of the arbitrary nature of this reference configuration, only changes in potential energy are physically significant.

A particular example, important in astrophysics, is the **gravitational potential energy** of an object of mass m at a distance r from the centre of a body of mass M , which is given by

$$E_{\text{GR}} = -\frac{GmM}{r} \quad (5.33)$$

such that $E_{\text{GR}} = 0$ at $r = \infty$.

- What is the gravitational potential energy of an apple of mass 100 g at the surface of the Earth? (Assume that the mass and radius of the Earth are 5.97×10^{24} kg and 6.38×10^6 m respectively and that the gravitational constant is $G = 6.67 \times 10^{-11}$ N m² kg⁻².)
- $E_{\text{GR}} = -(6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}) \times (0.1 \text{ kg}) \times (5.97 \times 10^{24} \text{ kg}) / (6.38 \times 10^6 \text{ m}) = 6.24 \times 10^6 \text{ N m}$ or 6.24 MJ (megajoules).

Another consequence of Newton's law of gravity is the fact that every massive body will possess a certain **escape speed**. This is the minimum speed an object must acquire for it to escape completely from the gravitational influence of another body. If an object has translational kinetic energy $E_{\text{KE}} = \frac{1}{2}mv^2$ at a distance r from the centre of a body of mass M , then in order to escape, it must be the case that $E_{\text{KE}} + E_{\text{GR}} \geq 0$. The limiting case is given by the equality and defines the escape speed such that $\frac{1}{2}mv_{\text{esc}}^2 - GmM/r = 0$. Rearranging this expression, we obtain the escape speed as

$$v_{\text{esc}} = (2GM/r)^{1/2} \quad (5.34)$$

- What is the escape speed at the surface of an asteroid whose radius is 10 km and whose mass is 10^{15} kg? ($G = 6.7 \times 10^{-11}$ N m² kg⁻²).
- Using Equation 5.34

$$v_{\text{esc}} = (2 \times 6.7 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2} \times 10^{15} \text{ kg} / 10^4 \text{ m})^{1/2} = 3.7 \text{ m s}^{-1}$$

or about 4 m s⁻¹.

If the potential energy associated with a particular conservative force (such as gravity) is a function of the single variable x , then the only non-zero component of the force will be F_x , and its value at any point will be given by *minus* the gradient of the E_{POT} versus x graph at that point:

$$F_x = -\frac{dE_{\text{POT}}}{dx} \quad (5.35)$$

A generalization of this to more than one dimension makes use of the nabla operator introduced in Section 4.8. The vector \mathbf{F} is given by

$$\mathbf{F} = -\nabla E_{\text{POT}} \quad (5.36)$$

Exercise 5.5 Use Equations 5.33 and 5.35 to derive an expression for the magnitude of the force of gravity experienced by a body of mass m at a distance r from the centre of a body of mass M .

For systems in which only conservative forces act, the total mechanical energy is conserved. That is,

$$\Delta E_{\text{POT}} + \Delta E_{\text{KE}} = 0$$

This equation provides the means of linking the speed of a body to its position in such a system. The principle of conservation of mechanical energy can be extended to cover all forms of energy, thus leading to the **principle of conservation of energy**.

In classical Newtonian mechanics the momentum of a body of mass m and velocity \mathbf{v} is given by

$$\mathbf{p} = m\mathbf{v} \quad (5.37)$$

This leads to an alternative expression of Newton's second law of motion, namely that the total force acting on a body is equal to the rate of change of momentum of the body:

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} \quad (5.38)$$

This is a more general statement of Newton's second law of motion than $\mathbf{F} = m\mathbf{a}$, since the latter only applies to situations in which the mass is constant.

According to the **principle of conservation of momentum**, the total momentum of any isolated system is constant. The principle of conservation of momentum may be used in the solution of a variety of problems. It is commonly used in the analysis of collision problems, often in association with some aspect of energy conservation. Collisions in which kinetic energy is conserved are said to be **elastic**. Collisions in which kinetic energy is not conserved are said to be **inelastic**.

5.4.2 Relativistic mechanics

In high-energy collisions, where collision speeds approach that of light in a vacuum ($c = 3.00 \times 10^8 \text{ m s}^{-1}$), it is necessary to use the relativistic definitions of momentum and translational kinetic energy, namely

$$\mathbf{p} = \frac{m\mathbf{v}}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (5.39)$$

$$E_{\text{KE}} = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} - mc^2 \quad (5.40)$$

The total relativistic energy is the sum of the relativistic translational kinetic energy and the mass energy of a body:

$$E_{\text{TOT}} = E_{\text{KE}} + E_{\text{MASS}} \quad (5.41)$$

where the mass energy is given by perhaps the most famous equation in all of physics:

$$E_{\text{MASS}} = mc^2 \quad (5.42)$$

and the total relativistic energy is therefore

$$E_{\text{TOT}} = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (5.43)$$

In high-energy collisions particles may be created (implying that mass energy increases) but total energy must be conserved, so there must be a corresponding decrease in kinetic energy.

By combining Equations 5.39 – 5.43, it can be seen that a general relationship between total relativistic energy, momentum and mass energy is

$$E_{\text{TOT}}^2 = p^2 c^2 + m^2 c^4 \quad (5.44)$$

Exercise 5.6 At what speed must a particle travel if its relativistic translational kinetic energy is equal to its mass energy?



5.5 Rotational motion

Many of the ideas of translational motion discussed in earlier sections have analogues in rotational motion. The first such concept to consider is that of **torque**, or the turning effect of a force. Torque is a vector quantity, and the torque Γ (the upper case Greek letter *gamma*) about a point O due to a force F is

$$\Gamma = \mathbf{r} \times \mathbf{F} \quad (5.45)$$

where \mathbf{r} is the displacement vector from O to the point of application of the force (see Figure 5.6). The direction of the torque vector is perpendicular to the plane containing the force and displacement vectors, as given by the right-hand rule.

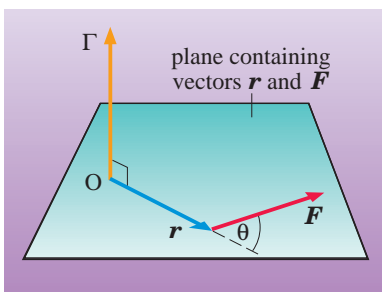


Figure 5.6 The magnitude of the torque Γ is equal to the product of the magnitude of the force F , the distance from the axis of rotation r , and the sine of the angle between these two vectors. The direction of Γ is perpendicular to the plane containing the other two vectors, given by the right-hand rule.

The rotational analogue of mass is **moment of inertia** I , which is defined about a given axis for a system of particles as

$$I = \sum_i m_i r_i^2 \quad (5.46)$$

where r_i is the perpendicular distance from the axis to the i th particle, and m_i is the mass of that particle. Note that the moment of inertia of a system depends on the axis about which it is determined. The moment of inertia I about a given axis for a rigid body has a similar significance, though its evaluation usually involves a definite integral rather than a sum. Some results are given in Figure 5.7.

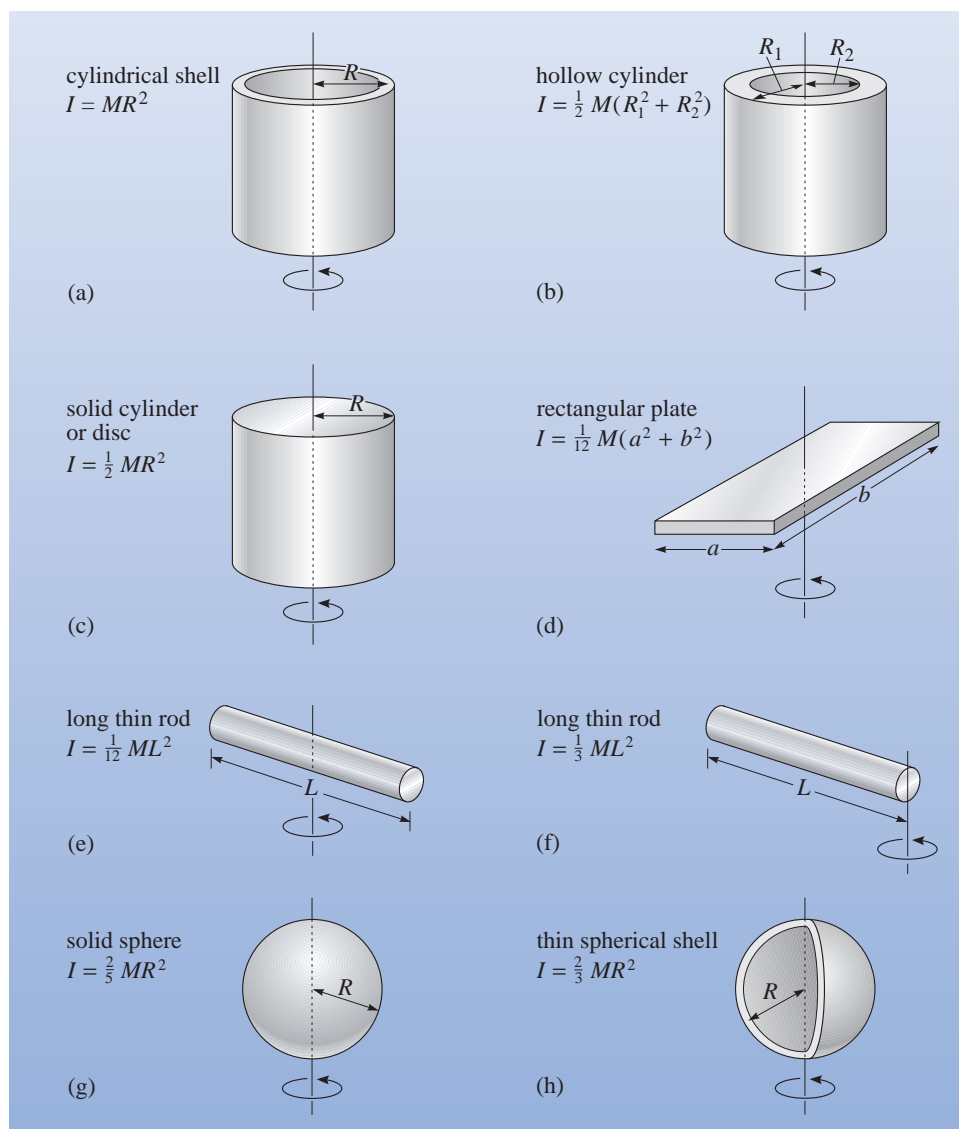


Figure 5.7 Moments of inertia about specified axes for some common uniform solids.

The angular analogue of translational kinetic energy is, not surprisingly, rotational kinetic energy. For a body rotating with angular speed ω , about a fixed axis

associated with moment of inertia I , the **rotational kinetic energy** is given by

$$E_{\text{rot}} = \frac{1}{2}I\omega^2 \quad (5.47)$$

The rotational analogue of linear momentum is **angular momentum**. The letter j is often used for the angular momentum of a particle, and \mathbf{J} for the angular momentum of a system such as a rigid body. (Nb. Note that sometimes the letters \mathbf{l} and \mathbf{L} are used instead for angular momentum.) It is a vector quantity and defined such that the angular momentum \mathbf{j} about a point O of a particle with linear momentum \mathbf{p} is

$$\mathbf{j} = \mathbf{r} \times \mathbf{p} \quad (5.48)$$

where \mathbf{r} is the displacement vector of the particle from O (see Figure 5.8). The direction of the angular momentum vector is perpendicular to the plane containing the linear momentum and displacement vectors, as given by the right-hand rule.

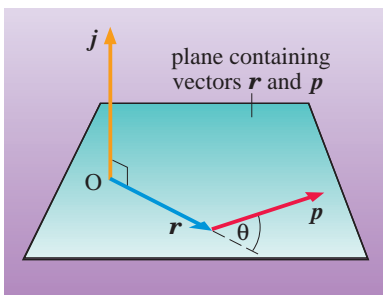


Figure 5.8 The magnitude of the angular momentum j is equal to the product of the magnitude of the linear momentum p , the distance from the axis of rotation r and the sine of the angle between these two vectors. The direction of \mathbf{j} is perpendicular to the plane containing the other two vectors, given by the right-hand rule.

For a rotating rigid body, the angular momentum \mathbf{J} about a given point depends on the way the body's mass is distributed, and on the components of its angular velocity ω . Although not generally true, for all the situations that you are likely to come across, \mathbf{J} is parallel to ω and

$$\mathbf{J} = I\omega \quad (5.49)$$

The angular momentum \mathbf{J} for *any* object subject to an external torque Γ satisfies the equation

$$\Gamma = \frac{d\mathbf{J}}{dt} \quad (5.50)$$

In situations where $\mathbf{J} = I\omega$, and I is a constant, it follows that

$$\Gamma = I \frac{d\omega}{dt} \quad (5.51)$$

where $d\omega/dt$ is the angular acceleration. This is in effect the rotational analogue of Newton's second law of motion.

According to the **principle of conservation of angular momentum**: for any system, the total angular momentum about any point remains constant as long as no net external torque acts on that system.

Exercise 5.7 (a) A truncated accretion disc orbiting around a compact star may be modelled as an annulus of material of mass M with inner and outer radii R_1 and R_2 respectively. Explain qualitatively why the moment of inertia of the

annulus about a central axis is different from that of a uniform disc of material of the same mass and radius R_2 . (b) Assuming an annulus and disc both rotate as solid bodies, and at the same angular speed as each other, explain which will have the greater angular momentum and which will have the greater rotational kinetic energy.



5.6 Properties of gases

The physics of matter is generally concerned with the three *phases* referred to as solids, liquids and gases. However, in astrophysics and cosmology we are almost always dealing solely with **gases** (or with **plasmas**, which are ionized gases).

The vast number of molecules present in any normal sample of gas means that the macroscopic properties of the gas can be deduced from the average behaviour of the molecules. Random fluctuations can be neglected. For instance, the pressure of a gas, detected on the walls of a container, is due to the ceaseless random bombardment of gas molecules. The **pressure** P at any point in a gas is the magnitude of the perpendicular force per unit area that the gas would exert on a surface at that point, so

$$P = F/A \quad (5.52)$$

The SI unit of pressure is the pascal (Pa), where $1 \text{ Pa} = 1 \text{ N m}^{-2}$. Another macroscopic property of a gas is its **density** which is simply the mass of the sample divided by its volume. It is usually represented by the symbol ρ . Thus

$$\rho = M/V \quad (5.53)$$

where the **volume** in question is often that of a spherical region and as such is given by

$$V = \frac{4}{3}\pi R^3 \quad (5.54)$$

where R is the radius of the sphere.

Another important macroscopic property of a gas is its temperature. Energy transferred from a warm body to a cool body, as a result of a difference in temperature, is known as **heat**. Energy transferred by non thermal means is classified as work. One way of characterizing **temperature** is to say that it is a label that determines the direction of heat flow: heat flows from a body with a higher temperature to a body with a lower temperature, and keeps on flowing until both bodies are at the same temperature. The lowest conceivable temperature is -273.15°C . This is known as the **absolute zero** of temperature. According to classical physics, molecules stop moving at absolute zero, and have zero kinetic energy. On the absolute temperature scale, absolute zero is taken to be zero kelvin (0 K), and 0°C becomes $+273.15 \text{ K}$.

One **mole** of a substance is an amount of it containing its relative atomic mass (for an element) or its relative molecular mass (for a compound) in grams. One mole of any substance contains the same number of basic particles (usually atoms or molecules). The number of basic particles per mole is called **Avogadro's**

constant N_m , equal to $6.02 \times 10^{23} \text{ mol}^{-1}$. Thus $M_m = M_r \times 10^{-3} \text{ kg mol}^{-1}$ and $M_m = N_m m$ where M_m is the molar mass of the element or compound, M_r is the relative atomic or molecular mass and m is the actual mass of the atom or molecule.

The simplest model of a gas treats molecules as structureless particles in random motion. An **ideal gas** is one in which the only interactions are elastic collisions (i.e. kinetic energy is conserved) and other molecular or gravitational interactions are neglected. Most gases under normal conditions are approximately ideal. In this model, the molecules collide with one another and with the walls of their container, subject to the laws of Newtonian mechanics. As a result, the model predicts that the pressure exerted by a gas depends on the average translational energy $\langle E_{\text{KE}} \rangle$ of the gas molecules:

$$PV = \frac{2}{3}N\langle E_{\text{KE}} \rangle \quad (5.55)$$

where P is the pressure exerted by the gas, V its volume and N the number of molecules. The Maxwell–Boltzmann energy distribution can be used to show that the average translational energy of the particles in a gas is

$$\langle E_{\text{KE}} \rangle = \frac{3}{2}kT \quad (5.56)$$

where T is the absolute temperature of the gas and k is the **Boltzmann constant** ($1.38 \times 10^{-23} \text{ J K}^{-1}$). These two equations lead to the **ideal gas equation of state** which links the macroscopic properties of a gas in equilibrium:

$$PV = NkT \quad (5.57)$$

Alternatively we may write

$$PV = nRT \quad (5.58)$$

where n is the number of moles of the gas, and R is the molar gas constant ($R = 8.314 \text{ J K}^{-1} \text{ mol}^{-1} = kN_m$).

- If a gas consists of N atoms, each of mass m , what is the relationship between the pressure (P) and density (ρ) of the gas and its temperature (T)?
- The total mass of the gas is Nm and so the density of the gas is $\rho = Nm/V$. Equation 5.57 may be rearranged as $P = NkT/V$, so combining these two equations we have

$$P = \rho kT/m \quad (5.59)$$

Exercise 5.8 The photosphere of the Sun contains about 10^{16} hydrogen atoms per cubic centimetre at a temperature of 5800 K. (a) What is the average density of the photosphere? (b) What is the average translational kinetic energy of the atoms in electronvolts? (c) What is the pressure in the gas? (For the purposes of this question you may neglect the helium and other atoms present in the Sun's photosphere.)

As noted above, although we may quantify the average translational kinetic of the particles in a gas, in equilibrium, there will be a distribution of speeds of

those particles, and a distribution of their translational kinetic energies. Such distributions can be represented by histograms, such as those in Figure 5.9, in which the height of each bar is the fractional frequency of particles with speeds or translational kinetic energies in the range of the width of the bar. The sum of the heights of all the histogram bars is equal to 1.

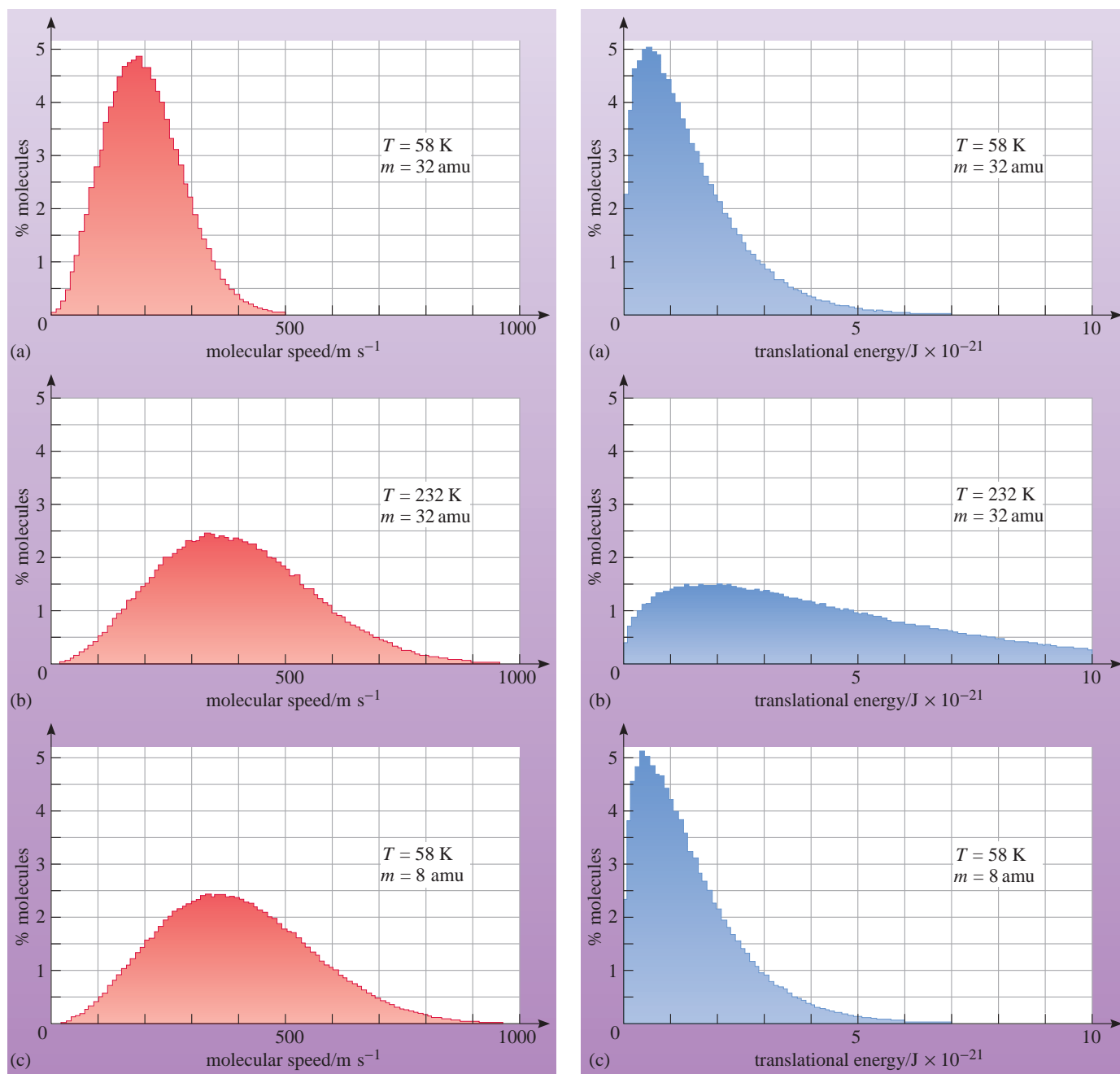


Figure 5.9 (Left) Histograms of the distribution of particle speeds in a gas in equilibrium for different temperatures and particle masses. (Right) Histograms of the distribution of translational kinetic energies in a gas in equilibrium for different temperatures and particle masses.

As can be seen in the left-hand panels of Figure 5.9 the speed distribution is

slightly asymmetric, with a tail extending to high speeds. As the temperature rises, the equilibrium speed distribution becomes broader and shifts to higher speeds, and the peak of the distribution becomes less pronounced (compare panels (a) and (b) of the left-hand Figure 5.9). In fact the average speed is proportional to the square root of the absolute temperature. Also, as the particle mass decreases, the equilibrium speed distribution becomes broader and shifts to higher speeds, and the peak of the distribution becomes less pronounced (compare panels (a) and (c) of the left-hand Figure 5.9). In more detail, the average speed is inversely proportional to the square root of the particle mass. (*Note:* Particle masses are given in units of atomic mass units (amu).) The distribution of particle speeds is quantified by the **Maxwell–Boltzmann speed distribution** function as

$$f(v) = 4\pi \left(\frac{m}{2\pi kT} \right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2kT}\right) \quad (5.60)$$

In this distribution, the most probable speed of the particles is $v_{\text{mp}} = \sqrt{2kT/m}$, the average speed of the particles is $\langle v \rangle = \sqrt{8kT/\pi m}$ and the root mean square speed of the particles is $v_{\text{rms}} = \sqrt{3kT/m}$.

Similarly, as can be seen in the right-hand panels of Figure 5.9, the translational kinetic energy distribution is very asymmetric, with a long tail extending to high energies. As temperature increases, the equilibrium energy distribution becomes broader and shifts to higher energies, and the peak of the distribution becomes less pronounced (compare panels (a) and (b) of the right-hand Figure 5.9). In fact the average energy is proportional to the absolute temperature. However, the equilibrium energy distribution is independent of particle mass: at a fixed temperature, all gases have the same equilibrium distribution of translational kinetic energy (compare panels (a) and (c) of the right-hand Figure 5.9). The distribution of translational kinetic energies is quantified by the **Maxwell–Boltzmann energy distribution** function as

$$g(E) = \frac{2}{\sqrt{\pi}} \left(\frac{1}{kT} \right)^{3/2} \sqrt{E} \exp\left(-\frac{E}{kT}\right) \quad (5.61)$$

In this distribution, the most probable translational kinetic energy of the particles is $E_{\text{mp}} = kT/2$ and as noted earlier, the average translational kinetic energy of the particles is $\langle E \rangle = 3kT/2$.

A final property of gases that it is important to consider is their ability to act as a medium in which sound waves can travel. A **sound wave** is a longitudinal variation in pressure and density consisting of periodic compressions and rarefactions of the gas, as shown in Figure 5.10 (overleaf).

A sound wave, like any wave, may be characterized by its wavelength and frequency (see Section 5.10) and by its speed of propagation. Although sound waves are most familiar to us as a means of communication (usually in air) they can and do exist in any gas (or plasma) and are generated in a variety of astrophysical and cosmological contexts. In general, the **sound speed** in a particular medium, denoted by the symbol c_s , depends on the ratio of the pressure to the density of the gas, or equivalently on the temperature of the gas (since $P/\rho \propto T$, see Equation 5.59):

$$c_s \propto (P/\rho)^{1/2} \propto T^{1/2} \quad (5.62)$$

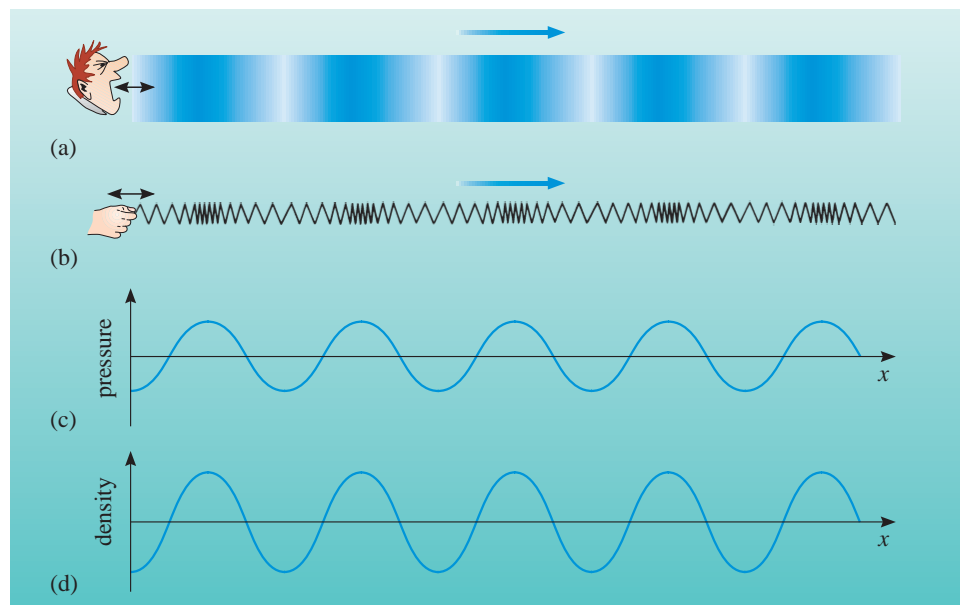


Figure 5.10 Sound waves (a) are analogous to longitudinal waves on a spring (b). A sound wave can be represented graphically by plotting the variation of (c) the gas pressure or (d) the gas density against position.

5.7 Atoms and energy levels

Although there are significant amounts of dark matter in the Universe, whose nature is at present largely unknown, the structures that we can observe directly (planets, stars, galaxies, clusters, etc.) are composed of atoms, and there are known to be around ninety different types of atom that occur naturally in the Universe. A material made of a single type of atom is known as an **element**. The most abundant elements in the Universe as a whole are those comprising the two simplest atoms: hydrogen and helium. Here on Earth, there are also significant amounts of other elements, in particular carbon, nitrogen, oxygen, sodium, magnesium, aluminium, silicon, sulfur, calcium and iron. This section describes the basic properties of atoms and energy levels, whilst the subsequent section describes the quantum physics that underlies these properties.

5.7.1 Atomic structure

Whatever the type of atom, each one has certain characteristic features. Each contains a central **nucleus**, which carries a positive electric charge as well as most of the atom's mass. The nucleus is surrounded by one or more negatively charged **electrons** (symbol: e) each of which has a much lower mass than the nucleus. The nucleus of an atom is what determines the type of element. The very simplest atoms of all, those of the element hydrogen, have a nucleus consisting of just a single **proton** (symbol: p). The next simplest atom, helium, has two protons in its nucleus; lithium has three protons; beryllium has four; boron has five; carbon has six; and so on. The number of protons in the nucleus of an atom is known as its **atomic number** (Z). The Periodic Table of the elements is displayed in Figure 5.11 where the atomic number and chemical symbol are shown for each element.

The electric charge of a proton is represented by the algebraic quantity $+e$ with a

numerical value of 1.60×10^{-19} C. The electric charge of an electron is exactly the same magnitude as that of a proton, but negative instead of positive, so is written as $-e$, which has a value of -1.60×10^{-19} C.

- What is the atomic number of carbon? What is the electric charge of a carbon nucleus?
- The nucleus of a carbon atom contains six protons, so the atomic number of carbon is 6 and the charge of the nucleus is $+6e$, which is equivalent to $6 \times (1.60 \times 10^{-19} \text{ C}) = 9.60 \times 10^{-19} \text{ C}$. To the nearest order of magnitude, this is therefore 10^{-18} C .

Group																	Group												
I	II																	III	IV	V	VI	VII	0						
Period 1																					1	2							
																					H	He							
Period 2	3	4																	5	6	7	8	9	10					
	Li	Be																	B	C	N	O	F	Ne					
Period 3	11	12																	13	14	15	16	17	18					
	Na	Mg																	Al	Si	P	S	Cl	Ar					
Period 4	19	20																	21	22	23	24	25	26	27	28	29	30	
	K	Ca																	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	
Period 5	37	38	Lanthanides																39	40	41	42	43	44	45	46	47	48	
	Rb	Sr																	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	
Period 6	55	56	Lanthanides																71	72	73	74	75	76	77	78	79	80	
	Cs	Ba																	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	
Period 7	87	88	Actinides																103	104	105	106	107	108	109	110	111	112	
	Fr	Ra																	Lr	Rf	Db	Sg	Bh	Hs	Mt				
		actinides																transition elements											
		typical elements																											

Figure 5.11 The Periodic Table showing the atomic number and chemical symbol for each element.

The other constituents of atomic nuclei are **neutrons** (symbol: n) which have a similar mass to protons, but have zero electric charge. Normal hydrogen atoms have no neutrons in their nuclei, although there is a form of hydrogen known as deuterium that does. The nucleus of a deuterium atom consists of a proton and a neutron. It is still the element hydrogen (since it contains one proton) but it is a heavy form of hydrogen, thanks to the extra neutron. Deuterium is said to be an isotope of hydrogen. Similarly, normal helium atoms contain two neutrons in their nucleus, along with the two protons; but a lighter isotope of helium, known as helium-3, contains only one neutron instead. The total number of protons and neutrons in the nucleus of an atom is the **mass number** (A) of the atom. **Isotopes** therefore denote forms of the same element with different mass numbers. The nucleus of a particular isotope is referred to as a **nuclide**.

As a shorthand, isotopes of each atomic element may be represented by a symbol. Letters are used to indicate the name of the element itself, and two numbers are used to indicate the atomic number (lower prefix) and mass number (upper prefix). So a normal hydrogen atom is represented as ${}^1_1\text{H}$, and an atom of the heavier isotope, deuterium, by ${}^2_1\text{H}$. When an isotope is written 'element- X ', for example helium-3, then the value of X is always the mass number A .

- What are the mass numbers of (a) normal hydrogen (b) heavy hydrogen (i.e. deuterium) (c) normal helium and (d) helium-3 ?
- (a) The nucleus of normal hydrogen contains one proton, so the mass number is 1. (b) The nucleus of heavy hydrogen contains one proton and one neutron so the mass number is 2. (c) The nucleus of normal helium contains two protons and two neutrons, so the mass number is 4. (d) The nucleus of helium-3 contains two protons and one neutron, so the mass number is 3.
- What is the symbol for the isotope of carbon which has six protons and eight neutrons in its nucleus?
- ${}^{14}_6\text{C}$.
- How many protons and how many neutrons does a nucleus of the uranium isotope ${}^{238}_{92}\text{U}$ contain?
- 92 protons and $(238 - 92) = 146$ neutrons.

Sometimes, protons and neutrons are collectively referred to as **nucleons** since both types of particle are found inside the nucleus of an atom. Similarly, electrons, protons and neutrons are often collectively referred to as sub-atomic particles, for obvious reasons.

Normal atoms are electrically neutral, so the positive electric charge of the nucleus is exactly balanced by the negative electric charge of the electrons surrounding it. Since each electron carries an electric charge of $-e$ and each proton carries an electric charge of $+e$, the number of electrons in a neutral atom is exactly the same as the number of protons in its nucleus.

- What is the difference between atoms of lithium-7 and beryllium-7?
- Both atoms have the same mass number, namely 7. However, the nucleus of the lithium atom has 3 protons and 4 neutrons, whilst the nucleus of the beryllium atom has 4 protons and 3 neutrons. Furthermore, the lithium atom contains 3 electrons whilst the beryllium atom contains 4 electrons.
- The element iron has an atomic number 26, and its most common isotope is known as iron-56. (a) How many protons and how many neutrons are there in a single nucleus of iron-56? (b) How many electrons are there in an electrically neutral atom of iron-56?
- (a) Since the atomic number is 26, the nucleus contains 26 protons. Since the mass number is 56, the total number of protons and neutrons in the nucleus is 56, and so the nucleus contains $56 - 26 = 30$ neutrons. (b) An electrically neutral atom contains the same number of electrons as protons, so the atom contains 26 electrons.

Finally, here is a reminder of the size of atoms and nuclei. Whereas a typical atomic nucleus has a size of around 10^{-14} m, the size of the atom itself is determined by the size of the region occupied by the electrons that surround the nucleus. The overall size of an atom is about 10^{-10} m across.

5.7.2 Photons and energy levels

Each type of atom can be characterized by the energies of the photons it can absorb or emit. In this section we look deeper into the structure of atoms and

show how atoms absorb and emit characteristic energies. As noted earlier, the explanation for this, in terms of quantum mechanics, is examined in the subsequent section.

A **photon** is a ‘particle’ of electromagnetic radiation. Monochromatic light, which has a single colour, consists of identical photons that each have exactly the same energy. The amount of energy carried by a single photon is called a **quantum** and quanta of visible light have energies of around 2 to 3 eV. The energy corresponding to each colour is shown in Figure 5.12.

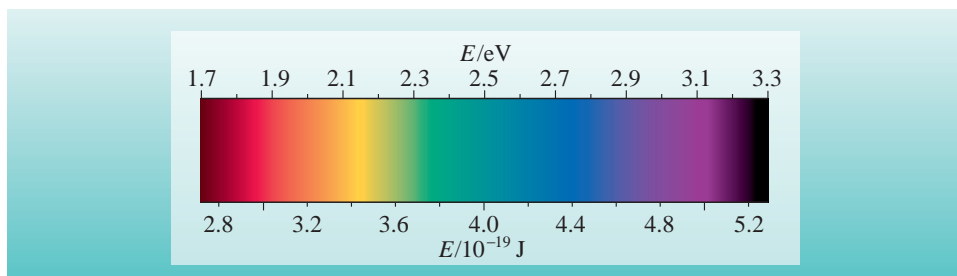


Figure 5.12 A continuous spectrum of visible light showing the corresponding photon energy for each colour.

When the photons emitted by a particular type of atom are dispersed to form a spectrum, the spectral lines of atomic spectra provide a unique ‘fingerprint’ of the atoms concerned (Figure 5.13).

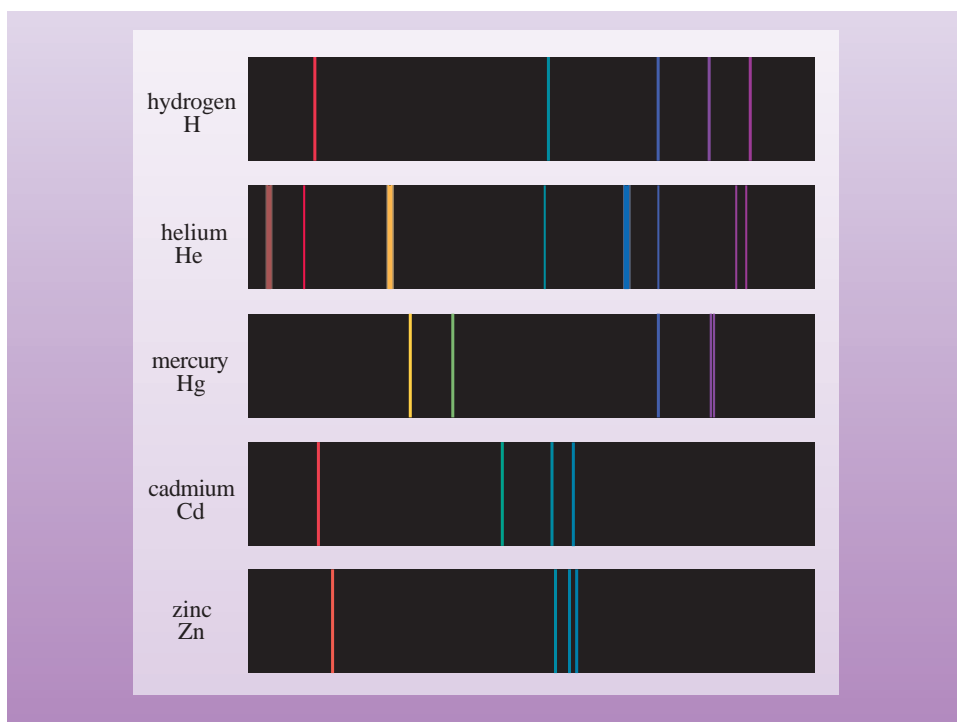


Figure 5.13 Each type of atom may be characterized by a unique spectrum of emission lines, some examples of which are shown here in the visible part of the spectrum.

The explanation for atomic spectra is that atoms can only exist with certain values of energy, known as **energy levels**. Transitions, often referred to as *quantum jumps*, can occur between these energy levels. When an atom has its lowest

possible energy, it is said to be in its **ground state**, whilst higher energy levels correspond to **excited states** of the atom. To make atoms jump to excited states, photons of the correct energies must be supplied, and the result is the **absorption** of a photon (Figure 5.14a). When an atom jumps from one energy level to another of lower energy, the energy that it loses is taken away by a photon, and the result is the **emission** of a photon (Figure 5.14b). In general, the energy of the photon is equal to the *change* in energy of the atom:

$$E_{\text{ph}} = \Delta E_{\text{atom}} \quad (5.63)$$

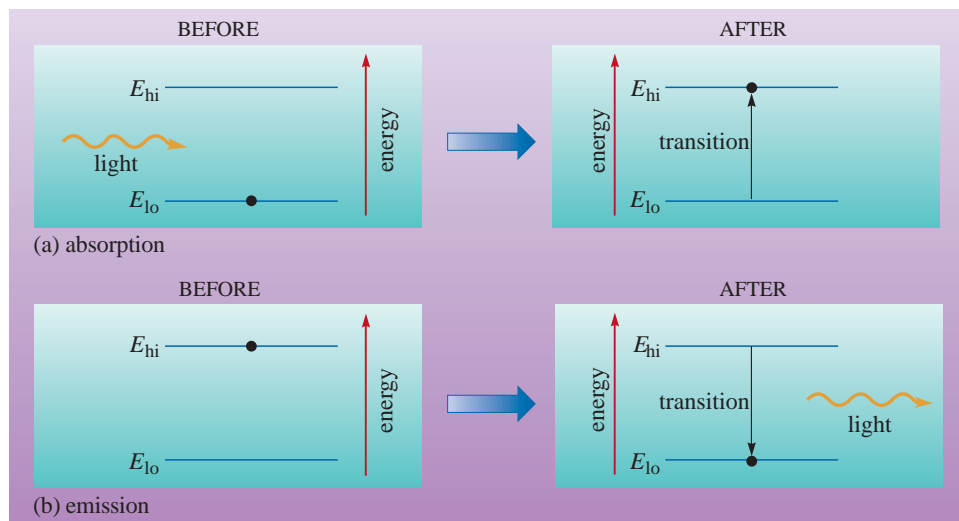


Figure 5.14 The two horizontal lines in each part of this figure represent two of the possible energies of an atom – two of the possible energy levels, labelled E_{hi} and E_{lo} . When the atom occupies an energy level, the energy level is marked with a dot. (The widths of the horizontal lines in this diagram are of no significance.) (a) When a photon is absorbed by an atom, the atom makes a transition from a lower energy level to a higher energy level. (b) A photon is emitted by an atom when the atom makes a transition from a higher energy level to a lower energy level.

Hydrogen is the simplest type of atom, consisting as it does of a single electron bound to a single proton. As you might expect, it therefore has the simplest energy-level diagram of any element (Figure 5.15). This consists of a series of levels which get progressively closer together at higher and higher energies. The energy E_n associated with the n th energy level of hydrogen is given by

$$E_n = \frac{-13.60}{n^2} \text{ eV} \quad (5.64)$$

Consequently, the energy of the ground state ($n = 1$) is -13.60 eV whilst higher energy levels have larger (i.e. less negative) energies, until the energy level identified as $n = \infty$ has an energy of 0 eV. Series of transitions whose lowest energy level corresponds to a particular value of n have been given names according to the scientists who first observed them, as shown in Figure 5.15.

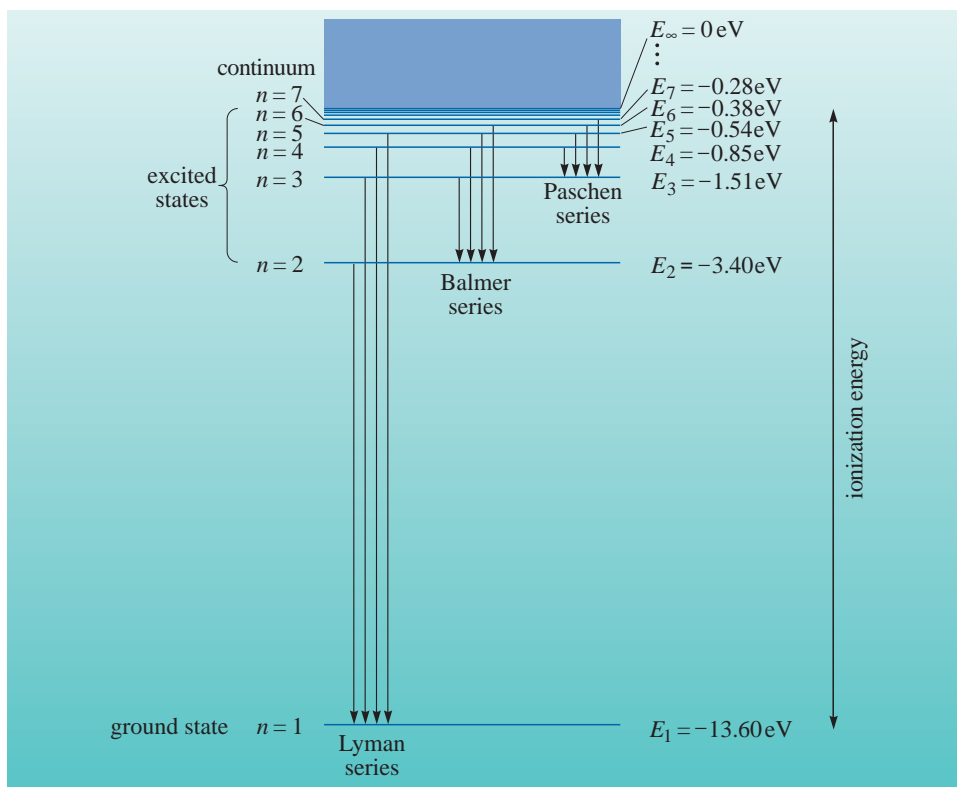


Figure 5.15 The energy-level diagram of hydrogen showing various series of transitions.

Exercise 5.9 Use the information presented in Figure 5.15 to calculate the photon energies corresponding to the first four lines in the hydrogen Balmer series.

If a hydrogen atom in its ground state absorbs a photon whose energy is greater than 13.60 eV, the atom will find itself in the continuum region shown at the top of Figure 5.15. In this state the atom is **ionized**, that is, the electron and proton which comprised the hydrogen atom will no longer be bound together. In effect, 13.60 eV of the photon's energy has been used to split the electron and proton apart, and any remaining amount of the photon's energy is imparted to the now free particles as kinetic energy. The **ionization energy** of a hydrogen atom is therefore 13.60 eV.

Different elements have different energy-level diagrams and different ionization energies. Furthermore, since all other elements possess more than one electron when neutral (helium has two electrons, lithium has three electrons, etc.) atoms of other elements can undergo successive ionizations as successive electrons are removed. A nomenclature used in astrophysics is that neutral hydrogen is denoted by HI, and ionized hydrogen by HII. Similarly, neutral helium is HeI, singly ionized helium is HeII, etc.

- What do you suppose is implied by the notation CIV?
- This is triply-ionized carbon, i.e. carbon from which three electrons have been removed.

If an atom has all but one of its electrons removed, as a result of ionization processes, it is referred to as a hydrogen-like ion. The energy levels of a hydrogen-like ion with nuclear charge Ze are given by

$$E_n = Z^2 \times \frac{-13.60}{n^2} \text{ eV} \quad (5.65)$$

which differs from Equation 5.64 (the corresponding equation for the hydrogen atom) only by the factor of Z^2 , where Z is the atomic number of the ion in question. Since $Z = 1$ for the hydrogen atom, the two equations are identical in that case, so Equation 5.65 is really just a generalisation of Equation 5.64.

The normal state of an atom is for it to sit in its ground state, which is the state of lowest possible energy for that atom. However, atoms in a gas will tend to occupy different states. The **Boltzmann equation** predicts the number of atoms N_n in a particular energy level (with energy E_n) relative to the number of atoms N_1 in the ground state (with energy E_1) for a gas at a temperature T :

$$\frac{N_n}{N_1} = \frac{g_n}{g_1} \exp\left(-\frac{(E_n - E_1)}{kT}\right) \quad (5.66)$$

The constants g_n and g_1 are weighting factors and k is Boltzmann's constant.

Similarly, the Saha ionization equation predicts the number of ions N^+ relative to the number of neutral atoms N_1 in the ground state for a sample of gas at a temperature T :

$$\frac{N^+}{N_1} = \left(\frac{2\pi m_e}{h^2}\right)^{3/2} \frac{(kT)^{3/2}}{N_e} \exp\left(-\frac{I}{kT}\right) \quad (5.67)$$

The number of free electrons is N_e and I is the ionization energy. Other quantities are all constants denoted by their usual symbols.

5.8 Quantum physics

The ideas of quantum physics are only required when we try to understand phenomena on an atomic scale. Since, in astrophysics and cosmology we are concerned with the microphysics of the interaction of matter and radiation within stars and galaxies, at the fundamental level we are indeed examining processes on an atomic scale, and so require aspects of quantum physics in order to understand what is going on. For the purposes of this document, we merely remind you of a few results from quantum physics, so that you can proceed to tackle further topics in astrophysics and cosmology where an application of these results is important.

Vital to quantum physics is an appreciation that, in order to interpret the behaviour of particles such as electrons, it is necessary to apply both *wave* and *particle* models. The two different models are required to explain different aspects of their behaviour: for example, broadly speaking, electrons propagate like waves but are absorbed or emitted like particles. In particular, any particle will have an associated wavelength, known as its **de Broglie wavelength**, given by

$$\lambda_{\text{dB}} = h/p \quad (5.68)$$

where h is Planck's constant (6.63×10^{-34} J s) and p is the magnitude of the particle's momentum.

- What is the de Broglie wavelength of a proton moving at a speed of 500 km s^{-1} , a typical speed for a proton in the core of the Sun?
($m_p = 1.67 \times 10^{-27} \text{ kg}$)
- The magnitude of the proton's momentum is
 $p = mv = (1.67 \times 10^{-27} \text{ kg}) \times (500 \times 10^3 \text{ m s}^{-1}) = 8.35 \times 10^{-22} \text{ kg m s}^{-1}$
so its de Broglie wavelength is
 $\lambda_{\text{dB}} = h/p = (6.63 \times 10^{-34} \text{ J s}) / (8.35 \times 10^{-22} \text{ kg m s}^{-1}) = 7.94 \times 10^{-13} \text{ m}$

5.8.1 Wave mechanics

As just noted, in order to interpret the behaviour of particles such as electrons, it is necessary to apply both wave and particle models. The two different models are required to explain different aspects of their behaviour. In experiments involving just a single electron, it is impossible to predict the outcome of that experiment, only the probability of the various possible outcomes can be predicted. The probability P of detecting a particle at a particular place is proportional to the square of the amplitude A of the particle's de Broglie wave at that place, $P \propto A^2$.

Quantum mechanics asserts that the behaviour of matter can be modelled using probability waves. A localized particle can be modelled as a wavepacket which can be produced by summing many infinitely long waves with a range of wavelengths and amplitudes. The range of wavelengths required to build up a localized particle implies an uncertainty in our knowledge of the particle's momentum. The more tightly localized a particle is, the greater is this uncertainty. This idea leads to **Heisenberg's uncertainty principle**, which has two formulations:

$$\Delta x \Delta p_x \geq \hbar/2 \quad (5.69)$$

and

$$\Delta E \Delta t \geq \hbar/2 \quad (5.70)$$

where $\hbar = h/2\pi$.

Exercise 5.10 An electron in an excited atom typically remains in an excited state for about 10^{-8} s before it loses the excess energy by emitting electromagnetic radiation. Use the energytime uncertainty relationship to estimate the indeterminacy that this implies for the energy of the excited atomic states. In physical terms, what effect do you think this indeterminacy will have on measurements of spectral lines?

According to Schrödinger's wave-mechanical approach to quantum mechanics, the information describing the behaviour of a particle is contained in its **wavefunction** Ψ , which is the solution to the **time-dependent Schrödinger equation** for that particle. The solution to Schrödinger's equation for a free particle of mass m involves travelling waves characterized by an angular wavenumber $k (= 2\pi/\lambda_{\text{dB}})$ that may have any positive value. The relationship between k and the particle's kinetic energy is

$$E_{\text{kin}} = \frac{\hbar^2 k^2}{2m} \quad (5.71)$$

So, for a free particle, any positive value of the kinetic energy (and hence total energy) is allowed.

For a particle of mass m in a stationary state, described by a wavefunction of the form $\Psi(x, t) = \psi(x)\phi(t)$, the time-independent wavefunction $\psi(x)$ will satisfy the **time-independent Schrödinger equation**:

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m}{\hbar^2} (E_{\text{tot}} - E_{\text{pot}}(x)) \psi(x) = 0 \quad (5.72)$$

where E_{pot} is the potential energy function of the particle.

For a particle in a *one-dimensional infinite square well* of width D (Figure 5.16), the time-independent wavefunctions take the form $\psi(x) = \psi_0 \sin kx$ and $\psi(x) = \psi_0 \cos kx$, but only certain values of k are allowed, and therefore only certain discrete values of the energy are allowed. That is, the particle's energy is quantized:

$$E_{\text{tot}} = \frac{n^2 h^2}{8mD^2} \quad (5.73)$$

where $n = 1, 2, 3$, etc.

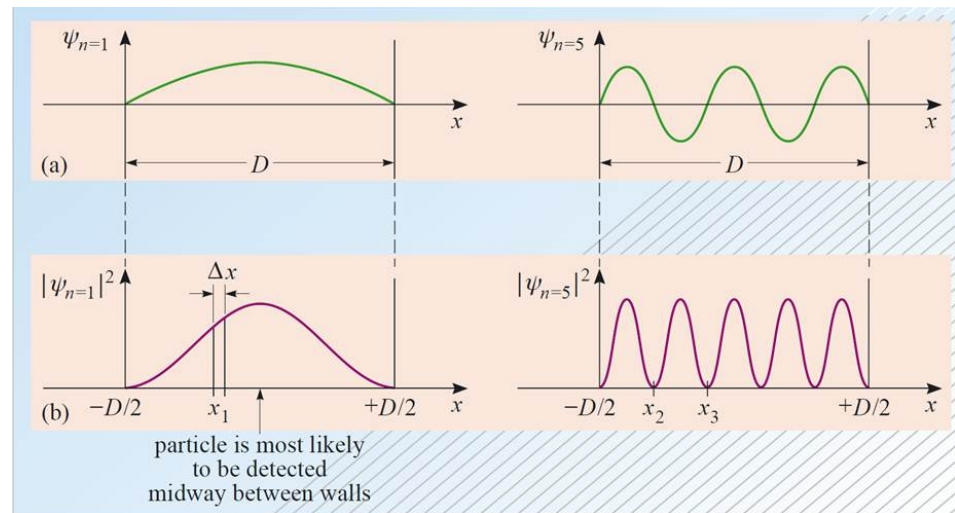


Figure 5.16 (a) Two of the standing probability waves, $\psi_{n=1}$ and $\psi_{n=5}$, that can describe the particle in a one-dimensional infinite square well. (b) The relative probability of detecting the particle in different regions, when the particle is described by the standing probability waves $\psi_{n=1}$ and $\psi_{n=5}$.

Exercise 5.11 An electron is confined between infinitely high, rigid walls positioned at locations $\pm D/2$ on the x -axis.

(a) Describe the time-independent wavefunction that describes the electron when it has a total energy of $E_{\text{tot}} = 9h^2/(8m_e D^2)$.

(b) Describe the positions where the electron is most likely to be detected when it has a total energy of $E_{\text{tot}} = 9h^2/(8m_e D^2)$.

(c) What would happen if the electron made a transition from the energy level $E_{\text{tot}} = 9h^2/(8m_e D^2)$ to the energy level $E_{\text{tot}} = h^2/(8m_e D^2)$.

In a *finite one-dimensional square well* the particle's energy is still quantized but the wavefunction penetrates to some extent into the classically forbidden region outside the well where $E_{\text{pot}} > E_{\text{tot}}$ (Figure 5.17). This means, of course, that there is a finite probability of finding the particle in this region.

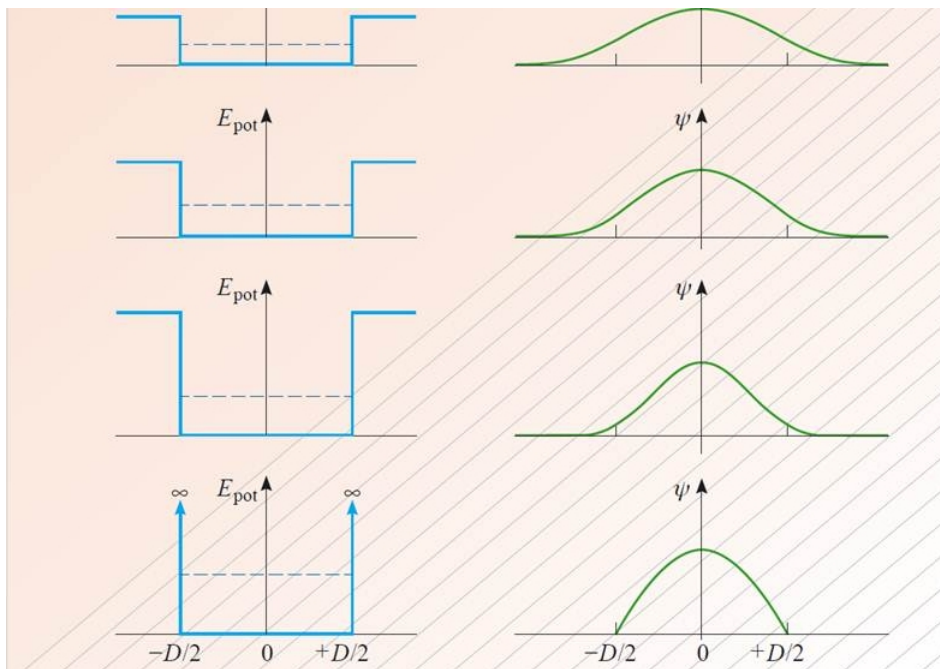


Figure 5.17 Sketches of the lowest energy time-independent wavefunction for a particle confined in a potential well of various heights. The wavefunction penetrates further into the classically forbidden region in a potential well of lower height. The dashed line indicates the energy of the ground state (state of lowest energy) in each case.

Conversely if the particle's energy is greater than the height of the walls of the well, $E_{\text{tot}} > W$, then the particle is unbound and the solutions are travelling waves. In each region separately, the wavefunction has an angular wavenumber given by $\sqrt{2m(E_{\text{tot}} - E_{\text{pot}})}/\hbar$.

One extraordinary consequence of the penetration into regions outside finite wells is that if the potential energy function is not just a *well* but a *barrier* of finite width and height, then the wavefunction can still have a finite value on the outside of the barrier. We can interpret this as meaning that there is a finite probability of the particle escaping from its confinement even though its total energy E_{tot} is nowhere near enough (in classical terms) for it to surmount the potential energy barrier of height W . This phenomenon is known as **barrier penetration** or **tunnelling** and has many important applications in physics. In the classically forbidden region within the barrier, the wavefunction decays exponentially at a rate which will depend on $W - E_{\text{tot}}$. So for walls much higher than E_{tot} or walls which are wide, the wavefunction will have decayed to a smaller value than for lower or narrower walls. In all cases (wells and barriers) also note that the continuity requirements are met at all the boundaries: the wavefunction and its first derivative (slope) match (i.e. they are continuous) at all points.

A summary of these results for Schrödinger's time-independent equation in one

dimension is provided in Figure 5.18.

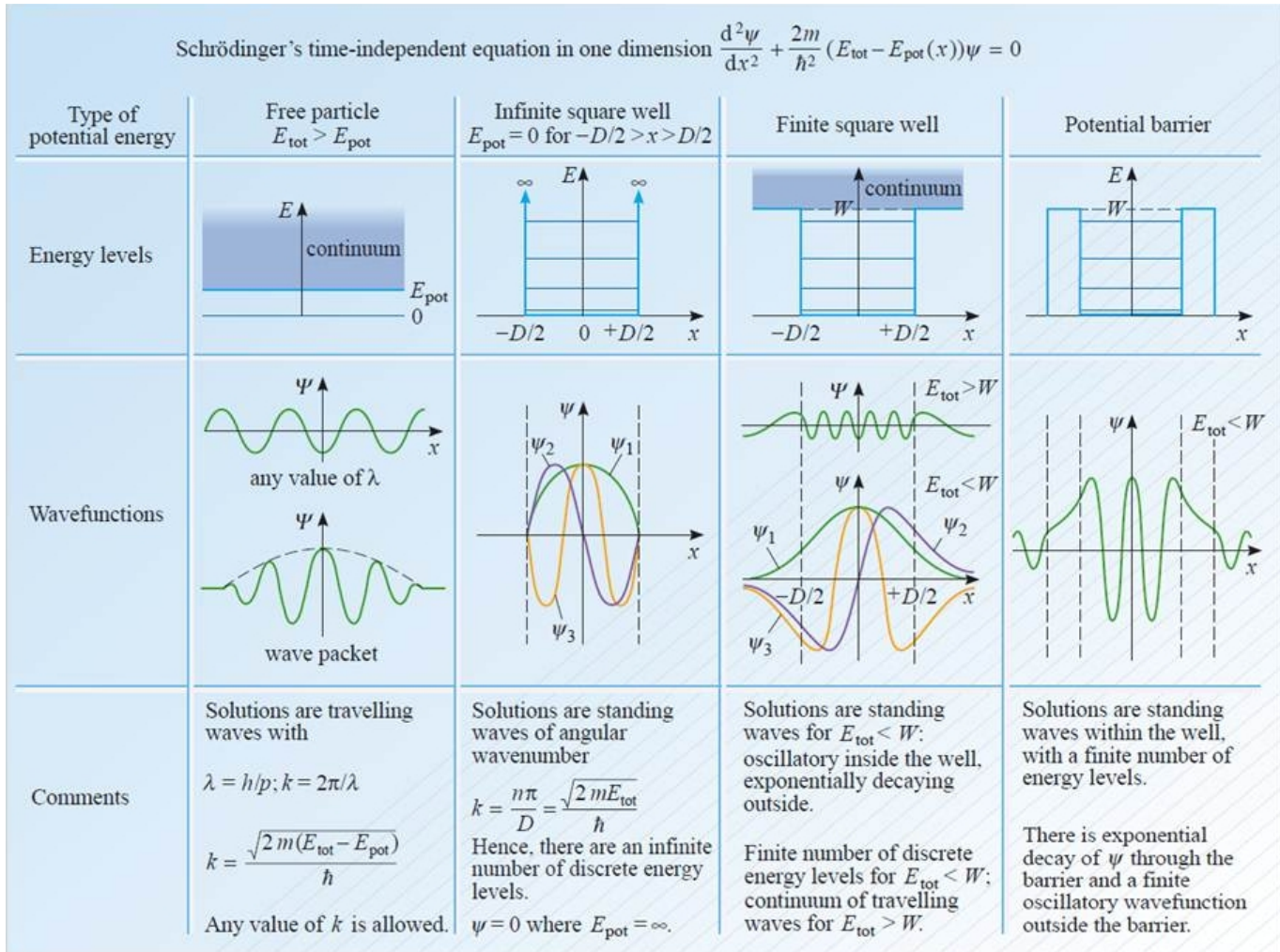


Figure 5.18 A summary of results for Schrödinger's time-independent equation in one dimension.

The probability P of finding a confined particle in a small region Δx at position x in a one-dimensional well is $P = |\psi(x)|^2 \Delta x$, if the time-independent wavefunction $\psi(x)$ is normalized. More generally, for particles described by a normalized time-dependent wavefunction $\Psi(x, t)$, the probability is $P(x, t) = |\Psi(x, t)|^2 \Delta x$. For a particle confined in three dimensions the energy levels are given by

$$E_{\text{tot}} = \frac{\hbar^2}{8mD^2} (n_1^2 + n_2^2 + n_3^2) \tag{5.74}$$

where n_1, n_2 and n_3 are positive integers. This leads to the phenomenon of degeneracy, where different combinations of the three quantum numbers n_1, n_2 and n_3 can lead to the same value for the energy, E_{tot} .

5.8.2 Quantum mechanics in atoms

The electron in the hydrogen atom is subject to the Coulomb interaction with the proton in the nucleus and it is the electrostatic potential energy associated with this interaction ($-e^2/4\pi\epsilon_0 r$) that is inserted into the Schrödinger equation for the electron in hydrogen. Because of the spherical symmetry of the hydrogen atom, the wavefunctions are most conveniently expressed in spherical polar coordinates (r, θ, ϕ) . Each stationary state wavefunction $\psi(r, \theta, \phi)$ can be written as the product of three wavefunctions that each depend on only one of the coordinates (r, θ, ϕ) . That is $\psi(r, \theta, \phi) = \psi_1(r) \times \psi_2(\theta) \times \psi_3(\phi)$.

When the electron is confined within the hydrogen atom (i.e. when it has $E_{\text{tot}} < 0$) the solutions to Schrödinger's equation are stationary state wavefunctions each of which has a definite energy. The energy levels of the hydrogen atom are given by

$$E_{\text{tot}} = -\frac{1}{n^2} \left(\frac{m_e e^4}{8h^2 \epsilon_0^2} \right) = -\frac{13.6}{n^2} \text{ eV} \quad (5.75)$$

(where $n = 1, 2, 3, \dots$ etc). There is also a continuum of positive energy states available to the electron. These correspond to the states of an ionized hydrogen atom.

Three **quantum numbers** n , l and m_l are required to specify the wavefunction of the electron in hydrogen according to Schrödinger's equation. The principal quantum number n is chiefly responsible for determining the energy of the electron. It can take any integer value. The orbital angular momentum quantum number l determines the magnitude of the electron's orbital angular momentum L according to

$$L = \sqrt{l(l+1)}\hbar \quad (5.76)$$

where $l = 0, 1, 2, \dots, n-1$. The orbital magnetic quantum number m_l determines the component of the electron angular momentum along an arbitrarily chosen z -axis, according to $L_z = m_l \hbar$ where m_l can take the values $m_l = 0, \pm 1, \pm 2, \dots, \pm l$. In the absence of a magnetic field, the energy levels of the electron in hydrogen are degenerate. That is, states with the same value of n but different values of l and m_l have the same energy.

- The electron in a hydrogen atom is described by a wavefunction that is characterized by the principal quantum number $n = 3$. What are the possible values of the magnitude L of the electrons angular momentum?
- The possible values of the electron's orbital angular momentum quantum number are $l = 0, l = 1$ and $l = 2$. So the possible values of the magnitude of the electron's orbital angular momentum are $L = 0, L = \sqrt{2}\hbar$ and $L = \sqrt{6}\hbar$.

When a magnetic field is applied, the energy levels of hydrogen are split due to a magnetic energy term E_{mag} that depends on m_l , namely

$$E_{\text{mag}} = m_l \left(\frac{e\hbar}{2m_e} \right) B_{\text{ext}} \quad (5.77)$$

The spatial distribution of the electron in hydrogen is often illustrated using the **electron cloud** picture, which represents the probability density $|\psi(r, \theta, \phi)|^2$ by the density of a pattern of dots (see Figure 5.19). All states with $l = 0$ are

spherically symmetric, whilst those with $l \geq 1$ have distributions that depend on θ . However, $|\psi_3(\phi)|^2$ is constant for all states. For spherically symmetric s states the radial probability density $4\pi r^2 |\psi|^2$ is the probability per unit radial distance of finding the particle at radius r .

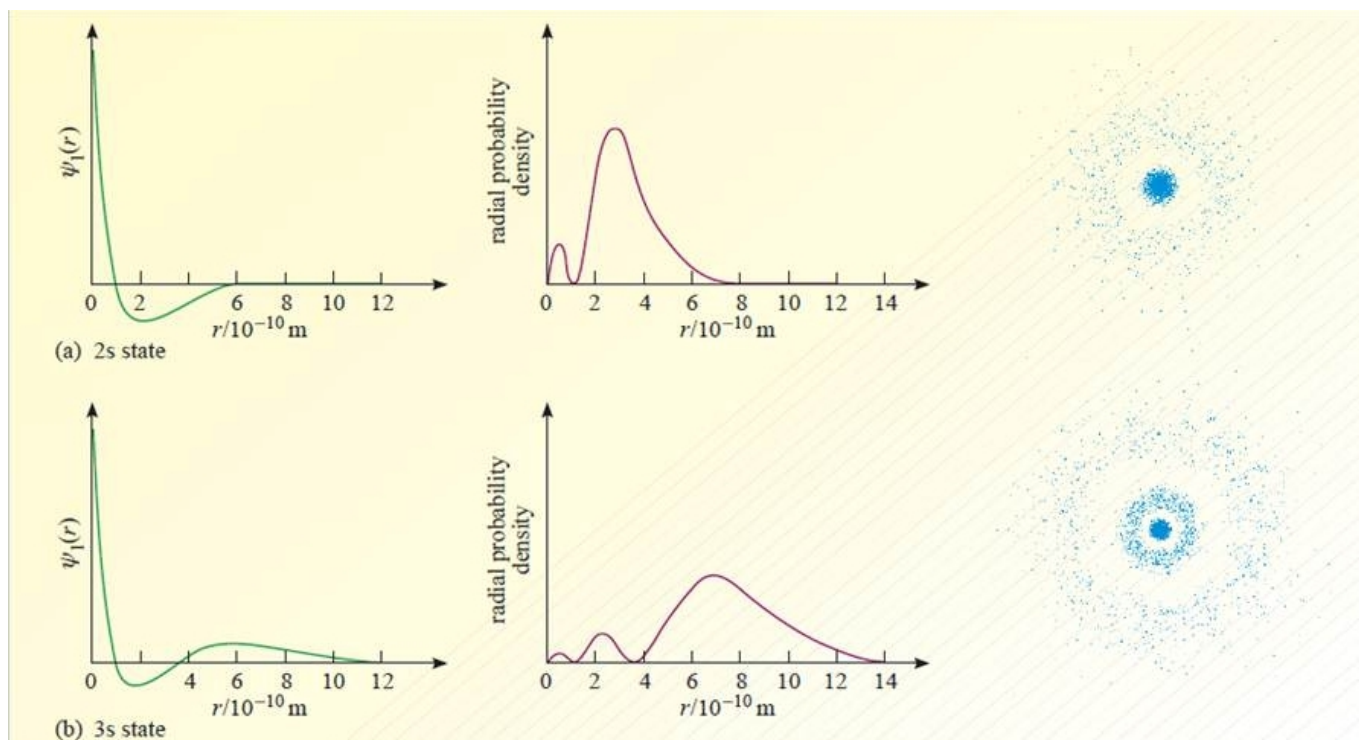


Figure 5.19 The radial wavefunctions, radial probability densities and electron cloud pictures for (a) the 2s and (b) the 3s states in hydrogen. Note that the scales of the electron cloud pictures are close to but not exactly the same as the scales for the graphs.

Atoms can make transitions between energy levels by absorbing or emitting photons of light of an appropriate energy. These radiative transitions are governed by **selection rules**, in particular, they must obey $\Delta l = \pm 1$ and $\Delta m_l = 0$ or ± 1 . Collisional transitions are not subject to the same selection rules.

The electron has an intrinsic angular momentum called spin which has magnitude S , and which is determined by its spin angular momentum quantum number s , such that

$$S = \sqrt{s(s+1)}\hbar \quad (5.78)$$

The quantum number s can take only one value: $1/2$. The z -component of the electron's spin relative to an arbitrarily defined z -axis is determined by the spin magnetic quantum number m_s such that $S_z = m_s\hbar$. The quantum number m_s has only two possible values: $+1/2$ and $-1/2$.

Schrödinger's equation can be applied to many-electron atoms but it is much more difficult to find the potential energy function. Nevertheless, the stationary states in heavy atoms can be specified by the same quantum numbers (n , l , m_l and m_s) as for the electron in a hydrogen atom. Since electrons are fermions and obey the

Pauli exclusion principle, no two electrons in an atom can have the same set of four quantum numbers. This means that they must occupy the available quantum states, filling from the lowest energy upwards and obeying Hund's rule. This states that, in the ground state of an atom, the total spin of the electrons always has its maximum possible value. This ordering in the filling up of available states gives rise to the Periodic Table of the elements.

5.9 Quantum physics of matter

In this section we summarise some results of the quantum physics of matter, focussing first on quantum gases (both bosons – like photons, and fermions – like electrons), before moving onto nuclear physics and particle physics.

5.9.1 Quantum gases

The translational energy of a molecule of mass m confined to a cubical container of side length L is quantized. The allowed energies of the translational quantum states are given by

$$E = \frac{h^2}{8mL^2}(n_1^2 + n_2^2 + n_3^2) \quad (5.79)$$

where the quantum numbers n_1 , n_2 and n_3 can be any positive integers. Each ordered set of three quantum numbers defines a translational quantum state. Most of the allowed energies are degenerate and the quantum states become more closely packed with increasing E .

The classical continuum approximation applies when the typical spacing between energy levels is small compared with kT , i.e. $h^2/8mL^2 \ll kT$ or $\lambda_{dB} \ll L$. The energy distribution of the translational quantum states is described by the **density of states** function $D(E) = B\sqrt{E}$ where $B = 2\pi V(2m)^{3/2}/h^3$. The number of quantum states with energies in the range E to $E + \Delta E$ is $D(E)\Delta E$.

The distribution of distinguishable particles amongst the allowed quantum states is given by **Boltzmann's law** which tells us that the average number of particles occupying a single quantum state of energy E is given by $F(E) = NA \exp(-E/kT)$ where N is the total number of particles, A is a constant, and the factor $\exp(-E/kT)$ is called the **Boltzmann factor**. We call $F(E)$ the **Boltzmann occupation factor**. The **Maxwell–Boltzmann energy distribution**, $G(E)$, is the product of the density of states and the Boltzmann occupation factor: $G(E) = B\sqrt{E} \times NA \exp(-E/kT)$ (see Figure 5.20). A configuration of a gas of distinguishable particles is a particular arrangement of the particles among their allowed quantum states. In thermal equilibrium all configurations are equally likely.

Boltzmann's law can be traced back to three fundamental assumptions: (i) the law of conservation of energy; (ii) the idea that all configurations are equally likely; (iii) the idea that the particles are distinguishable from one another.

In quantum mechanics, identical particles cannot be distinguished from one another. The indistinguishability of identical particles invalidates Boltzmann's law and requires a revision of the definition of a configuration. A configuration of

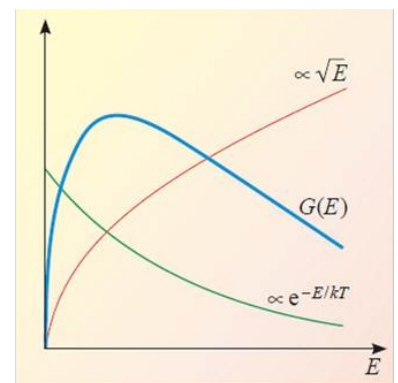


Figure 5.20 The Maxwell-Boltzmann distribution $G(E)$ is the product of two energy-dependent factors, one proportional to $\exp(-E/kT)$ and the other proportional to \sqrt{E} .

identical particles in quantum mechanics is defined by giving the numbers of particles in each quantum state. Every particle in physics is either a **boson** or a **fermion**. Identical fermions obey the Pauli exclusion principle, and so only one fermion can occupy any quantum state. Bosons do not obey the exclusion principle and so any number of bosons can occupy any quantum state. A composite particle is a fermion if it contains an odd number of fermions, or a boson if it contains an even number of fermions.

- A hypothetical gas contains three particles. Assume that each particle has three quantum states of energies 0 J, ε and 2ε . Suppose the total energy of the gas is fixed at $E_T = 3\varepsilon$. What is the probability of finding all three particles in the same quantum state if the particles are (a) distinguishable, (b) identical bosons, (c) identical fermions?
- (a) If the particles are distinguishable, they can be labelled A, B and C. There are seven configurations with total energy $E_T = 3\varepsilon$: six of these have one particle in each quantum state (so that $E_T = 0 + \varepsilon + 2\varepsilon$ in each case) and one of these has all three particles in the same state (i.e. the middle one, so that $E_T = 3 \times \varepsilon$). Since all seven configurations are equally likely, the probability of finding all three particles in the same state is $1/7$.

(b) If the particles are identical bosons, they cannot be labelled so there are only two configurations with energy 3ε : one in which there is one particle in each state, and one in which all three particles are in the middle state. Each of these configurations is equally likely so the probability of finding all three particles in the same state is $1/2$.

(c) There is no possibility of finding three identical fermions in the same state because this would contravene the Pauli exclusion principle.

The average number of identical bosons in a single quantum state of energy E is given by the **Bose occupation factor**. The Bose occupation factor for photons is

$$F_B = \frac{1}{\exp(E/kT) - 1} \quad (5.80)$$

The Bose occupation factor expresses the fact that bosons have a tendency to congregate together in the low-energy quantum states (bosons are sociable). The average number of identical fermions in a single quantum state of energy E is given by the **Fermi occupation factor**,

$$F_F = \frac{1}{\exp((E - E_F)/kT) + 1} \quad (5.81)$$

where E_F is the **Fermi energy**.

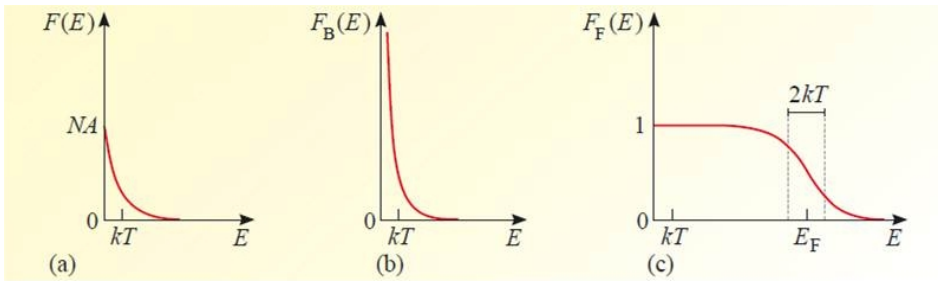


Figure 5.21 The average number of particles in a quantum state of energy E at temperature T : (a) the Boltzmann occupation factor $F(E)$, (b) the Bose occupation factor $F_B(E)$ and (c) the Fermi occupation factor $F_F(E)$.

The Fermi occupation factor is the embodiment of the exclusion principle. At $T = 0$ K, $F_F(E) = 1$ for E less than E_F and $F_F(E) = 0$ for E greater than E_F . This shows that, at $T = 0$ K, states of energy up to the Fermi energy are all occupied and those above are all empty. At higher temperatures the fall-off of $F_F(E)$ at the Fermi energy E_F is less abrupt and occurs over an energy range of a few kT , showing that a few electrons just below E_F have been excited to empty states just above E_F . The effects of indistinguishability and the exclusion principle can be neglected when the quantum states are sparsely occupied. The criterion for this is

$$N \ll \frac{V}{3\pi^2} \left(\frac{8m\pi^2 kT}{h^2} \right)^{3/2} \quad (5.82)$$

or $\lambda_{dB} \ll d$. Here λ_{dB} is the typical de Broglie wavelength and d is the typical distance between molecules. This criterion is satisfied for gases of molecules under normal conditions, but is not satisfied for gases of photons or electrons.

Thermal radiation (also called blackbody radiation or cavity radiation) is radiation that is in thermal equilibrium with matter at a fixed temperature T . A clean-cut example of thermal radiation is the radiation inside a cavity, for example an oven. Photons are bosons and so the photons of thermal radiation can be treated as a boson gas. The energy distribution law for photons, called Planck's radiation law, has the form $G_p(E) = D_p(E) \times F_B(E)$ where $D_p(E)$ is the density of states for photons and $F_B(E)$ is the Bose occupation factor. Thus Planck's radiation law is

$$G_p(E) = CE^2 \times \frac{1}{\exp(E/kT) - 1} \quad (5.83)$$

Here $C = 8\pi V/h^3 c^3 = (3.206 \times 10^{75} \text{ J}^{-3} \text{ m}^{-3})V$ and V is the volume of the cavity.

The number of photons of thermal radiation in a volume V and at a temperature T is (to two significant figures) $N = 2.4C(kT)^3 = (2.0 \times 10^7 \text{ m}^{-3} \text{ K}^{-3})VT^3$.

The energy of thermal radiation is $U = (\pi^4/15)C(kT)^4 = aVT^4$ where $a = (7.566 \times 10^{-16} \text{ J m}^{-3} \text{ K}^{-4})$ and is known as the radiation constant.

The average photon energy of blackbody radiation at a temperature T is therefore $\langle E \rangle = U/N \simeq (\pi^4/36)kT \simeq 2.7kT$.

Finally, the pressure of thermal radiation is $P = U/3V = aT^4/3 = (2.52 \times 10^{-16} \text{ J m}^{-3} \text{ K}^{-4})T^4$.

Pauli's quantum theory of the electron gas recognizes that electrons are fermions and therefore obey the exclusion principle. The energy distribution function for the electron gas, called **Pauli's distribution**, has the form $G_e(E) = D_e(E) \times F_F(E)$. Here $D_e(E)$ is the density of states for free electrons and $F_F(E)$ is the Fermi occupation factor. Thus Pauli's distribution is

$$G_e = B' \sqrt{E} \times \frac{1}{\exp((E - E_F)/kT) + 1} \quad (5.84)$$

where $B' = 4\pi V(2m_e)^{3/2}/h^3$. The Fermi energy E_F is found by equating the total number of electrons N to the total number of electron states up to E_F , at $T = 0$ K. The result is

$$E_F = \frac{h^2}{8m_e} \left(\frac{3n}{\pi} \right)^{2/3} \quad (5.85)$$

where n is the number density $n = N/V$. The total translational energy of the electron gas at $T = 0$ K is $U = \frac{3}{5}NE_F$ and the pressure of the electron gas at $T = 0$ K is $P = \frac{2}{5}nE_F$.

The electron energy distribution changes very little with temperature since only those electrons with energies within a few kT of E_F can be excited into empty states above E_F . For this reason the above values of E_F , U and P , which are calculated for $T = 0$ K, remain good approximations at higher temperatures. The free electrons contribute only slightly to the heat capacities of metals for the same reason. Another example of a Fermi gas is a neutron star where the neutron pressure prevents the star from collapsing.

As you have seen, the Pauli exclusion principle asserts that for certain particles (i.e. fermions such as electrons, protons or neutrons) no two particles of the same type in a system can occupy the *same* quantum state (i.e. they must have different combinations of quantum numbers). Since different quantum states will in general correspond to different energies, this means that there is a maximum allowed density of such particles corresponding to any given energy. Under normal conditions inside stars this is not a problem: for instance in a star like the Sun, the number of electrons per unit volume, distributed according to Maxwell–Boltzmann energy distribution, is less than the limit set by the Pauli exclusion principle at all energies. As the number density increases though, the Maxwell–Boltzmann energy distribution would normally constrain a large number of particles to relatively low energies. At a given number density, decreasing the temperature will also normally mean that more particles have relatively low energies (recall Figure 5.9). However, the Pauli exclusion principle sets an upper limit on the number of quantum states available at these low energies and consequently a large number of electrons are forced into higher energy states than they would otherwise occupy, simply because there are not enough low-energy states available. When the majority of electrons are forced into these high-energy states, the gas of electrons is said to be **degenerate** and exerts a new form of pressure known as **degeneracy pressure**. An equivalent way of visualizing the situation in a degenerate gas is that the de Broglie wavelength of the particles is larger than their separation, that is to say the particles 'overlap'. White dwarfs and neutron stars are essentially composed of degenerate gases of electrons and neutrons respectively and the bizarre behaviour these stars display is partly due to the conditions imposed by the results of quantum physics.

5.9.2 Nuclear physics

Nuclei of atoms are formed of two building blocks: **neutrons** and **protons**, and these are referred to collectively as nucleons. As noted earlier, protons carry a single unit of positive charge ($e = 1.60 \times 10^{-19}$ C), whilst neutrons have zero charge. The masses of the two types of particle are similar, with the neutron being slightly more massive than the proton: $m_p = 1.6726 \times 10^{-27}$ kg, $m_n = 1.6749 \times 10^{-27}$ kg.

An atomic nucleus is composed of Z protons and N neutrons held together by the strong nuclear force in a region of radius $r \approx (1.21 \text{ fm})A^{1/3}$ where the mass number $A = Z + N$. The atomic number Z specifies the particular chemical element and is also equal to the number of electrons in the neutral atom. A particular nuclear species is specified by a symbol ${}^A_Z X_N$, or in the non-redundant form ${}^A X$, where X denotes the chemical symbol of the element.

One aim of this section is to show you how atoms can transform from one type to another as a result of radioactive decays or nuclear fusion processes. Figure 5.22 shows a chart of the various isotopes of each element, colour-coded according to their principle mode of decay.

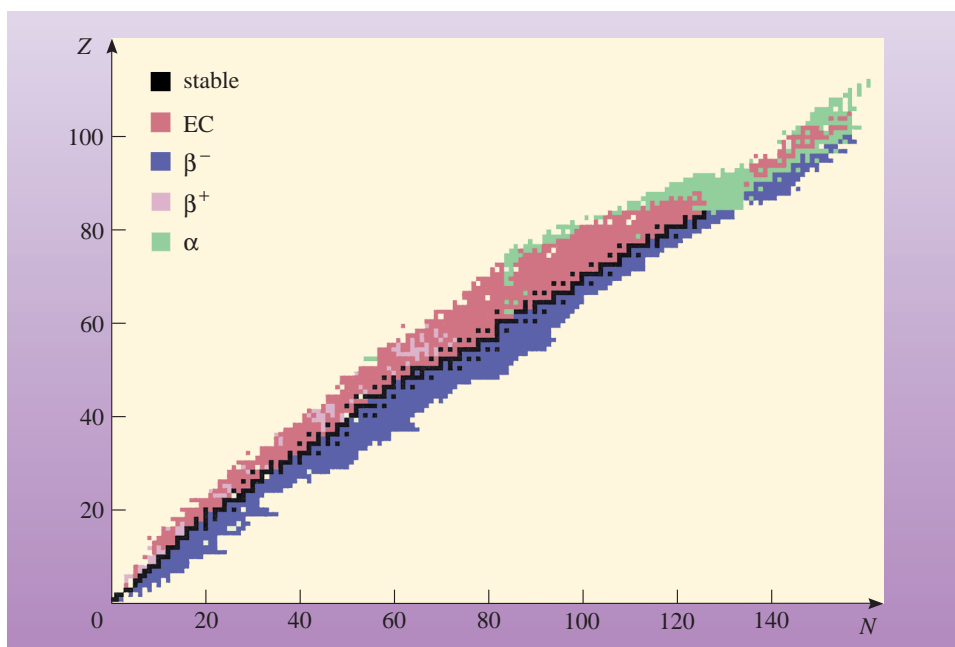


Figure 5.22 The isotopes of each element plotted on a chart of number of protons Z versus number of neutrons N . Stable nuclei are indicated by black squares, nuclei which undergo α -decay by green squares, and nuclei which undergo β^- -decay, β^+ -decay, and electron capture by blue, pink and red squares respectively.

- On Figure 5.22, where will all the isotopes of a single element lie?
- Along a single horizontal row, corresponding to a particular value of Z .

As you will see later, in all nuclear decays certain principles are obeyed:

- Electric charge is always conserved: the net charge of the products of a nuclear decay is the same as the net charge of the original nucleus.
- The mass number is conserved: the total number of nucleons in the products is the same as that in the original nucleus.
- As in all physical processes, energy is conserved.

The one complication here though, is that since the energies involved in nuclear decays are so large, we need to take account of the relationship between energy and mass, $E = mc^2$. Since nuclear energies are generally measured in units of MeV or GeV, convenient units in which to measure nuclear masses are MeV/c^2 or GeV/c^2 . In these units, the mass of a proton is $938.3 \text{ MeV}/c^2$ and that of a neutron is $939.6 \text{ MeV}/c^2$, or around $1 \text{ GeV}/c^2$ in each case.

Radioactive decay is governed by the exponential decay law

$$N(t) = N_0 \exp(0.693t/T_{1/2}) \quad (5.86)$$

where $0.693 = \log_e(2)$ to 3 decimal places, N_0 is the initial number of radioactive nuclei and $N(t)$ is the average number left at time t . The **half-life**, $T_{1/2}$, characterises the nucleus and decay mode. The number of undecayed nuclei is halved during any period equal to the half-life.

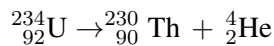
The **binding energy** B of a nucleus is the minimum energy required to disassemble the nucleus into its constituent nucleons; it is also equal to the energy released when the nucleus is formed from its constituents. The binding energy per nucleon, B/A , is a measure of the relative stability of nuclei. B/A for most nuclei is within 10% or so of 8 MeV per nucleon, with a maximum value of about 8.8 MeV per nucleon near $A = 56$ (iron and nickel), then falling off slowly for heavier nuclei and falling off quite rapidly for the light nuclei. This means that energy can be released when heavy nuclei undergo fission or when light nuclei undergo fusion. When the values of the total energy per nucleon ($-B/A$) for all known nuclei are plotted as points above the ZN plane they lie near a surface having the shape of a valley called the valley of stability, with the valley floor lying directly above the path of stability. Then the β^- -decay of neutron-rich isobars and the β^+ -decay (or electron capture) of proton-rich isobars can be thought of as streams running down the valley walls, while α -decay of heavy nuclei is like a stream running down the valley floor.

The semi-empirical model of the nucleus incorporates the short-range strong nuclear force and the repulsive Coulomb force and is based on an analogy between a nucleus and a charged liquid drop. It identifies four contributions to B/A : a volume energy, a surface energy, a Coulomb energy and a symmetry energy. Using empirical parameters, the model fits the overall trend of measured B/A values quite well but there are discrepancies in the small-scale structure suggestive of shell effects. In the nuclear shell model each proton and neutron is represented by a wavefunction and occupies an energy level in a potential energy well produced by the other nucleons. The potential energy wells for neutrons follow the nuclear matter densities, while those for protons have contributions from Coulomb repulsion, resulting in a bump or barrier at the nuclear edge, and wells that are less deep than the corresponding neutron wells. Nuclei with magic numbers of nucleons, corresponding to closed shells, are particularly stable. Because of pairing, nuclei with even numbers of protons and neutrons tend to be more stable than those with odd numbers. The proton and neutron energy levels are filled starting with the lowest energies and subject to the Pauli exclusion principle. Nuclear stability favours the filling of the neutron wells and the shallower proton wells to the same energy. This explains the neutron excess in medium and heavy nuclei and the tendency for isobars far from the valley floor to exhibit β -decay.

α -decay

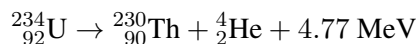
The α -particle is simply the nucleus of the helium atom, with mass number $A = 4$ and atomic number $Z = 2$. It consists of $Z = 2$ protons and $A - Z = 4 - 2 = 2$ neutrons. It is a very tightly bound arrangement: this ground state for two protons and two neutrons has an energy that is about 28 MeV lower than the energy of the four free nucleons.

In some cases, it is energetically favourable for a nucleus of mass number A and atomic number Z to emit an α -particle, thereby producing a new nucleus, with mass number $A - 4$ and atomic number $Z - 2$. A case in point is the unstable isotope of uranium ${}^{234}_{92}\text{U}$, containing 92 protons and 142 neutrons. It undergoes α -decay (**alpha-decay**) to produce an isotope of thorium, with 90 protons and 140 neutrons



Exercise 5.12 The nucleus of thorium-230 subsequently undergoes four more α -decays. What isotope of lead results? ■

You have seen that electric charge and mass number are conserved in an α -decay process, but what of energy conservation? Well, nuclear decays will clearly involve changes in energy, much as atomic transitions that you saw earlier. For example, the α -decay of ${}^{234}_{92}\text{U}$ liberates 4.77 MeV of kinetic energy, carried away (almost exclusively) by the α -particle. We can write the decay as



and making use of Einstein's mass–energy relation, we have the mass equation

$$(\text{mass of } {}^{234}_{92}\text{U}) = (\text{mass of } {}^{230}_{90}\text{Th}) + (\text{mass of } {}^4_2\text{He}) + 4.77 \text{ MeV}/c^2$$

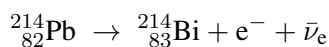
So a liberated energy of 4.77 MeV is produced by mass loss of $4.77 \text{ MeV}/c^2$. To appreciate how substantial this change in mass is, it may be compared with the mass of a proton, $m_p \approx 1 \text{ GeV}/c^2$. So, in the α -decay above, the decrease in mass is about 0.5% of the proton's mass, and hence about 0.002% of the mass of the original uranium nucleus, with $A = 234$.

Quantum-mechanical tunnelling is an essential mechanism for many nuclear reactions involving charged particles. α -decay occurs when α -particles inside nuclei tunnel to the outside through the Coulomb barrier. The tunnelling probability depends very sensitively on the energy of the α -particle relative to the top of the barrier. This explains the huge range of α -decay lifetimes corresponding to a small range of α -particle energies.

 β -decay and electron capture

The usual type of β -decay (**beta-decay**) involves the emission of an electron from the nucleus of an atom. The process occurs when a neutron in the original nucleus transforms into a proton, so *increasing* the atomic number by one. For reasons that are not important here, another particle is created in the beta-decay process too. It is called the electron antineutrino and it has zero electric charge. Creation

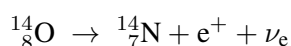
of an electron and an electron antineutrino occurs in what is called β^- -decay (**beta-minus decay**), the minus sign indicating that the electron is negatively charged. A nucleus that undergoes β^- -decay is the unstable lead isotope $^{214}_{82}\text{Pb}$ which transforms into a stable bismuth isotope $^{214}_{83}\text{Bi}$. The decay in this case can be represented as



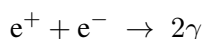
The rather clumsy symbol $\bar{\nu}_e$ represents an electron antineutrino which has zero charge. The subscript e indicates that it is associated with an electron, and the bar over the top of the letter indicates that it is an antiparticle.

Exercise 5.13 A nucleus of the unstable nitrogen isotope $^{16}_7\text{N}$ undergoes β^- -decay. Write down an expression for this nuclear decay, indicating what nucleus is formed as a result.

The process described above is only half of the story as far as beta-decay is concerned. There is a very closely related process, called β^+ -decay (**beta-plus decay** or sometimes *inverse beta decay*), in which a positively charged particle, called a positron, is created, along with an electron neutrino, which has zero charge. In this process, a proton in the original nucleus transforms into a neutron, so *decreasing* the atomic number by one. A nucleus that undergoes β^+ -decay is the unstable oxygen isotope $^{14}_8\text{O}$ which transforms into a stable nitrogen isotope $^{14}_7\text{N}$. The decay in this case can be represented as

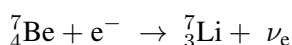


Here, the symbol e^+ is used to represent the positron (also known as an antielectron) and ν_e is the electron neutrino. The observable consequence of β^+ -decay is that the positron produced will immediately combine with an electron to produce gamma-rays:



Exercise 5.14 A nucleus of the unstable phosphorus isotope $^{30}_{15}\text{P}$ undergoes β^+ -decay. Write down an expression for this decay process, indicating what nucleus is formed as a result.

Related to β^+ -decay is the process of **electron capture**, in fact the outcome is virtually the same. Some nuclei which have too many protons, rather than undergoing β^+ -decay, instead capture an electron into the nucleus. As a result, a proton is transformed into a neutron and an electron neutrino is emitted. A nucleus that undergoes electron capture is the unstable beryllium isotope ^7_4Be which transforms into a stable lithium isotope ^7_3Li . The decay in this case can be represented as



As in β^+ -decay, the atomic number decreases by one and the mass number remains the same.

γ -decay

The final type of nuclear decay that we consider here is γ -decay (**gamma-decay**). In contrast to the two processes of α -decay and β -decay, this involves no change in the numbers of neutrons and protons. Gamma-decay occurs when a nucleus finds itself in an excited state. A quantum jump down to the ground-state configuration of the same number of neutrons and protons is accompanied by the emission of a photon, as with electron transitions in atoms. This time however, the photon energy is around a million times larger – it is a gamma-ray photon. Such excited states of nuclei may be created in the process of α -decay or β -decay, or by the collisions of nuclei at high kinetic energies.

Exercise 5.15 The unstable isotope of caesium ${}_{55}^{137}\text{Cs}$ undergoes β^- -decay to produce an excited state of the barium isotope ${}_{56}^{137}\text{Ba}$. The barium nucleus then decays to its ground state with the emission of a gamma-ray photon of energy 662 keV. What are the atomic number and mass number of the barium nucleus *after* the γ -decay?

**Nuclear fusion**

A final type of nuclear process which is important in astrophysics is that of nuclear fusion. In fact it is processes of this kind that provide the energy source for stars. Many nuclear reactions between charged particles occur by the particles tunnelling inwards through their mutual Coulomb barrier so that they can interact by the strong nuclear force. Fusion reactions in the Sun can only occur at the prevailing temperatures by tunnelling.

As noted above, certain configurations of nucleons are energetically more favourable than others. Hence, by forcing certain nuclei sufficiently close together (overcoming their mutual electrical repulsion), they may fuse to form a single more massive nucleus which is at a lower energy than the original two nuclei. Once again, an example should make the process clear.

A nuclear fusion process which occurs in the majority of stars is the fusion of two nuclei of helium-3. It may be represented as



The mass of each nucleus of helium-3 is $2808 \text{ MeV}/c^2$, whilst the mass of a nucleus of helium-4 is $3727 \text{ MeV}/c^2$ and the mass of a proton is $938 \text{ MeV}/c^2$. Simply adding up these masses, there is a deficit of $13 \text{ MeV}/c^2$ on the right-hand side. This lost mass appears as energy, liberated by the fusion process.

To initiate such reactions, high temperatures (of the order 10^7 K or higher) are required to provide the necessary kinetic energy to overcome the electrical repulsion between nuclei (all of which are positively charged). Such conditions generally only occur in the cores of stars.

Fusion can liberate energy when various light nuclei are combined together. So, in the cores of stars, nuclear fusion processes liberate energy to power the star and also convert light elements into heavier ones. The limiting mass at which the process ceases to be energetically favourable is around that of the nuclei of iron,

cobalt and nickel. So, nuclear fusion provides an energy source for stars only until their cores are composed of nuclei such as iron. For many stars, nuclear fusion will end long before iron is formed as the temperatures in their cores are not sufficient to trigger fusion reactions beyond, say, helium or carbon. The life cycles of stars are closely dependent on these energetic processes.

Exercise 5.16 The triple-alpha process involves the fusing together of three helium-4 nuclei to form a single nucleus of carbon-12. Show that, whilst the fusion of two helium-4 nuclei to form a nucleus of beryllium-8 involves a slight energy deficit, the subsequent fusion of a beryllium-8 nucleus with another helium-4 nucleus is energetically favoured. (You may assume: mass of ${}^4_2\text{He} = 3.7274 \text{ GeV}/c^2$, mass of ${}^8_4\text{Be} = 7.4549 \text{ GeV}/c^2$, mass of ${}^{12}_6\text{C} = 11.1749 \text{ GeV}/c^2$.)



5.9.3 Particle physics

There are four **fundamental forces** or interactions through which particles can interact: the strong, weak, electromagnetic and gravitational interactions. There are exchange particles associated with each of the fundamental forces: the photon (for electromagnetism), the W^+ , W^- and Z^0 particles (for the weak interaction), and eight kinds of gluon (for the strong interaction). These are all spin 1 particles.

The strong, electromagnetic and weak interactions of the fundamental particles (and their antiparticles) may be described theoretically by quantum field theories in which the forces are mediated by exchange particles. These quantum field theories have been developed into the standard model of particle physics. This is not thought to be the final theory, but its construction is regarded as a major triumph since, by using **Feynman diagrams** and other techniques, it permits the evaluation of measurable quantities such as cross-sections and mean lifetimes. Going beyond the standard model might require the formulation of a grand unified theory, or a string theory that might involve supersymmetry.

Today, it is believed that there are two families of fundamental particles, called **leptons** and **quarks**. By fundamental we mean that there is no evidence that these particles are composed of smaller or simpler constituents. There are just six leptons and six quarks, together with an equal number of their antiparticles. All the familiar forms of matter are ultimately composed of these particles.













	1st generation	2nd generation	3rd generation
leptons with charge $-e$			
leptons with charge 0			
(a)			
	1st generation	2nd generation	3rd generation
quarks with charge $+\frac{2}{3}e$			
quarks with charge $-\frac{1}{3}e$			
(b)			

Figure 5.23 (a) The three generations of leptons and their electric charges. (b) The three generations of quarks and their electric charges.

The six different types are often referred to as different flavours of lepton, and the three pairs are said to represent the three generations of leptons. The first generation consists of the familiar electron (e) and its β -decay partner, the electron neutrino (ν_e). The second pair of leptons consists of the muon (μ) and another type of neutrino called a muon neutrino (ν_μ). The muon is similar to the electron except that it is about 200 times heavier and unstable with a fairly long lifetime of a few microseconds. The third generation of leptons consists of a particle called a tauon (τ) and a third type of neutrino called a tauon neutrino (ν_τ). The tauon is similar to and even heavier than the muon and has a much shorter lifetime. These two heavier leptons, being unstable, are not normally constituents of matter, but are created in high-energy collisions between other subatomic particles. Associated with these six leptons are the six antileptons, particles of antimatter. These include the positron (e^+) which is the antiparticle of the electron and the electron antineutrino ($\bar{\nu}_e$). Leptons are spin $1/2$ particles and have lepton number $L = 1$ whilst the corresponding family of 6 antileptons have $L = -1$. Leptons feel the weak interaction and the charged ones also feel the electromagnetic interaction, but leptons do not feel the strong interaction.

The pattern of the leptons is repeated for the quarks. The six types (or flavours) of quark are labelled (for historical reasons) by the letters u, d, c, s, t and b, which stand for up, down, charm, strange, top and bottom. Like the leptons, the quarks are paired off in three generations on the basis of their mass. To each quark, there corresponds an antiquark, with the opposite electric charge and the same mass. The antiquarks are denoted by \bar{u} , \bar{d} , \bar{c} , \bar{s} , \bar{t} and \bar{b} . Unlike leptons, the quarks and antiquarks have never been observed in isolation. They only seem to occur bound together in combinations held together by gluons. For example, the familiar proton is a combination of two up quarks and a down quark, which we can write as uud. Note that each up quark carries an electric charge of $2e/3$ and a down quark carries a charge of $-e/3$, so the combination uud does indeed give a net electric charge equal to the charge e on a proton. Similarly, a neutron is the combination udd which has a net electric charge of zero. Observable particles

consisting of combinations of quarks are collectively called **hadrons** and there are literally hundreds of them, the proton and neutron being the most familiar. There are three recipes for building hadrons from quarks: A hadron can consist of: three quarks (in which case it is called a **baryon**); three antiquarks (in which case it is called an **antibaryon**); or one quark and one antiquark (in which case it is called a **meson**). Baryons have half odd-integer spin ($1/2$, $3/2$, etc.) and baryon number $B = 1$ whilst the corresponding family of antibaryons have $B = -1$. Mesons have zero or integer spin (0 , 1 , 2 , etc.) and baryon number $B = 0$. Baryons and mesons can interact through the strong interaction as well as the weak interaction and, if charged, the electromagnetic interaction.

When particles collide, certain quantities such as total electric charge, total (relativistic) momentum and total (relativistic) energy are always conserved. The particles may undergo elastic collisions where the colliding particles simply exchange energy and momentum, or inelastic collisions where (relativistic) kinetic energy may be converted into rest energy and so the nature and number of particles may change. The likelihood of a particle being scattered in a given process at a specified energy is described by a measurable quantity called a **cross-section**. Cross-sections are measured in barns ($1 \text{ barn} = 10^{-28} \text{ m}^2$).

5.10 Electromagnetism

As noted earlier, virtually the only information about the Universe which we receive here on Earth is that which arrives in the form of electromagnetic radiation. In order to understand the processes occurring in stars and galaxies it is therefore vital to have an appreciation of the ways in which electromagnetic radiation arises and interacts with matter. To understand electromagnetic radiation, we first consider the nature of electricity and magnetism.

5.10.1 Electricity and magnetism

The topic of electricity and magnetism forms a huge part of physics, and it is therefore important in many areas of astrophysics and cosmology. At the root of many electromagnetic phenomena are the key results that static electric charges give rise to electric forces, whilst moving electric charges (i.e. currents) give rise to magnetic forces.

A crucial equation is **Coulomb's law**, discovered by Charles Augustin de Coulomb in 1785. Following a similar formulation to Newton's law of gravity, it may be expressed as follows. Two particles of unlike (or like) electric charge, at rest, separated by a distance r , attract (or repel) each other with an electrostatic force that is inversely proportional to the square of their separation and is proportional to the product of the charges. Coulomb's law of force between charged particles situated in a vacuum may be expressed in the form of the vector equation

$$\mathbf{F}_{21} = \frac{q_1 q_2}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \quad (5.87)$$

where q_1 and q_2 are the electric charges of the two particles and ϵ_0 is a constant known as the permittivity of free space with a value $8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$,

\hat{r} is a unit vector directed from q_1 towards q_2 . Electric charges are measured in the SI unit of coulomb (symbol C) and the charge of a single electron, denoted by $-e$, is -1.602×10^{-19} C.

Electric field $\mathcal{E}(\mathbf{r})$ is a vector quantity, defined at all points in space, such that, at any particular point \mathbf{r} , its value is given by the electrostatic force per unit charge at that point. Equivalently, the electrostatic force experienced by a particle whose charge is of magnitude q is given by

$$\mathbf{F}_{\text{el}} = q\mathcal{E}(\mathbf{r}) \quad (5.88)$$

where $\mathcal{E}(\mathbf{r})$ is the electric field. The direction of \mathbf{F}_{el} is parallel to the direction of \mathcal{E} at the point in question. The two vectors will be in the same direction if the charge is positive, and in the opposite direction if the charge is negative.

Magnetic field is also a vector quantity denoted by $\mathbf{B}(\mathbf{r})$ and measured in the SI unit of tesla (symbol T). Magnetic fields are generated by, and give rise to forces which act upon, *moving* electric charges. In particular, the magnetic force on a particle moving with velocity \mathbf{v} and whose charge is of magnitude q , passing through a magnetic field $\mathbf{B}(\mathbf{r})$, is given by

$$\mathbf{F}_{\text{mag}} = q\mathbf{v} \times \mathbf{B}(\mathbf{r}) \quad (5.89)$$

where the direction of the force is found using the right-hand rule, as with all vector products.

Combining the equations for electric and magnetic force, gives a general equation known as the **Lorentz force law**, for the force on a charged particle in a region containing both electric and magnetic fields:

$$\mathbf{F} = q[\mathcal{E}(\mathbf{r}) + \mathbf{v} \times \mathbf{B}(\mathbf{r})] \quad (5.90)$$

The direction of the magnetic force vector is always at right angles to the plane containing the velocity vector of the charged particle and the magnetic field vector, as shown in Figure 5.24a. The magnetic force will therefore give rise to an acceleration of the particle (by Newton's second law) which is at right angles to the original direction of motion. So if the direction of motion of the charged particle is at right angles to a uniform magnetic field, the particle will trace out a circular path orbiting around the magnetic field lines (Figure 5.24b) as its direction of motion is constantly altered by the action of the magnetic force. For an arbitrary initial direction of motion, a charged particle will travel in a *helical* path around magnetic field lines (Figure 5.24c) as it travels through space.

Worked Example 5.2

What is the radius of the circular path traced out by a charged particle of charge q and mass m travelling at a speed v at right angles to a magnetic field of strength B ?

Solution

The magnitude of the magnetic force (qvB) must be equated to the magnitude of the centripetal force acting on the particle, where the magnitude of the centripetal force is equal to the mass of the particle (m) multiplied by the magnitude of the centripetal acceleration (v^2/r),

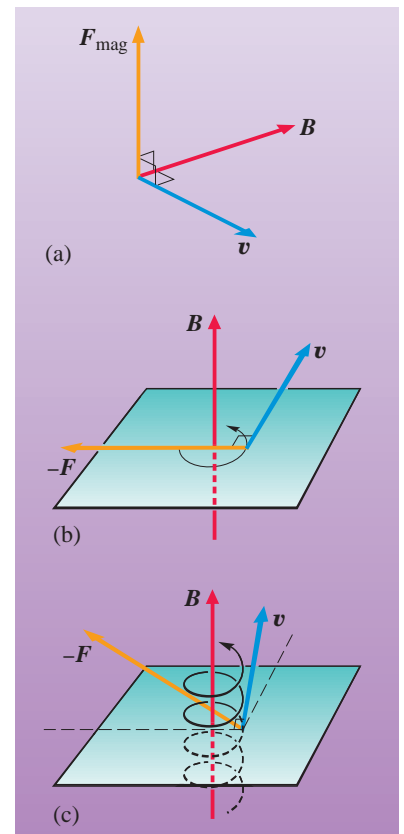


Figure 5.24 (a) The direction of the magnetic force \mathbf{F} experienced by a moving charged particle is at right angles to the plane containing the magnetic field vector \mathbf{B} and the velocity vector \mathbf{v} of the particle. (b) For motion at right angles to the magnetic field, the particle will travel in a circle. (c) In general, a charged particle passing through a region of uniform magnetic field will travel in a *helical* path around magnetic field lines.

therefore $mv^2/r = qvB$. Rearranging this gives the radius of the circle as $r = mv/qB$. This is known as the **cyclotron radius** of the particle.

Exercise 5.17 (a) What is the cyclotron radius of an electron travelling with speed $3.0 \times 10^6 \text{ m s}^{-1}$ (i.e. 1% of the speed of light) at right angles to the magnetic field of a pulsar whose strength is $3.0 \times 10^8 \text{ T}$? (b) What is the frequency at which the electron completes its orbits? (Assume $m_e = 9.1 \times 10^{-31} \text{ kg}$ and $-e = -1.6 \times 10^{-19} \text{ C}$.)

The frequency at which the electron orbits the magnetic field lines will correspond to the fundamental frequency of the cyclotron radiation produced; this is known as the **cyclotron frequency**. An electron travelling through a magnetic field of strength a few hundred megatesla will therefore radiate electromagnetic radiation with a frequency of nearly 10^{19} Hz , which corresponds to the X-ray part of the spectrum, as you will see in the next Section.

5.10.2 Electromagnetic waves

The discussion earlier about atoms referred to light in terms of ‘particles’ of electromagnetic radiation called photons. This was an appropriate description because we were concerned with the interaction of light with matter (i.e. its absorption or emission by atoms). However, although light interacts with matter as though it’s composed of a stream of particles, light propagates like a wave.

The **wavelength** of a wave is the distance between two similar points on the wave profile, and it is given the symbol λ . The **frequency** of a wave is the number of cycles of the wave that pass a given point in one second, and in astrophysics it is usually given the symbol ν . The wavelength and frequency of an electromagnetic wave are related by the equation

$$c = \lambda\nu \quad (5.91)$$

where c is the speed of light and has a value of $3.00 \times 10^8 \text{ m s}^{-1}$.

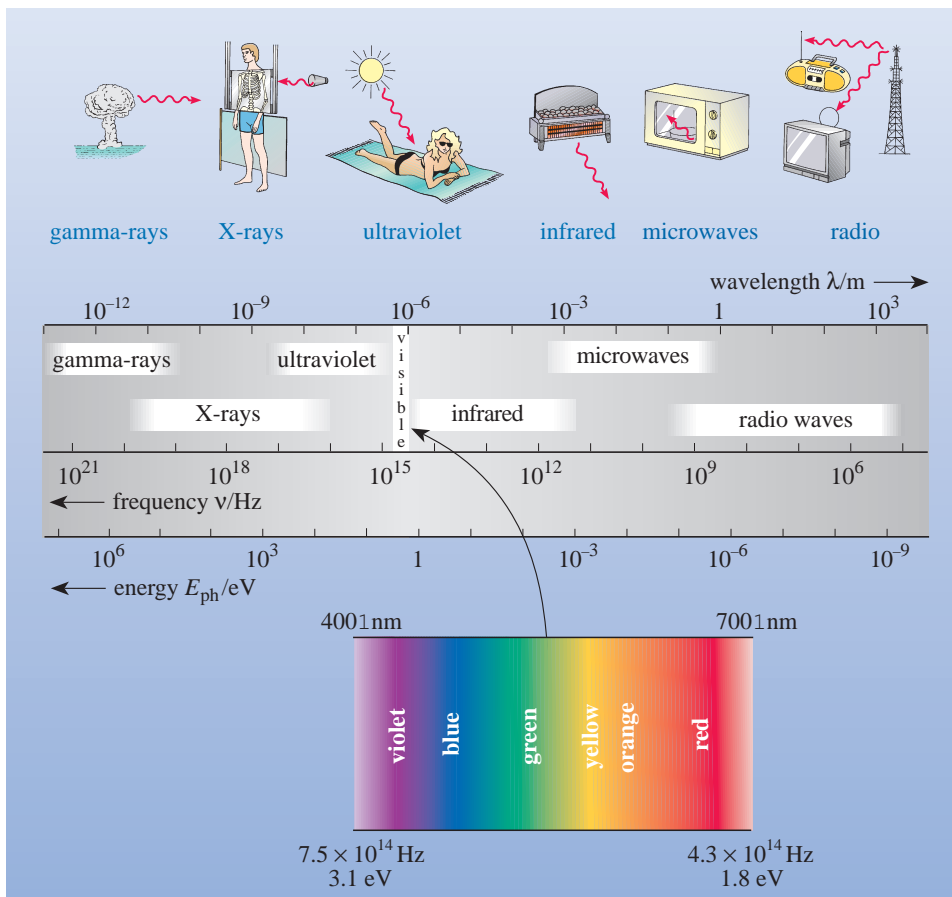


Figure 5.25 The electromagnetic spectrum showing the wavelengths, frequencies and photon energies appropriate to each region.

Different regions of the electromagnetic spectrum are distinguished by the different wavelengths and frequencies of radiation (Figure 5.25). The longest wavelength, lowest frequency, radiation is referred to as radio waves. Moving to shorter wavelengths and higher frequencies the radiation is referred to as microwaves, infrared radiation, (visible) light, ultraviolet radiation, and X-rays. The radiation with the shortest wavelength and highest frequency is referred to as gamma-rays.

The link between the photon and wave pictures of electromagnetic radiation is provided by the relationship that each photon carries an amount of energy E_{ph} that is determined by the frequency or wavelength of the radiation used to characterize its propagation, namely

$$E_{ph} = h\nu = \frac{hc}{\lambda} \quad (5.92)$$

where h is Planck's constant and equal to 6.63×10^{-34} J s or 4.14×10^{-15} eV Hz $^{-1}$ in alternative units. Related to this is the expression for the momentum of a photon. Since photons have zero mass, their momentum cannot be given by a relationship like $p = mv$. Instead, the relationship for the magnitude of a photon's momentum is

$$p_{ph} = \frac{E_{ph}}{c} = \frac{h\nu}{c} = \frac{h}{\lambda} \quad (5.93)$$

Exercise 5.18 What is the longest wavelength spectral line corresponding to

each of the (a) Lyman and (b) Paschen series of the hydrogen spectrum (see Figure 5.15), and in which regions of the electromagnetic spectrum do these lines occur? (Assume $h = 4.14 \times 10^{-15} \text{ eV Hz}^{-1}$ and $c = 3.00 \times 10^8 \text{ m s}^{-1}$.)

Since photons possess momentum, they can exert a pressure. **Radiation pressure** is the pressure which electromagnetic radiation exerts on a surface. For blackbody radiation in equilibrium with a surface, the radiation pressure exerted on the surface is given by

$$P_{\text{rad}} = \frac{4\sigma T^4}{3c} \quad (5.94)$$

where σ is the Stefan-Boltzmann constant ($5.671 \times 10^{-8} \text{ J s}^{-1} \text{ m}^{-2} \text{ K}^{-4}$) and c is the speed of light.

A final feature of all electromagnetic radiation is that it can be polarized. Electromagnetic radiation propagates as mutually interacting electric and magnetic fields (see Section 5.10.1), and a snapshot of such a wave is shown in Figure 5.26. The electric field and magnetic field are each perpendicular to the direction of propagation, so the wave is said to be transverse. In this case we are supposing that the electric field is confined to the vertical plane, and that its direction and magnitude at each point along the propagation axis are indicated by the orange curve. The magnetic field at each point along the propagation axis is indicated by the green curve.

Polarization occurs when there is a restriction placed on the direction in which the vibrations in such a wave can take place. The vibrations *must* be at right angles to the direction of propagation, but that still leaves them free, in principle, to take up a variety of orientations. Two such orientations are shown in Figure 5.27, but there are, in principle, an infinite number of other orientations that are intermediate between these extremes. Figure 5.27a shows the electric field vibrating in the vertical plane and in Figure 5.27b it vibrates in the horizontal plane. In each case the wave is said to be plane (or linearly) polarized along the direction of oscillation.

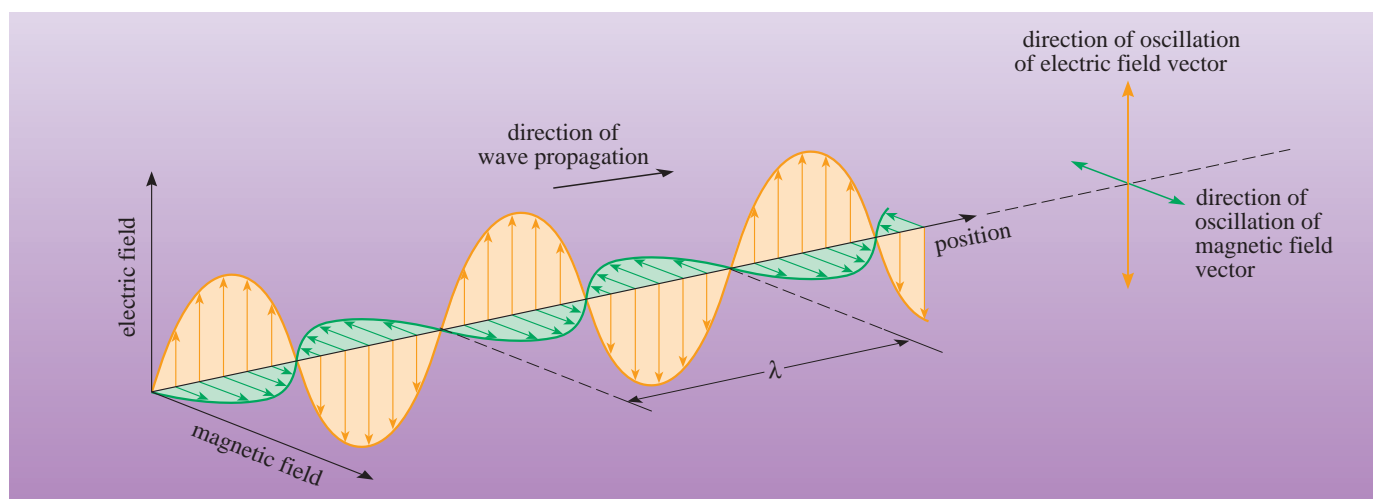


Figure 5.26 The variations of electric and magnetic fields in an electromagnetic wave.

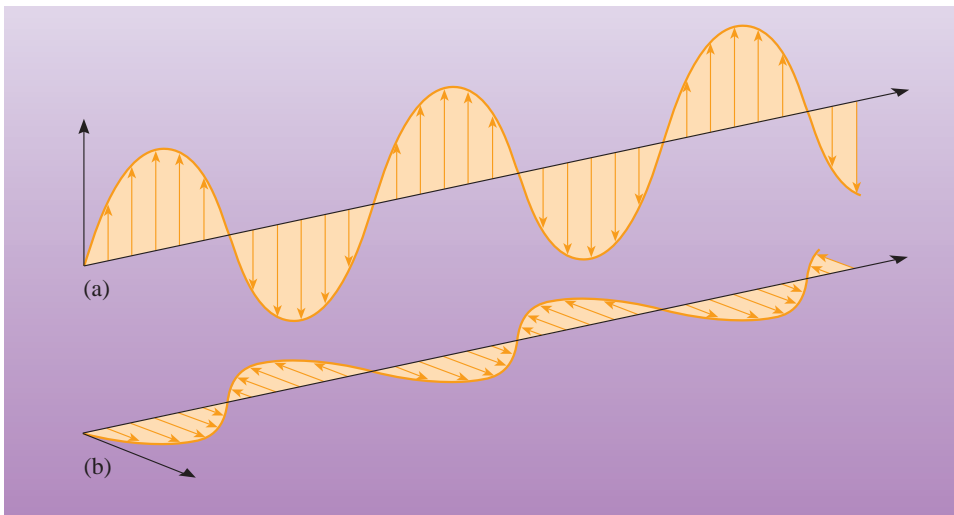


Figure 5.27 A plane polarized electromagnetic wave: vibrating (a) in a vertical plane, (b) in a horizontal plane.

Light and other electromagnetic radiation is generally unpolarized, i.e. it is a superposition of a large number of polarized waves that have electric fields in every possible direction perpendicular to the direction of propagation. This is because each atom emits light quite independently of the surrounding atoms, and though the light emitted from an individual atom is polarized, when the light from a vast number of atoms is combined, there is no overall preferred polarization direction. Some other processes by which electromagnetic radiation is emitted however do give rise to polarized electromagnetic radiation.

5.10.3 Spectra

There are three basic types of spectra, as shown in Figure 5.28, encountered in astrophysics. **Continuous spectra** can be produced by hot objects in which there are many energy levels with extremely small separations between them. These levels form a continuous energy band within which transitions are possible, so a distribution of photons with a continuous range of energies is emitted. If light from such a source passes through a gas then photons of some particular energies are absorbed by the gas, and are then emitted in all directions. Thus if we look towards the gas in a direction other than towards the source, we see the photon energies emitted by the gas as a set of bright lines on a dark background. This is an **emission line spectrum**. If we look towards the gas in the direction of the source, the continuous spectrum displays dark lines at the absorbed photon energies. This is an **absorption line spectrum**.

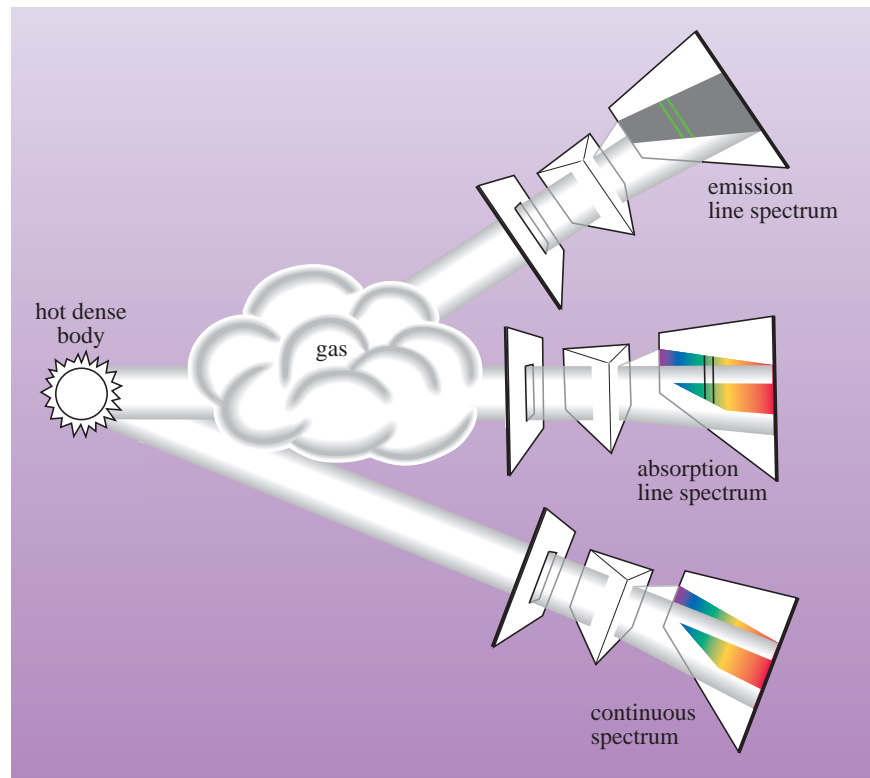


Figure 5.28 Three types of spectra encountered in astrophysics. In many cases, real spectra are superpositions of these types.

- Why is the spectrum of the Sun (and other stars) an absorption line spectrum?
- A continuous spectrum is produced deep inside the Sun (or other star) but as the photons emerge through the cooler outer layers, some photon energies are absorbed by the atoms present in the photosphere.

The continuous spectra emitted by many hot objects resemble so-called **black-body** spectra. Black-body spectra all have similar shapes (Figure 5.29) with the property that the higher the temperature of the object, the higher the frequency (or shorter the wavelength) at which the peak intensity in the black-body distribution occurs. The name arises because a 'black-body' is one which absorbs all the radiation falling on it and is also a perfect emitter of radiation. As you saw earlier, the spectra of stars are approximately those of black-bodies (with absorption, and sometimes emission, lines superimposed).

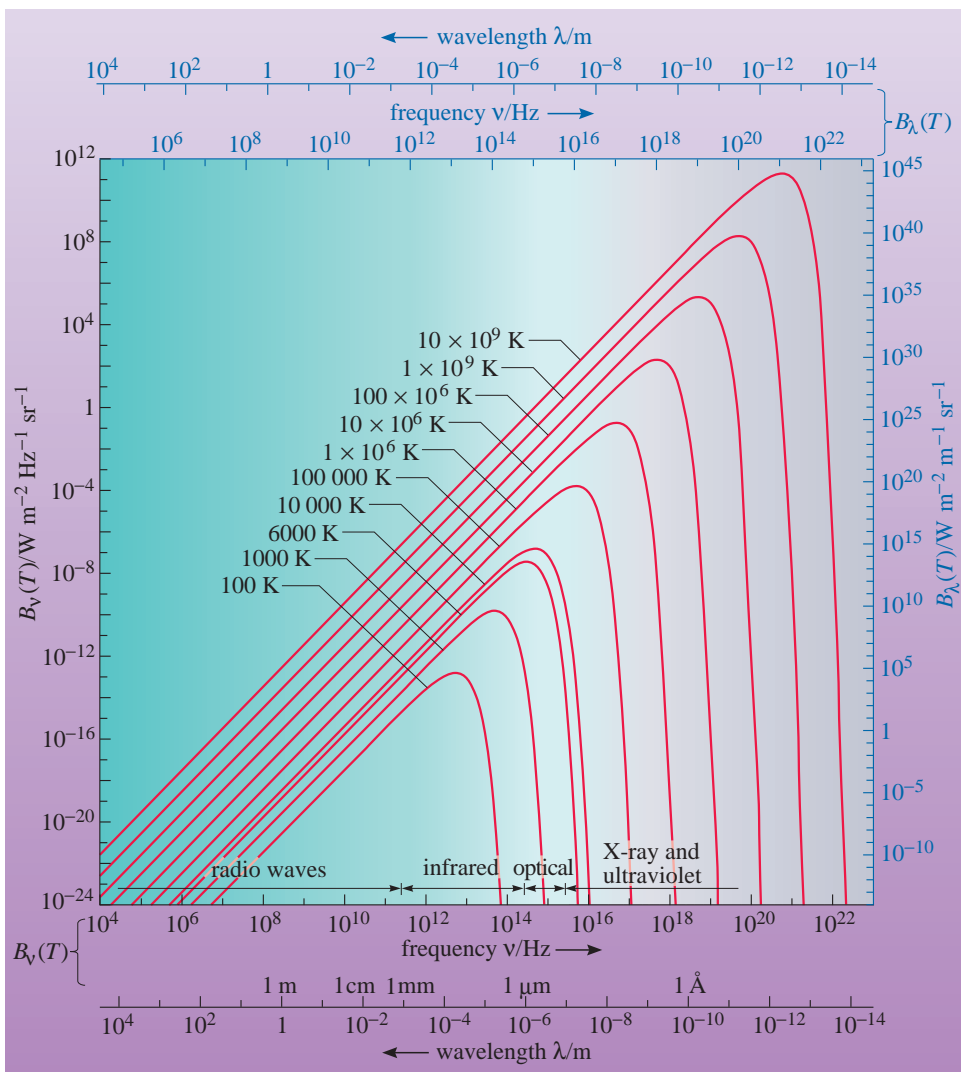


Figure 5.29 Black-body spectra corresponding to different temperatures.

Black-body radiation has a continuous distribution of photon energies and the graph of the spectrum has a characteristic shape. The shape of the spectrum is described by the **Planck function**, which can be written in one of two ways. (These are related to the two forms of spectral flux density which you read about in Section 3.4.) The power per unit area per unit frequency (or per unit wavelength) per unit solid angle is:

$$B_{\nu}(T) = \left(\frac{2h\nu^3}{c^2} \right) \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} \text{ W m}^{-2} \text{ Hz}^{-1} \text{ sr}^{-1} \quad (5.95)$$

$$B_{\lambda}(T) = \left(\frac{2hc^2}{\lambda^5} \right) \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1} \text{ W m}^{-2} \text{ m}^{-1} \text{ sr}^{-1} \quad (5.96)$$

where k is Boltzmann's constant, h is Planck's constant, c the speed of light, λ and ν are the wavelength and frequency in question, and T is the temperature of the black-body source.

Exercise 5.19 Verify that $\nu B_\nu(T) = \lambda B_\lambda(T)$ for all temperatures, wavelengths and frequencies.

Although the Planck function formulae look quite fearsome, you will rarely need to use them explicitly. There is a much simpler equation, known as the **Wien displacement law**, which describes the wavelength at which a black-body spectrum reaches a peak, and this is often more useful. In the two representations, the maximum value of B_ν occurs at a wavelength determined by

$$\lambda_{\max} T = 5.1 \times 10^{-3} \text{ m K (maximizing } B_\nu) \quad (5.97)$$

and the maximum value of B_λ occurs at a wavelength determined by

$$\lambda_{\max} T = 2.9 \times 10^{-3} \text{ m K (maximizing } B_\lambda) \quad (5.98)$$

Another useful expression is that for the mean photon energy in the black-body spectrum. Irrespective of whether a $B_\nu(T)$ or a $B_\lambda(T)$ function is used, the mean photon energy (as noted earlier) is

$$\langle E_{\text{ph}} \rangle = 2.7kT \quad (5.99)$$

Exercise 5.20 A particular black-body spectrum is produced by a body at a temperature of 10^8 K.

- What is the photon energy in electronvolts corresponding to the wavelength at which B_ν reaches a maximum?
- What is the photon energy in electronvolts corresponding to the wavelength at which B_λ reaches a maximum?
- What is the mean photon energy of this spectrum in electronvolts?
- In what region of the electromagnetic spectrum does the peak of the black-body curve lie?

(Assume $h = 6.63 \times 10^{-34}$ J s, $c = 3.00 \times 10^8$ m s⁻¹, $k = 1.38 \times 10^{-23}$ J K⁻¹, 1 eV = 1.60×10^{-19} J.)

Black-body spectra are referred to as **thermal spectra** because they are produced by matter which has a characteristic temperature. Other types of spectra that are encountered in astrophysics include those due to bremsstrahlung and synchrotron radiation.

Bremsstrahlung (the German for ‘braking radiation’) arises when charged particles (such as electrons) are decelerated when passing close to other charged particles (such as atomic nuclei). Bremsstrahlung has a continuous spectrum, and when produced in a plasma whose electrons have an energy distribution described by the Maxwell–Boltzmann distribution, is described as **thermal bremsstrahlung**. In astrophysics and cosmology, thermal bremsstrahlung is observed in the X-ray region of the electromagnetic spectrum from plasma at temperatures of $10^7 - 10^8$ K.

Synchrotron radiation is generated by highly relativistic electrons that are accelerated by magnetic fields. As such it gives rise to spectra which are described

as **non-thermal**. Like black-body radiation and bremsstrahlung, synchrotron radiation also has a continuous spectrum, and in an astrophysical context may be observed across the whole electromagnetic spectrum from objects ranging from pulsars to active galaxies. A synchrotron spectrum typically has a power-law form on a spectral energy distribution, and the radiation is also polarized.

5.10.4 Opacity and optical depth

Whenever electromagnetic radiation passes through a medium (which in astrophysics is usually in the form of a gas) some of the radiation will generally be absorbed. The quantity describing the amount of absorption by the medium is known as the **opacity**. Unfortunately though, there are *two* definitions of opacity in use in astrophysics.

1. The first definition of opacity is the ratio of the total radiant energy received by a body to the amount transmitted through it. Thus, under this definition, a totally transparent body has an opacity of one (all radiation transmitted) whilst an opaque body has an opacity of infinity (all radiation absorbed).
2. A second definition of opacity is the absorption probability per unit time divided by the flux of photons per unit time. Under this definition therefore, a totally transparent body has an opacity of zero (zero probability of absorption).

Related to opacity is the concept of **optical depth**. Denoted by the symbol τ , it is defined by the following:

$$I_x/I_0 = \exp(-\tau) \quad (5.100)$$

where I_0 and I_x are amount of radiant energy incident on a gaseous body and the amount remaining after travelling for a distance x within the gas. Optical depth, like opacity, depends on the chemical composition, density and temperature of the gas in question.

- What are the optical depths of a transparent body and an opaque body?
- For a transparent body, $I_x = I_0$, so $\exp(-\tau) = 1$ and therefore $\tau = 0$. For an opaque body, $I_x = 0$, so $\exp(-\tau) = 0$ and therefore $\tau = \infty$.

An optical depth of one implies that the radiation is reduced by a factor of $\exp(-1)$ or $1/e$ as a result of transmission through the gas. If the optical depth is much greater than one, the body is referred to as being **optically thick** and if the optical depth is much less than one, the body is referred to as being **optically thin**.

Despite its name, optical depth can refer to radiation from any region of the electromagnetic spectrum. For instance the optical depth to X-rays of energy 1 keV through the Crab nebula supernova remnant is $\tau \approx 0.2$. This implies that the intensity of 1 keV X-rays is reduced to about $\exp(-0.2) = 0.8$ or 80% as they pass through the nebula.

Summary of Chapter 5

1. The velocity of a particle is its rate of change of position with respect to time, $v = dr/dt$, whilst the acceleration of a particle is its rate of change of

velocity with respect to time, $\mathbf{a} = d\mathbf{v}/dt$.

- For uniform motion in a circle, the magnitude of the centripetal acceleration is $a = r\omega^2 = v^2/r$, where ω is the angular speed ($= 2\pi/P$) and v is the magnitude of the instantaneous velocity tangential to the circle.
- Newton's three laws of motion are the key to predicting how bodies will move. The second law may be written $\mathbf{F} = m\mathbf{a}$ if m is constant, or more generally $\mathbf{F} = d\mathbf{p}/dt$, where \mathbf{F} is a force acting on a body, m its mass, \mathbf{a} its acceleration and \mathbf{p} its linear momentum.
- Newton's law of universal gravitation describes the gravitational force between two objects of mass m_1 and m_2 whose centres are separated by a distance r :

$$\mathbf{F}_{21} = -\frac{Gm_1m_2}{r^2}\hat{\mathbf{r}}$$

- According to Einstein's theory of special relativity, the transformations between the coordinates measured in two frames of reference in standard configuration are:

$$\begin{aligned}x' &= \frac{x - Vt}{\sqrt{1 - \frac{V^2}{c^2}}} \\y' &= y \\z' &= z \\t' &= \frac{t - Vx/c^2}{\sqrt{1 - \frac{V^2}{c^2}}}\end{aligned}$$

where V is the speed of one frame of reference relative to the other.

- These transformations lead to effects described as time dilation and length contraction whereby 'moving clocks run slow' and 'moving rods contract in their direction of motion'.
- The translational kinetic energy of a body is given by $E_{\text{KE}} = \frac{1}{2}mv^2$ whilst the magnitude of its linear momentum is $p = mv$. Both the energy and linear momentum of an isolated system are conserved. A collision in which kinetic energy is conserved is said to be elastic.
- The equations for translational kinetic energy and momentum need to be modified as the speeds involved approach the speed of light. The total relativistic energy of a body is the sum of its relativistic translational kinetic energy and its mass energy.
- The work done on any body by a force is the energy transferred to or from the body. The energy of a system is a measure of its capacity for doing work and power is the rate at which work is done or energy is transferred.
- The gravitational potential energy of a body of mass m at a distance r from the centre of a body of mass M is $E_{\text{GR}} = -GmM/r$ and the escape speed from the surface of a body of mass M and radius r is $v_{\text{esc}} = (2GM/r)^{1/2}$.
- The rotational analogues of force, mass, linear momentum and translational kinetic energy are torque, moment of inertia, angular momentum and rotational kinetic energy respectively. Angular momentum is conserved for a system on which no external torques act.

12. Macroscopic properties of a gas include its pressure ($P = F/A$) and density ($\rho = M/V$). Another macroscopic property, the temperature of a gas, is usefully defined as a label that determines the direction of heat flow: heat flows from a body with a higher temperature to a body with a lower temperature, and keeps on flowing until both bodies are at the same temperature.
13. One mole of a substance is an amount of it containing its relative atomic mass (for an element) or its relative molecular mass (for a compound) in grams. The number of basic particles per mole is called Avogadro's constant, $N_m = 6.02 \times 10^{23} \text{ mol}^{-1}$. Thus $M_m = M_r \times 10^{-3} \text{ kg mol}^{-1}$ and $M_m = N_m m$ where M_m is the molar mass of the element or compound, M_r is the relative atomic or molecular mass and m is the actual mass of the atom or molecule.
14. An ideal gas obeys the relationship $PV = NkT$, or $PV = nRT$, or equivalently $P = \rho kT/m$, and has an average translational kinetic energy per molecule of $\langle E_{KE} \rangle = 3kT/2$.
15. The sound speed in a gas depends on the ratio of the pressure to the density of the gas or equivalently on the temperature of the gas.
16. Atoms consist of positively charged nuclei (containing protons and (usually) neutrons) surrounded by a cloud of negatively charged electrons. The atomic number Z quantifies the number of protons in the nucleus and determines the type of atom. The mass number A quantifies the total number of protons and neutrons in the nucleus and specifies the particular isotope of the atom. In a neutral atom, the number of electrons is equal to the number of protons.
17. When atoms make transitions between different energy levels, photons are absorbed or emitted such that $E_{ph} = \Delta E_{atom}$. Hydrogen has the simplest energy level diagram with the energy of the n th level determined by $E_n = -13.60 \text{ eV}/n^2$.
18. The Balmer series of lines in the hydrogen spectrum originates in transitions down to or up from the $n = 2$ energy level. Such transitions give rise to emission or absorption lines in the visible part of the spectrum.
19. The Boltzmann equation describes how the relative proportion of atoms in different energy levels vary with temperature. The Saha ionization equation describes how the relative numbers of ions and neutral atoms vary with temperature.
20. Quantum physics implies that any particle will have an associated wavelength, known as its de Broglie wavelength, given by $\lambda_{dB} = h/p$.
21. The information describing the behaviour of a particle is contained in its wave function which is the solution to Schrödinger's equation for that particle. When the particle has particular, allowed, values of energy, momentum, spin, etc., it is described as being in a particular quantum state and may be characterized by a set of quantum numbers. The Pauli exclusion principle asserts that for certain particles, such as electrons, no two particles in a system can occupy the *same* quantum state.
22. Quantum gases may be composed of either bosons or fermions, each of which are indistinguishable particles. Only one fermion may occupy a given

quantum state, but any number of bosons may do so. Bose gases and Fermi gases each have a characteristic occupation factor. The energy distribution law for each is the product of a density of states function and an occupation factor. Since photons are bosons, thermal (blackbody) radiation is an example of a Bose gas. Electrons in a conductor are an example of a Fermi gas and obey the Pauli distribution: all states up to the Fermi energy are typically occupied, whilst those above it are not.

23. In a degenerate gas, particles are forced into higher energy quantum states than they would normally occupy at a particular density and temperature. This gives rise to a degeneracy pressure. In this situation, the de Broglie wavelength of the particles is larger than their separation.
24. Nuclei can be transformed from one element to another as a result of α -decay, β -decay or electron capture. In these processes, and also in γ -decay, energy is released.
25. Nuclear energy can also be liberated as a result of nuclear fusion when lighter nuclei fuse to form heavier nuclei. This is the main source of energy in stars.
26. Coulomb's law describes the electrostatic force between two particles of charge q_1 and q_2 whose centres are separated by a distance r :

$$\mathbf{F}_{21} = \frac{q_1 q_2}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}}$$

27. The electrostatic force experienced by a charged particle in an electric field \mathcal{E} is given by $\mathbf{F}_{\text{el}} = q\mathcal{E}$, whereas the magnetic force on a charged particle moving with velocity \mathbf{v} in a magnetic field \mathbf{B} , is given by $\mathbf{F}_{\text{mag}} = q\mathbf{v} \times \mathbf{B}$. The direction of the electric force is the same as that of the electric field (or opposite if the charge is negative), whilst the direction of the magnetic force is at right angles to the plane containing the velocity vector of the particle and the magnetic field vector.
28. Electromagnetic radiation propagates as mutually perpendicular, self-propagating electric and magnetic fields. It is a transverse wave and may be polarized.
29. Although electromagnetic radiation interacts with matter as though it is composed of a stream of particles, called photons, it propagates as a wave. The wavelength, frequency and photon energy are related by $c = \lambda\nu$ and $E_{\text{ph}} = h\nu$.
30. Electromagnetic radiation with the longest wavelengths, lowest frequencies and lowest energy photons is referred to as radio waves; electromagnetic radiation with the shortest wavelengths, highest frequencies and highest energy photons is referred to as gamma-rays. Visible light refers to electromagnetic radiation with wavelengths between around 400 nm and 700 nm or with photon energies between around 2 eV and 3 eV.
31. Continuous spectra of electromagnetic radiation can be produced by hot, dense objects in which there are many energy levels with extremely small separations between them. If a continuous spectrum passes through a gas then photons of some particular energies are absorbed by the gas, and are then emitted in all directions. Depending on the direction in which this gas

is viewed, either an emission line spectrum or an absorption line spectrum may be observed.

32. A black-body spectrum has a characteristic continuous shape, described by the Planck function. Higher temperature black-body spectra peak at shorter wavelengths (higher frequencies or photon energies) as determined by the Wien displacement law. The mean photon energy of a black-body spectrum is $\langle E_{\text{ph}} \rangle = 2.7kT$.
33. The amount by which a gaseous body absorbs radiation is characterized in terms of its optical depth, defined by $I_x/I_0 = \exp(-\tau)$. If $\tau \ll 1$ the body is said to be optically thin and if $\tau \gg 1$ the body is optically thick.

Appendix

Table 6.1 Common SI unit conversions and derived units

Quantity	Unit	Conversion
speed	m s^{-1}	
acceleration	m s^{-2}	
angular speed	rad s^{-1}	
angular acceleration	rad s^{-2}	
linear momentum	kg m s^{-1}	
angular momentum	$\text{kg m}^2 \text{s}^{-1}$	
force	newton (N)	$1 \text{ N} = 1 \text{ kg m s}^{-2}$
energy	joule (J)	$1 \text{ J} = 1 \text{ N m} = 1 \text{ kg m}^2 \text{s}^{-2}$
power	watt (W)	$1 \text{ W} = 1 \text{ J s}^{-1} = 1 \text{ kg m}^2 \text{s}^{-3}$
pressure	pascal (Pa)	$1 \text{ Pa} = 1 \text{ N m}^{-2} = 1 \text{ kg m}^{-1} \text{s}^{-2}$
frequency	hertz (Hz)	$1 \text{ Hz} = 1 \text{ s}^{-1}$
charge	coulomb (C)	$1 \text{ C} = 1 \text{ A s}$
potential difference	volt (V)	$1 \text{ V} = 1 \text{ J C}^{-1} = 1 \text{ kg m}^2 \text{s}^{-3} \text{A}^{-1}$
electric field	N C^{-1}	$1 \text{ N C}^{-1} = 1 \text{ V m}^{-1} = 1 \text{ kg m s}^{-3} \text{A}^{-1}$
magnetic field	tesla (T)	$1 \text{ T} = 1 \text{ N s m}^{-1} \text{C}^{-1} = 1 \text{ kg s}^{-2} \text{A}^{-1}$

Table 6.2 Other unit conversions

wavelength

1 nanometre (nm) = $10 \text{ \AA} = 10^{-9} \text{ m}$
 1 ångstrom = $0.1 \text{ nm} = 10^{-10} \text{ m}$

angular measure

$1^\circ = 60 \text{ arcmin} = 3600 \text{ arcsec}$
 $1^\circ = 0.01745 \text{ radian}$
 1 radian = 57.30°

temperature

absolute zero: $0 \text{ K} = -273.15^\circ\text{C}$
 $0^\circ\text{C} = 273.15 \text{ K}$

spectral flux density

1 jansky (Jy) = $10^{-26} \text{ W m}^{-2} \text{ Hz}^{-1}$
 $1 \text{ W m}^{-2} \text{ Hz}^{-1} = 10^{26} \text{ Jy}$

cgs units

1 erg = 10^{-7} J
 1 dyne = 10^{-5} N
 1 gauss = 10^{-4} T
 1 emu = 10 C

mass-energy equivalence

$1 \text{ kg} = 8.99 \times 10^{16} \text{ J}/c^2$ (c in m s^{-1})
 $1 \text{ kg} = 5.61 \times 10^{35} \text{ eV}/c^2$ (c in m s^{-1})

distance

1 astronomical unit (AU) = $1.496 \times 10^{11} \text{ m}$
 1 light-year (ly) = $9.461 \times 10^{15} \text{ m} = 0.307 \text{ pc}$
 1 parsec (pc) = $3.086 \times 10^{16} \text{ m} = 3.26 \text{ ly}$

energy

1 eV = $1.602 \times 10^{-19} \text{ J}$
 $1 \text{ J} = 6.242 \times 10^{18} \text{ eV}$

cross-section area

1 barn = 10^{-28} m^2
 $1 \text{ m}^2 = 10^{28} \text{ barn}$

pressure

1 bar = 10^5 Pa
 $1 \text{ Pa} = 10^{-5} \text{ bar}$
 1 atm pressure = 1.01325 bar
 1 atm pressure = $1.01325 \times 10^5 \text{ Pa}$

Table 6.3 Constants

Name of constant	Symbol	SI value
Fundamental constants		
gravitational constant	G	$6.673 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$
Boltzmann constant	k	$1.381 \times 10^{-23} \text{ J K}^{-1}$
speed of light in vacuum	c	$2.998 \times 10^8 \text{ m s}^{-1}$
Planck constant	h	$6.626 \times 10^{-34} \text{ J s}$
	$\hbar = h/2\pi$	$1.055 \times 10^{-34} \text{ J s}$
fine structure constant	$\alpha = e^2/4\pi\epsilon_0\hbar c$	1/137.0
Avogadro's constant	N_m	$6.022 \times 10^{23} \text{ mol}^{-1}$
Stefan-Boltzmann constant	$\sigma = \frac{2\pi^5 k^4}{15h^3 c^2}$	$5.671 \times 10^{-8} \text{ J m}^{-2} \text{ K}^{-4} \text{ s}^{-1}$
Radiation constant	$a = \frac{4\sigma}{c} = \frac{8\pi^5 k^4}{15h^3 c^3}$	$7.566 \times 10^{-16} \text{ J m}^{-3} \text{ K}^{-4}$
Thomson cross-section	σ_T	$6.652 \times 10^{-29} \text{ m}^2$
permittivity of free space	ϵ_0	$8.854 \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$
permeability of free space	μ_0	$4\pi \times 10^{-7} \text{ T m A}^{-1}$
Particle constants		
charge of proton	e	$1.602 \times 10^{-19} \text{ C}$
charge of electron	$-e$	$-1.602 \times 10^{-19} \text{ C}$
electron rest mass	m_e	$9.109 \times 10^{-31} \text{ kg}$ $0.511 \text{ MeV}/c^2$
proton rest mass	m_p	$1.673 \times 10^{-27} \text{ kg}$ $938.3 \text{ MeV}/c^2$
neutron rest mass	m_n	$1.675 \times 10^{-27} \text{ kg}$ $939.6 \text{ MeV}/c^2$
atomic mass unit	u	$1.661 \times 10^{-27} \text{ kg}$
Astronomical constants		
mass of the Sun	M_\odot	$1.99 \times 10^{30} \text{ kg}$
radius of the Sun	R_\odot	$6.96 \times 10^8 \text{ m}$
luminosity of the sun	L_\odot	$3.83 \times 10^{26} \text{ J s}^{-1}$
mass of the Earth	M_\oplus	$5.97 \times 10^{24} \text{ kg}$
radius of the Earth	R_\oplus	$6.37 \times 10^6 \text{ m}$
mass of Jupiter	M_J	$1.90 \times 10^{27} \text{ kg}$
radius of Jupiter	R_J	$7.15 \times 10^7 \text{ m}$
astronomical unit	AU	$1.496 \times 10^{11} \text{ m}$
light-year	ly	$9.461 \times 10^{15} \text{ m}$
parsec	pc	$3.086 \times 10^{16} \text{ m}$
Hubble parameter	H_0	$67.74 \pm 0.46 \text{ km s}^{-1} \text{ Mpc}^{-1}$ $2.195 \pm 0.015 \times 10^{-18} \text{ s}^{-1}$
age of Universe	t_0	$13.80 \pm 0.02 \times 10^9 \text{ years}$
current critical density	$\rho_{\text{crit},0}$	$8.62 \pm 0.12 \times 10^{-27} \text{ kg m}^{-3}$
current dark energy density	$\Omega_{\Lambda,0}$	$69.1 \pm 0.6\%$
current matter density	$\Omega_{m,0}$	$30.9 \pm 0.6\%$
current baryonic matter density	$\Omega_{b,0}$	$4.9 \pm 0.1\%$
current non-baryonic matter density	$\Omega_{c,0}$	$25.9 \pm 0.6\%$

Table 6.4 Greek letters and the quantities they represent

α	fine structure constant <i>or</i> viscosity parameter
β	speed as a fraction of speed of light <i>or</i> angular displacement <i>or</i> $1/kT$
γ	Lorentz factor <i>or</i> adiabatic index / ratio of heat capacities <i>or</i> shear
δ	small change <i>or</i> Kronecker delta
ε	permittivity of free space <i>or</i> energy generation rate <i>or</i> complex ellipticity
ζ	mass-radius index <i>or</i> equation of state parameter
η	efficiency <i>or</i> dynamical viscosity <i>or</i> Minkowski metric <i>or</i> entropy per baryon
θ	arbitrary angle <i>or</i> angular coordinate <i>or</i> degeneracy parameter <i>or</i> Einstein temperature
κ	opacity <i>or</i> convergence <i>or</i> performance
λ	wavelength <i>or</i> length scale <i>or</i> affine parameter <i>or</i> decay constant
μ	mean molecular weight <i>or</i> proper motion <i>or</i> relative mass <i>or</i> reduced mass <i>or</i> chemical potential <i>or</i> magnification
ν	frequency <i>or</i> kinematic viscosity <i>or</i> temperature index of reaction rate
ξ	duty cycle <i>or</i> correlation function <i>or</i> impact parameter <i>or</i> projected distance between star and planet in stellar radii
π	3.141... <i>or</i> parallax angle
ρ	mass density <i>or</i> electric charge density
σ	Stefan-Boltzmann constant <i>or</i> cross section <i>or</i> shear stress <i>or</i> standard deviation <i>or</i> velocity dispersion
τ	timescale <i>or</i> half-life <i>or</i> optical depth
ϕ	phase <i>or</i> azimuthal coordinate <i>or</i> work function <i>or</i> inflaton field <i>or</i> luminosity function
χ	statistical test <i>or</i> energy term in wave function
ψ	time independent wave function <i>or</i> scaled projected Newtonian potential
ω	angular speed <i>or</i> angular frequency <i>or</i> longitude of pericentre
Γ	torque <i>or</i> connection coefficient
Δ	change in a quantity <i>or</i> Gamow width
Λ	cosmological constant <i>or</i> Lorentz transformation matrix
Π	product
Σ	summation <i>or</i> surface density
Φ	phase <i>or</i> gravitational potential
Ψ	wave function
Ω	angular speed <i>or</i> density parameter <i>or</i> longitude of ascending node

Table 6.5 Lower case Roman letters and the quantities they represent

<i>a</i>	acceleration <i>or</i> semi-major axis <i>or</i> dimensionless scale factor <i>or</i> radiation constant
<i>b</i>	width <i>or</i> semi-minor axis <i>or</i> impact parameter <i>or</i> bias parameter <i>or</i> Galactic latitude
<i>c</i>	speed of light
<i>d</i>	distance
<i>e</i>	proton charge <i>or</i> eccentricity <i>or</i> 2.718...
<i>f</i>	general function <i>or</i> mass function <i>or</i> focal length <i>or</i> degrees of freedom <i>or</i> Maxwell speed distribution
<i>g</i>	acceleration due to gravity <i>or</i> surface gravity <i>or</i> number of polarizations <i>or</i> metric tensor <i>or</i> Maxwell-Boltzmann energy distribution
<i>h</i>	Planck constant <i>or</i> relative Hubble constant
<i>i</i>	integer index <i>or</i> inclination angle <i>or</i> imaginary number
<i>j</i>	angular momentum
<i>k</i>	Boltzmann constant <i>or</i> curvature <i>or</i> wave number <i>or</i> Love number
<i>l</i>	length <i>or</i> path length <i>or</i> semi-latus rectum <i>or</i> Galactic longitude <i>or</i> angular momentum <i>or</i> orbital angular momentum quantum number
<i>m</i>	mass <i>or</i> apparent magnitude <i>or</i> magnification <i>or</i> magnetic dipole moment <i>or</i> magnetic quantum number
<i>n</i>	number density <i>or</i> number of moles <i>or</i> quantum concentration <i>or</i> principal quantum number
<i>p</i>	momentum <i>or</i> probability <i>or</i> pressure <i>or</i> planet radius/stellar radius <i>or</i> geometric albedo
<i>q</i>	charge <i>or</i> mass ratio <i>or</i> deceleration parameter
<i>r</i>	radius <i>or</i> radial coordinate
<i>s</i>	displacement <i>or</i> standard error <i>or</i> spacetime separation <i>or</i> spin quantum number
<i>t</i>	time
<i>u</i>	initial speed <i>or</i> atomic mass unit
<i>v</i>	speed
<i>w</i>	equation of state parameter
<i>x</i>	position <i>or</i> coordinate
<i>y</i>	position <i>or</i> coordinate
<i>z</i>	position <i>or</i> coordinate <i>or</i> redshift

Table 6.6 Upper case Roman letters and the quantities they represent

<i>A</i>	area <i>or</i> mass number <i>or</i> amplitude <i>or</i> absorption <i>or</i> Bond albedo <i>or</i> action <i>or</i> adiabatic accessibility index
<i>B</i>	magnetic field strength <i>or</i> blackbody radiation formula <i>or</i> binding energy
<i>C</i>	integration constant <i>or</i> circumference <i>or</i> heat capacity
<i>D</i>	diameter <i>or</i> dissipation rate <i>or</i> density of states function
<i>E, \mathcal{E}</i>	energy <i>or</i> electric field strength <i>or</i> event
<i>F</i>	force <i>or</i> flux <i>or</i> electromagnetic field tensor <i>or</i> occupation factor <i>or</i> Maxwell speed distribution
<i>G</i>	gravitational constant <i>or</i> viscous torque <i>or</i> Einstein tensor <i>or</i> Maxwell-Boltzmann energy distribution
<i>H</i>	Hubble constant <i>or</i> scale height
<i>I</i>	moment of inertia <i>or</i> intensity <i>or</i> image scale <i>or</i> ionization energy
<i>J</i>	angular momentum <i>or</i> electric current density
<i>K</i>	Gaussian curvature <i>or</i> constant in degenerate gas equation of state
<i>L, \mathcal{L}</i>	luminosity <i>or</i> length interval <i>or</i> angular momentum <i>or</i> Lagrangian
<i>M</i>	mass <i>or</i> absolute magnitude <i>or</i> angular magnification
<i>N, \mathcal{N}</i>	number <i>or</i> column density <i>or</i> Avogadro number <i>or</i> neutron number
<i>P</i>	pressure <i>or</i> period <i>or</i> probability <i>or</i> power <i>or</i> four momentum
<i>Q</i>	charge <i>or</i> energy released <i>or</i> heat <i>or</i> tidal dissipation parameter
<i>R</i>	radius <i>or</i> scale factor <i>or</i> fusion rate <i>or</i> Riemann curvature tensor <i>or</i> Ricci tensor <i>or</i> molar gas constant
<i>S</i>	flux <i>or</i> entropy <i>or</i> nuclear S-factor <i>or</i> surface <i>or</i> spin angular momentum
<i>T</i>	temperature <i>or</i> time period <i>or</i> generic tensor <i>or</i> energy momentum tensor
<i>U</i>	four velocity <i>or</i> internal energy
<i>V</i>	volume <i>or</i> potential energy
<i>W</i>	work <i>or</i> equivalent width <i>or</i> number of microstates
<i>X</i>	mass fraction
<i>Y</i>	number of electrons per nucleon
<i>Z</i>	atomic number <i>or</i> partition function

Solutions to exercises

Ex 1.1 (a)

$$t \left(2 - \frac{k}{t^2} \right) = 2t - \frac{kt}{t^2} = 2t - \frac{k}{t}$$

(b)

$$\begin{aligned} (a - 2b^2) &= (a - 2b)(a + 2b) = a(a + 2b) - 2b(a + 2b) \\ &= a^2 + 2ab - 2ba - 4b^2 = a^2 - 4b^2 \end{aligned}$$

Ex 1.2 (a)

$$\frac{2xy}{z} \div \frac{z}{2} = \frac{2xy}{z} \times \frac{2}{z} = \frac{4xy}{z^2}$$

(b)

$$\frac{a^2 - b^2}{a + b} = \frac{(a + b)(a - b)}{a + b} = a - b$$

(c)

$$\begin{aligned} \frac{2}{3} + \frac{5}{6} &= \frac{2 \times 6}{3 \times 6} + \frac{5 \times 3}{6 \times 3} \\ &= \frac{12}{18} + \frac{15}{18} = \frac{27}{18} = \frac{27/9}{18/9} = \frac{3}{2} \end{aligned}$$

(d)

$$\frac{a}{b} - \frac{c}{d} = \frac{ad}{bd} - \frac{cb}{db} = \frac{ad - cb}{bd}$$

Ex 1.3 (a) Since $E = -GmM/r$, therefore $Er = -GmM$ and so $m = -Er/GM$.

(b) Since $E^2 = p^2c^2 + m^2c^4$ therefore $E^2 - p^2c^2 = m^2c^4$ and so $m^2 = (E^2 - p^2c^2)/c^4$. Hence $m = \pm\sqrt{(E^2 - p^2c^2)/c^4}$. Since m is declared to be a mass, it must be positive so the negative square root is rejected on physical grounds. Therefore $m = \sqrt{(E^2 - p^2c^2)/c^4}$.

(c) Since $T = 2\pi\sqrt{m/k}$ therefore $T^2 = (2\pi)^2(m/k)$ and so $m = k(T/2\pi)^2$.

Note: The solutions have been written out step by step. You may have arrived at the correct solution in fewer steps.

Ex 1.4 (a) Given (i) $a - b = 1$ and (ii) $a + b = 5$, from Equation (i), $a = 1 + b$, so substituting into Equation (ii), $(1 + b) + b = 5$, therefore $2b = 4$ and so $b = 2$. Therefore from (i), $a = 1 + b = 1 + 2 = 3$, and the solution is $a = 3$ and $b = 2$.

(b) Given (i) $2a - 3b = 7$ and (ii) $a + 4b = 9$, use Equation (ii) to express a in terms of b , so $a = 9 - 4b$ then substitute this expression for a into Equation (i) to evaluate b . This gives $2(9 - 4b) - 3b = 7$, so $18 - 8b - 3b = 7$, which leads to $18 - 7 = 11b$, and therefore $11b = 11$, i.e. $b = 1$.

Now substitute the known value of b into *either* of the original equations to obtain the value for a . Equation (ii) will be fastest, giving $a = 9 - 4b = 9 - 4 = 5$. The solution is therefore $a = 5, b = 1$.

Ex 1.5 (a) $10^2 \times 10^3 = 10^{2+3} = 10^5$

(b) $10^2/10^3 = 10^{2-3} = 10^{-1} = 1/10 = 0.1$

(c) $t^2/t^{-2} = t^{2-(-2)} = t^{2+2} = t^4$

(d) $1000^{1/3} = \sqrt[3]{1000} = 10$

(e) $(10^4)^{1/2} = 10^{4 \times 0.5} = 10^2 = 100$

(f) $125^{-1/3} = 1/\sqrt[3]{125} = 1/5 = 0.2$

(g) $(x^4/4)^{1/2} = x^{4 \times 0.5}/4^{0.5} = x^2/2$

(h) $(2 \text{ kg})^2/(2 \text{ kg})^{-2} = 2^2 \text{ kg}^2/2^{-2} \text{ kg}^{-2} = 2^2 \times 2^2 \times \text{kg}^2 \times \text{kg}^2 = 4 \times 4 \text{ kg}^4 = 16 \text{ kg}^4$

Note: These solutions have been written out using many steps as an aid to your working, but in many cases you may have been able to write down the answer immediately. In laying out calculations, you may include as many (or as few) steps as you feel comfortable with.

Ex 1.6 (a) Using Equation 1.11, $x = (-10 \pm \sqrt{10^2 - (4 \times 4 \times -6)})/(2 \times 4) = (-10 \pm \sqrt{196})/8 = (-10 \pm 14)/8$. So $x = -3$ or $x = 1/2$. The quadratic equation may therefore be written as $4(x+3)(x-0.5) = 0$ or equivalently (multiplying through by 4) as $(x+3)(4x-2) = 0$.

(b) Using Equation 1.11, $x = (0.9 \pm \sqrt{(-0.9)^2 - (4 \times 1 \times -17.86)})/(2 \times 1) = (0.9 \pm \sqrt{72.25})/2 = (0.9 \pm 8.5)/2$. So $x = 4.7$ or $x = -3.8$. The quadratic equation may therefore be written as $(x-4.7)(x+3.8) = 0$.

(c) This quadratic equation may be solved simply writing $8x^2 = 50$, so $x = \pm\sqrt{50/8}$. Hence the solutions are $x = \pm 2.5$. The quadratic equation may therefore be written as $8(x+2.5)(x-2.5) = 0$, or equivalently $(2x+5)(4x-10) = 0$.

Ex 1.7 1 light-year = 9.46×10^{15} m, so 4.2 light-years = $4.2 \times 9.46 \times 10^{15}$ m = 39.732×10^{15} m.

Converting to kilometres, this is 39.732×10^{12} km. Writing this in scientific notation, it becomes 3.9732×10^{13} km, and rounding to 2 significant figures, the final answer is 4.0×10^{13} km.

Note: This result is given to 2 s.f., the same precision as for the distance in light-years. For further discussion of this point, see Section 1.7.

Ex 1.8 (a) $6093 \text{ km}/500 \text{ s} = 12.2 \text{ km s}^{-1} = 1.22 \times 10^1 \text{ km s}^{-1}$.

(b) $2.0 \times 10^{18} \text{ km}/8.86 \times 10^{15} \text{ s} = 230 \text{ km s}^{-1} = 2.3 \times 10^2 \text{ km s}^{-1}$.

(c) $3000 \text{ km}/0.01 \text{ s} = 300\,000 \text{ km s}^{-1} = 3 \times 10^5 \text{ km s}^{-1}$.

In part (c), convention suggests that the distance is known to 4 s.f., i.e. it lies between 3000.5 km and 2999.5 km. The time, however, appears to be quoted to just 1 s.f., so the speed should be given to the same precision: $3 \times 10^5 \text{ km s}^{-1}$ is the only unambiguous way to write that result.

Ex 1.9 (a) The mean value $\langle m_v \rangle$ is found by adding all of the magnitude values and dividing by the number of values (ten). Thus $\langle m_v \rangle = (220.0)/10 = 22.0$.

(b) The measurements are spread over a range from 21.6 to 22.3, a range of about ± 0.35 magnitudes. It is conventional to quote the uncertainty in a measured value as about $2/3$ of the spread, in recognition of the fact that values from the extremes of the range are not very likely. So in this case we would estimate the uncertainty as $2/3 \times (\pm 0.35 \text{ magnitudes})$, which is ± 0.2 magnitudes, to one significant figure.

(c) The standard deviation is calculated using Equation 1.12. The mean value was calculated in part (a). The sum of the squared deviations of the measurements from the mean is 0.40 magnitudes². The mean of the squared deviations is therefore $(0.40 \text{ magnitudes}^2)/10 = 0.04 \text{ magnitudes}^2$, and the standard deviation (calculated by taking the square root of this) is 0.2 magnitudes. Note that the size of the uncertainty estimated in (b) is approximately the same as the standard deviation, and this is why it is often sufficient to use the simpler $2/3$ spread procedure.

(d) The uncertainty in the mean magnitude is $\sigma_m = s_n/\sqrt{n} = 0.2/\sqrt{10} \sim 0.06$ magnitudes.

Ex 1.10 (a) $\log_{10} 1000 = 3$, since $1000 = 10^3$

(b) $\log_{10} 0.001 = -3$, since $0.001 = 10^{-3}$

(c) $\log_{10} \sqrt{10} = 0.5$, since $\sqrt{10} = 10^{0.5}$

Ex 1.11 (a) (i) $\log_{10} 200 = \log_{10}(2 \times 10^2) = \log_{10} 2 + \log_{10} 10^2 = 0.301 + 2 = 2.301$

(ii) $\log_{10} 32 = \log_{10}(2^5) = 5 \times \log_{10} 2 = 5 \times 0.301 = 1.505$

(iii) $\log_{10} 0.25 = \log_{10}(2^{-2}) = -2 \times \log_{10} 2 = -2 \times 0.301 = -0.602$

(b) (i) $\log_{10} 3 + \log_{10} 8 = \log_{10}(3 \times 8) = \log_{10} 24$

(ii) $\log_{10} 4 - \log_{10} 3 - \log_{10} 5 = \log_{10}(4/(3 \times 5)) = \log_{10}(4/15)$

(iii) $3 \log_{10} 2 = \log_{10}(2^3) = \log_{10} 8$

Ex 1.12 You could plot U against x , this would give a curve (more specifically a parabola, Figure S1.1a) since U depends linearly on x^2 (not x). A smooth curve would certainly suggest a simple relationship between U and x , but just by looking at a curve, it is difficult to be sure of its exact shape. To test equations, therefore, it is always best to plot straight-line graphs. In this case, a graph of U against x^2 should give a straight line, with gradient $k/2$ and intercept c on the U -axis (Figure S1.1b). According to the question, x is the independent variable; U is therefore plotted along the vertical axis.

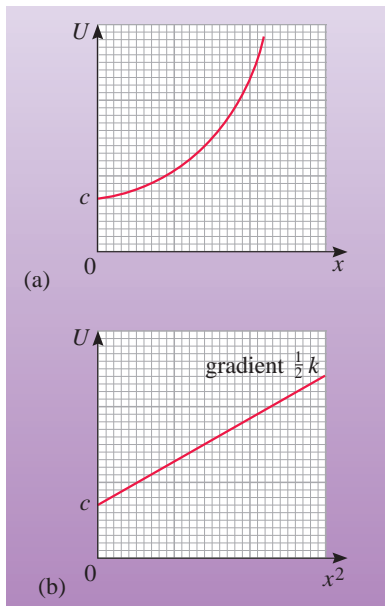


Figure S1.1 Two ways of plotting $U = kx^2/2 + c$.

Ex 1.13 (a)

$$90^\circ = \frac{\pi}{2} \text{ radians}$$

$$30^\circ = \frac{\pi}{6} \text{ radians}$$

$$180^\circ = \pi \text{ radians}$$

(b)

$$\frac{\pi}{8} \text{ radians} = \frac{180^\circ}{8} = 22.5^\circ$$

$$\frac{3\pi}{2} \text{ radians} = \frac{3 \times 180^\circ}{2} = 270^\circ$$

Ex 1.14 At a distance of 1.0×10^{17} m away from the star, a detector of unit area (i.e. 1 m^2) subtends a solid angle of $1.0 \text{ m}^2 / (1.0 \times 10^{17} \text{ m})^2$ steradians = 1.0×10^{-34} sr. The power per unit area received by the detector in this particular frequency range is therefore

$$(1.4 \times 10^6 \text{ W sr}^{-1}) \times (1.0 \times 10^{-34} \text{ sr}) / 1 \text{ m}^2 = 1.4 \times 10^{-28} \text{ W m}^{-2}$$

Ex 1.15 We can construct a right-angled triangle as shown in Figure S1.2.

Clearly the Sun's radius subtends an angle of $(31.9'/2) = 15.95'$, and $15.95'$ is equivalent to $(15.95/60.0) = 0.2658^\circ$. So we can write $\tan 0.2658^\circ = R / (1.50 \times 10^{11} \text{ m})$, from which $R = 6.96 \times 10^8 \text{ m}$ and the diameter of the Sun is therefore $2 \times 6.96 \times 10^8 \text{ m} = 1.39 \times 10^9 \text{ m}$.

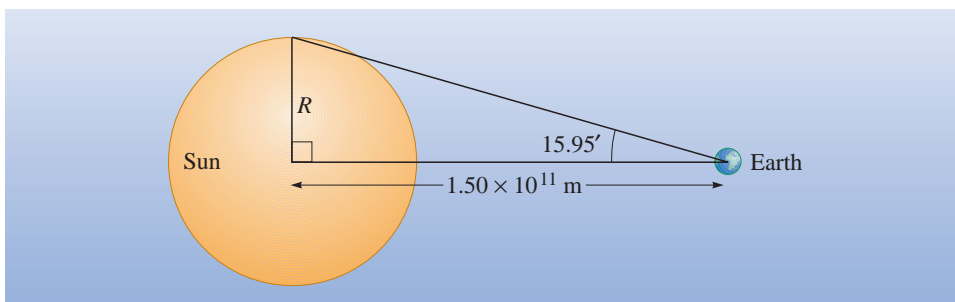


Figure SI.2 The disc of the Sun subtends an angle of $31.9'$ at the Earth's orbit.

Ex 1.16 (a) The amplitude of the velocity is $A = 20.0 \text{ km s}^{-1}$

(b) The period of the motion is $P = 8.0 \text{ days}$

(c) The frequency of the motion is $f = 1/P = (1/8.0) \text{ days}^{-1} = 0.125 \text{ days}^{-1}$ or $1/(8.0 \times 24 \times 3600) \text{ s}^{-1} = 1.4 \times 10^{-6} \text{ Hz}$.

(d) The angular frequency of the motion is $\omega = 2\pi/P = 2\pi/(8.0 \times 24 \times 3600) \text{ rad s}^{-1} = 9.1 \times 10^{-6} \text{ rad s}^{-1}$.

Ex 1.17 (a) $\sin^{-1}(2/\sqrt{5}) = 63.4^\circ$.

(b) The hypotenuse has a relative length of $\sqrt{5}$ units and one of the other sides of the triangle has a relative length of 2 units. By Pythagoras's theorem, the square of the relative length of the third side is given by $(\sqrt{5})^2 - 2^2 = 5 - 4 = 1$. So the relative length of the third side is $\sqrt{1} = 1$ unit. The three sides therefore have lengths in the ratio $1:2:\sqrt{5}$.

(c) The tangent of the smallest angle in the triangle (i.e. $90^\circ - 63.4^\circ = 26.6^\circ$) is given by $1/2 = 0.5$.

Ex 1.18 The x -component of the force vector is given by

$$\begin{aligned} F_x &= F \cos \theta_x \\ &= (3.50 \times 10^{22} \text{ N}) \times \cos 30^\circ \\ &= 3.03 \times 10^{22} \text{ N} \end{aligned}$$

and the y -component of the force vector is given by

$$\begin{aligned} F_y &= F \cos \theta_y \\ &= (3.50 \times 10^{22} \text{ N}) \times \cos 60^\circ \\ &= 1.75 \times 10^{22} \text{ N} \end{aligned}$$

(Notice that $\theta_x + \theta_y = 90^\circ$ since the axes must be at right angles to each other.)

Ex 1.19 The magnitude of the position vector is given by

$$\begin{aligned} a &= (a_x^2 + a_y^2)^{1/2} \\ &= (0.90^2 + 1.20^2)^{1/2} \times 10^{11} \text{ m} \\ &= 1.50 \times 10^{11} \text{ m} \end{aligned}$$

It represents the distance of the Sun from the Earth.

Ex 1.20 (a) $\mathbf{a} \cdot \mathbf{a} = a^2$, a positive scalar, since $\cos 0 = 1$.

(b) $\mathbf{b} \times \mathbf{b} = \mathbf{0}$, i.e. it vanishes since $\sin 0 = 0$. The vector (cross) product of any parallel vectors vanishes.

Ex 1.21 The range (r) coordinate is given by

$$\begin{aligned} r &= (x^2 + y^2 + z^2)^{1/2} \\ &= (1.2^2 + 1.6^2 + 0.0^2)^{1/2} \times 10^{10} \text{ m} = 2.0 \times 10^{10} \text{ m} \end{aligned}$$

The elevation (θ) coordinate is found from $\cos \theta = z/r$, but since $z = 0.0$, clearly $\cos \theta = 0.0$ and so $\theta = 90^\circ$.

Finally the azimuthal (ϕ) coordinate is found from

$$\begin{aligned} \sin \phi &= y/(r \sin \theta) \\ &= (1.6 \times 10^{10} \text{ m}) / (2.0 \times 10^{10} \text{ m} \times \sin 90^\circ) = 0.80 \end{aligned}$$

so $\phi = 53^\circ$. The spherical coordinates of the star are therefore $(2.0 \times 10^{10} \text{ m}, 90^\circ, 53^\circ)$.

Ex 1.22 Temperature and pressure within a star are scalar fields – they can be represented simply by a number at each point inside the star. Gravity and magnetic field are vector fields – they have a magnitude *and* a direction at each point inside the star.

Ex 1.23 The matrix PQ is a 3×3 matrix whose elements are given by

$$PQ = \begin{pmatrix} (1 \times 1) + (0 \times -1) + (1 \times 0) & (1 \times -1) + (0 \times 0) + (1 \times 1) & (1 \times 1) + (0 \times -1) + (1 \times 0) \\ (0 \times 1) + (-1 \times -1) + (0 \times 0) & (0 \times -1) + (-1 \times 0) + (0 \times 1) & (0 \times 1) + (-1 \times -1) + (0 \times 0) \\ (1 \times 1) + (0 \times -1) + (1 \times 0) & (1 \times -1) + (0 \times 0) + (1 \times 1) & (1 \times 1) + (0 \times -1) + (1 \times 0) \end{pmatrix}$$

So the resulting matrix is

$$PQ = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Similarly, the matrix QP is a 3×3 matrix whose elements are given by

$$QP = \begin{pmatrix} (1 \times 1) + (-1 \times 0) + (1 \times 1) & (1 \times 0) + (-1 \times -1) + (1 \times 0) & (1 \times 1) + (-1 \times 0) + (1 \times 1) \\ (-1 \times 1) + (0 \times 0) + (-1 \times 1) & (-1 \times 0) + (0 \times -1) + (-1 \times 0) & (-1 \times 1) + (0 \times 0) + (-1 \times 1) \\ (0 \times 1) + (1 \times 0) + (0 \times 1) & (0 \times 0) + (1 \times -1) + (0 \times 0) & (0 \times 1) + (1 \times 0) + (0 \times 1) \end{pmatrix}$$

so the resulting matrix is

$$QP = \begin{pmatrix} 2 & 1 & 2 \\ -2 & 0 & -2 \\ 0 & -1 & 0 \end{pmatrix}$$

Clearly $PQ \neq QP$.

Ex 1.24 (a) $|A| = (2 \times 5) - (4 \times 3) = -2$.

(b) $|B| = 1(3 \times 3 - 4 \times 2) - 2(2 \times 3 - 4 \times 1) + 3(2 \times 2 - 3 \times 1) = 1 - 4 + 3 = 0$.
The matrix is singular.

Ex 2.1 Moving at 33 km s^{-1} , Tau Ceti would cover a path length of $(33 \text{ km s}^{-1} \times 3600 \text{ seconds per hour} \times 24 \text{ hours per day} \times 365 \text{ days per year}) =$

1.04×10^9 km in one year. If the angle subtended by a length of 1.04×10^9 km at a distance of 1.12×10^{14} km away is θ radians, then

$$\tan \theta = (1.04 \times 10^9 \text{ km}) / (1.12 \times 10^{14} \text{ km}) = 9.29 \times 10^{-6}$$

This is clearly a small angle, so $\tan \theta = \theta$ and the angle is 9.29×10^{-6} radians or 5.32×10^{-4} degrees or $1.92''$. Therefore Tau Ceti moves less than $2''$ per year.

Ex 2.2 The magnitude of the radial velocity is

$$\begin{aligned} v_r &= (3.00 \times 10^8 \text{ m s}^{-1}) \times \\ & (4863.5 \text{ \AA} - 4861.3 \text{ \AA}) / 4861.3 \text{ \AA} \\ &= 1.36 \times 10^5 \text{ m s}^{-1} \end{aligned}$$

Since the difference in wavelengths is accurate to only 2 s.f., the magnitude of the radial velocity can be stated as 140 km s^{-1} . As the shifted wavelength is larger than the rest wavelength, the star is moving *away* from the Earth. (Notice that both a magnitude and a direction are required to fully characterize the radial velocity of the star.)

Ex 2.3 (a) Figure 2.8 shows that the strengths of the hydrogen and neutral helium lines are equal at a photospheric temperature of about 20 000 K.

(b) Table 2.1 shows that a B5 star has a temperature of 17 000 K whilst an O5 star has a temperature of 40 000 K. This indicates that a star with a temperature of 20 000 K would have a spectral classification of an early B-type, say B3.

Ex 2.4 The effective photospheric temperature of the Sun is

$$\begin{aligned} T_{\text{eff}} &= \sqrt[4]{\frac{L}{4\pi\sigma R^2}} \\ &= \sqrt[4]{\frac{3.83 \times 10^{26} \text{ W}}{4\pi(5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4})(6.96 \times 10^8 \text{ m})^2}} = 5772 \text{ K} \end{aligned}$$

The effective photospheric temperature of the Sun is therefore 5770 K (to 3 s.f.).

Ex 2.5 We have already calculated the absolute visual magnitude of Rigel as $M_{\text{Rigel}} = -7.12$. Using the same approach, the absolute visual magnitude of Ross 154 is $M_{\text{Ross154}} = 10.45 + 5 - (5 \log_{10} 2.9) + 0 = 13.14$.

So using Equation 2.12, the ratio of luminosities can be calculated as $10^{(M_{\text{Rigel}} - M_{\text{Ross154}})/2.5} = L_{\text{Ross154}}/L_{\text{Rigel}}$ so,
 $L_{\text{Ross154}}/L_{\text{Rigel}} = 10^{(-7.12 - 13.14)/2.5} = 10^{-8.10} = 7.9 \times 10^{-9}$.

The ratio of the luminosity of Rigel to that of Ross 154 is therefore $L_{\text{Rigel}}/L_{\text{Ross154}} = 1/7.9 \times 10^{-9} = 1.3 \times 10^8$. Therefore Rigel is about 130 million times more luminous than Ross 154.

Ex 2.6 Using Equation 2.7,

$$\begin{aligned}
 L &\approx 4\pi R^2 \sigma T^4 \\
 &\approx 4\pi (3.5 \times 6.96 \times 10^8 \text{ m})^2 \\
 &\quad \times (5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}) \times (14\,500 \text{ K})^4 \\
 &\approx 1.87 \times 10^{29} \text{ W} \\
 &= [(1.87 \times 10^{29}) / (3.83 \times 10^{26})] L_{\odot} \\
 &= 488 L_{\odot}
 \end{aligned}$$

Given that the luminosity depends on the fourth power of the temperature, which is here given only roughly, the luminosity is close enough to that in Table 2.4.

Ex 2.7 (a) Light-gathering power is proportional to (aperture diameter)², so the ratio of light-gathering powers for the two telescopes is $(5.0/1.0)^2 = 25$.

(b) The (theoretical) limit of angular resolution is inversely proportional to the aperture of the objective lens or objective mirror. Thus, a telescope with $D_o = 5$ m can theoretically resolve two stars with an angular separation five times smaller than a telescope with $D_o = 1$ m (neglecting air turbulence and aberrations). In practice, of course, for ground-based telescopes, atmospheric seeing is usually the limiting factor.

Ex 3.1 The typical timescale for the X-ray flux to double is $\sim 10^4$ seconds. The radius of the region producing the radiation must be less than the distance light can travel in this time. Hence, using Equation 3.7,

$$R_{\text{max}} \sim 3.00 \times 10^8 \text{ s}^{-1} \times 10^4 \text{ s} = 3 \times 10^{12} \text{ m}$$

This corresponds to only about 20 AU.

Ex 3.2 (a) There are no such ranges, all the FRW models are consistent with the cosmological principle which demands homogeneity and isotropy.

(b) All ranges of k and Λ allow a big bang, but $k = +1$ models with $0 < \Lambda < \Lambda_E$ allow the possibility of universes that began without a big bang. The case $k = +1$, $\Lambda = \Lambda_E$ allows the possibility that the universe might be static (hence no big bang) or that there might not have been a big bang in a non-static universe. Among the limiting cases that arise as the density approaches zero, there are cases in which the big bang happened an infinitely long time ago.

(c) This is possible for $k = +1$ and $0 < \Lambda \leq \Lambda_E$

(d) There are no ranges that allow the big bang to be associated with a unique point in space. Such an association would violate the cosmological principle. Take good note of this since it is a widespread misconception to suppose that the big bang was the ‘explosion’ of a dense primeval ‘atom’ located at some particular point in space. Rather than thinking of the big bang as an event in spacetime you should think of it as giving rise to spacetime.

(e) This is true in any model with $k = 0$.

(f) This is true in any model with $k = +1$.

(g) This is true in all models with $k = 0$ or -1 . Of course, due to the finite speed of light we can have no direct observational knowledge of those parts of the Universe that are so distant that light emitted from them has not yet reached us.

Ex 3.3 The relationship between temperature and scale factor is given by Equation 3.18, $T \propto 1/R(t)$. Thus the relationship between the temperature of the background radiation at the present time T_0 and that at the time of last scattering T_{last} is

$$\frac{T_{\text{last}}}{T_0} = \frac{R(t_0)}{R(t_{\text{last}})}$$

where t_{last} is the time at which the last scattering of photons occurred. The relationship between redshift and scale factor is given by Equation 3.11. In this case, the time at which the photon is observed is t_0 and the time at which the photon was emitted is t_{last} , so Equation 3.11 can be written as

$$z = \frac{R(t_0)}{R(t_{\text{last}})} - 1$$

and so

$$z = \frac{T_{\text{last}}}{T_0} - 1$$

We know that $T_{\text{last}} = 3000 \text{ K}$ and $T_0 \sim 2.7 \text{ K}$, so

$$z = \frac{3000}{2.7} - 1 \sim 1100$$

So the redshift at which the last scattering of cosmic background photons occurred is 1.1×10^3 (to 2 significant figures).

Ex 3.4 Following the same reasoning as used in the answer to the previous exercise, the redshift z is related to the temperature T_{em} of the background radiation at that redshift by

$$z = \frac{T_{\text{em}}}{T_0} - 1$$

This can be rearranged to give

$$T_{\text{em}} = T_0(z + 1)$$

Inserting the given values leads to

$$T_{\text{em}} = (2.73 \text{ K}) \times (2.5 + 1) = 9.56 \text{ K}$$

So, their measurement of the temperature of the cosmic microwave background would be 9.6 K (to 2 significant figures). (Although we don't have any communication with astronomers anywhere else in the Universe, a similar principle applies to a real observational technique: it is possible to measure the temperature of the cosmic background radiation as experienced by Lyman α clouds at redshifts of $z \sim 2$. Such measurements, which are based on detailed analysis of spectral lines, show that the temperature of the cosmic background does increase with redshift in this way.)

Ex 3.5 Using Equation 3.16

$$\begin{aligned} \rho_{\text{crit}} &= \frac{3H_0^2}{8\pi G} \\ &= \frac{3 \times (67.74 \text{ km s}^{-1} \text{ Mpc}^{-1})^2}{8 \times \pi \times (6.67 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2})} \\ &= \frac{1.442 \times 10^{-35} \text{ s}^{-2}}{1.676 \times 10^{-9} \text{ m}^3 \text{ s}^{-2} \text{ kg}^{-1}} = 8.60 \times 10^{-27} \text{ kg m}^{-3} \end{aligned}$$

This is equivalent to about 5 hydrogen atoms per cubic metre.

Ex 4.1 (a) We can write $F(r) = k/r$ as $F(r) = kr^{-1}$. So, using the rule that

$$\frac{d(at^n)}{dt} = nat^{n-1}$$

we have

$$\frac{dF}{dr} = (-1) \times kr^{-1-1} = -kr^{-2} = -\frac{k}{r^2}$$

(b) This time we use the product rule,

$$\frac{d(uv)}{dt} = u \frac{dv}{dt} + v \frac{du}{dt}$$

where $u = ax^2$ and $v = \sin(bx)$. Now

$$\frac{d(ax^2)}{dx} = 2ax$$

$$\text{and } \frac{d[\sin(bx)]}{dx} = b \cos(bx)$$

$$\text{so } \frac{dG}{dx} = [ax^2 \times b \cos(bx)] + [\sin(bx) \times (2ax)]$$

Ex 4.2 Following the hint given at the end of the question, we equate the new value of y after one half-life with half the original value, i.e.

$$\exp[-(t + \tau_{1/2})/\tau] = 0.5 \exp(-t/\tau)$$

Taking natural logarithms, we obtain

$$(-t - \tau_{1/2})/\tau = \log_e 0.5 - (t/\tau)$$

and multiplying both sides by τ gives

$$(-t - \tau_{1/2}) = \tau \times \log_e 0.5 - t$$

then adding t to both sides gives

$$-\tau_{1/2} = \tau \times \log_e 0.5$$

and therefore

$$\tau_{1/2}/\tau = -\log_e 0.5$$

Finally, recall that $\log_e(1/x) = -\log_e x$, so

$$\tau_{1/2}/\tau = \log_e(1/0.5) = \log_e 2 = 0.69.$$

Now reading from the graph, the time for the radioactivity to halve is 5600 years, and the time for it to reach a factor $1/e$ ($= 0.37$) is 8100 years. The ratio $\tau_{1/2}/\tau = 5600/8100 = 0.69$, as expected.

Ex 4.3 (a) As indicated in the question, we first put $p = c\beta \sin \theta$ and $q = (1 - \beta \cos \theta)^{-1}$ so that $V = pq$.

Now, using the rules from Table 4.1, $dp/d\theta = c\beta \cos \theta$.

To calculate $dq/d\theta$ we first put $u = 1 - \beta \cos \theta$ so that $q = u^{-1}$. Now, by the chain rule,

$$\frac{dq}{d\theta} = \frac{dq}{du} \times \frac{du}{d\theta}$$

So $\frac{dq}{d\theta} = (-u^{-2}) \times (-\beta \times -\sin \theta)$

or $\frac{dq}{d\theta} = \frac{-\beta \sin \theta}{(1 - \beta \cos \theta)^2}$

Now we can finally apply the product rule:

$$\frac{d(pq)}{d\theta} = p \frac{dq}{d\theta} + q \frac{dp}{d\theta}$$

So

$$\frac{dV}{d\theta} = (c\beta \sin \theta) \times \frac{-\beta \sin \theta}{(1 - \beta \cos \theta)^2} + \frac{1}{(1 - \beta \cos \theta)} \times c\beta \cos \theta$$

$$\frac{dV}{d\theta} = \frac{-c(\beta \sin \theta)^2}{(1 - \beta \cos \theta)^2} + \frac{c\beta \cos \theta}{(1 - \beta \cos \theta)}$$

Finding a common denominator in order to add the two terms, this becomes

$$\frac{dV}{d\theta} = \frac{-c(\beta \sin \theta)^2 + c\beta \cos \theta (1 - \beta \cos \theta)}{(1 - \beta \cos \theta)^2}$$

$$= \frac{-(\beta \sin \theta)^2 + \beta \cos \theta - (\beta \cos \theta)^2}{(1 - \beta \cos \theta)^2} \times c$$

$$= \frac{-\beta^2(\sin^2 \theta + \cos^2 \theta) + \beta \cos \theta}{(1 - \beta \cos \theta)^2} \times c$$

We recall the rule that $\sin^2 \theta + \cos^2 \theta = 1$ for any value of θ (this comes from Pythagoras's theorem) and therefore

$$\frac{dV}{d\theta} = \frac{-\beta^2 + \beta \cos \theta}{(1 - \beta \cos \theta)^2} \times c$$

$$\frac{dV}{d\theta} = \frac{\beta(\cos \theta - \beta)}{(1 - \beta \cos \theta)^2} \times c$$

(b) The angle at which the maximum value of V is observed is found by setting $dV/d\theta = 0$. So

$$\frac{\beta(\cos \theta - \beta)}{(1 - \beta \cos \theta)^2} \times c = 0$$

and therefore $\cos \theta - \beta = 0$, so $\theta = \cos^{-1} \beta$.

(Note: The fact that this complicated question resulted in a simple answer at the end is a good indication that the calculation has been carried out correctly!)

Ex 4.4 (a) We write $f(\theta) = \sin \theta$ and therefore $f'(\theta) = \cos \theta$. The first-order Maclaurin expansion is therefore,

$$f(\theta) = f(0) + \theta f'(0) = \sin 0 + \theta \cos 0$$

Since $\sin 0 = 0$ and $\cos 0 = 1$, this becomes

$$f(\theta) = \theta$$

which is the small-angle approximation.

(b) For a slightly more accurate expansion we will need to calculate more terms in the expansion. So first we note that $f''(\theta) = -\sin \theta$ and $f'''(\theta) = -\cos \theta$. Therefore

$$f(\theta) = f(0) + \theta f'(0) + \theta^2 f''(0)/2 + \theta^3 f'''(0)/6$$

$$f(\theta) = \sin 0 + \theta \cos 0 - \theta^2 \sin 0/2 - \theta^3 \cos 0/6$$

Once again we note that $\sin 0 = 0$ and $\cos 0 = 1$, so

$$f(\theta) = \theta - \theta^3/6$$

This is a slightly more accurate approximation to $\sin \theta$ and will apply equally well for somewhat larger angles.

Ex 4.5 (a)

$$\begin{aligned} \frac{\partial y(x, t)}{\partial x} &= \frac{\partial}{\partial x} A \sin(kx + \omega t) \\ &= Ak \cos(kx + \omega t) \end{aligned}$$

(b)

$$\begin{aligned} \frac{\partial y(x, t)}{\partial t} &= \frac{\partial}{\partial t} A \sin(kx + \omega t) \\ &= A\omega \cos(kx + \omega t) \end{aligned}$$

(c)

$$\begin{aligned} \frac{\partial^2 y(x, t)}{\partial x^2} &= \frac{\partial}{\partial x} \left(\frac{\partial y(x, t)}{\partial x} \right) \\ &= \frac{\partial}{\partial x} [Ak \cos(kx + \omega t)] \\ &= -Ak^2 \sin(kx + \omega t) \\ &= -k^2 y(x, t) \\ \frac{\partial^2 y(x, t)}{\partial t^2} &= \frac{\partial}{\partial t} \left(\frac{\partial y(x, t)}{\partial t} \right) \\ &= \frac{\partial}{\partial t} [A\omega \cos(kx + \omega t)] \\ &= -A\omega^2 \sin(kx + \omega t) \\ &= -\omega^2 y(x, t) \end{aligned}$$

Ex 4.6 The gradient of the scalar field ∇h represents the way in which the altitude changes with distance in each direction – the *slope* of the land in a given direction at each point.

Ex 4.7 Given

$$x = A \sin(\omega t + \phi)$$

so

$$\frac{dx}{dt} = A\omega \cos(\omega t + \phi)$$

and

$$\frac{d^2x}{dt^2} = -A\omega^2 \sin(\omega t + \phi)$$

Substituting into this last equation using the original proposed solution, we have

$$\frac{d^2x}{dt^2} = -\omega^2 x$$

Comparing this result with the original differential equation, $m d^2x/dt^2 = -kx$, it is clear that the two equations have the same form provided that $\omega^2 = k/m$, as required.

Ex 4.8 (a) Using the rule that

$$\int at^n dt = \frac{at^{n+1}}{n+1} + C$$

in this case the power on the variable r is $n = -2$, so

$$\int \frac{GMm}{r^2} dr = -\frac{GMm}{r} + C$$

(b) Using the rule that

$$\int (u + v) dt = \int u dt + \int v dt$$

in this case

$$\int \left(b \exp x + \frac{1}{x} \right) dx = b \exp x + \log_e x + C$$

Ex 4.9 Using the suggested substitution, $x = a \sin \theta$, we have $dx/d\theta = a \cos \theta$.

So substituting into the original integral:

$$\begin{aligned} \int_{x=0}^{x=a} \frac{1}{\sqrt{a^2 - x^2}} dx &= \int_{x=0}^{x=a} \frac{1}{\sqrt{a^2 - a^2 \sin^2 \theta}} \times a \cos \theta d\theta \\ &= \int_{x=0}^{x=a} \frac{a \cos \theta}{a \sqrt{1 - \sin^2 \theta}} d\theta \end{aligned}$$

Now, since $\sin^2 \theta + \cos^2 \theta = 1$, so $1 - \sin^2 \theta = \cos^2 \theta$. Making this substitution inside the square root gives

$$\begin{aligned} \int_{x=0}^{x=a} \frac{1}{\sqrt{a^2 - x^2}} dx &= \int_{x=0}^{x=a} \frac{a \cos \theta}{a \sqrt{\cos^2 \theta}} d\theta \\ &= \int_{x=0}^{x=a} \frac{a \cos \theta}{a \cos \theta} d\theta \\ &= \int_{x=0}^{x=a} d\theta = [\theta]_{x=0}^{x=a} \end{aligned}$$

Reversing the original substitution we have $\theta = \sin^{-1}(x/a)$, so the integral becomes

$$\begin{aligned} \int_{x=0}^{x=a} \frac{1}{\sqrt{a^2 - x^2}} dx &= \left[\sin^{-1} \left(\frac{x}{a} \right) \right]_{x=0}^{x=a} \\ &= \sin^{-1} \left(\frac{a}{a} \right) - \sin^{-1} \left(\frac{0}{a} \right) \\ &= \sin^{-1}(1) - \sin^{-1}(0) \end{aligned}$$

Since $\sin^{-1}(1) = 90^\circ$ or $\pi/2$ radians and $\sin^{-1}(0) = 0$, the final answer is simply $\pi/2$.

Ex 4.10 Putting $u = \log_e x$ and $dv = dx$ we have $du/dx = 1/x$ and $v = x$. So we can write

$$\begin{aligned} \int u dv &= uv - \int v du \\ \int \log_e x dx &= x \log_e x - \int x \frac{dx}{x} \\ &= x \log_e x - \int dx \\ &= x \log_e x - x + C \end{aligned}$$

Ex 4.11 The mass of the star is given by

$$M_{\text{star}} = \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} \int_{\theta=0}^{\theta=\pi} \frac{R^2 \rho_0 r^2 \sin \theta}{r^2} dr d\phi d\theta$$

which simplifies immediately to

$$M_{\text{star}} = \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} \int_{\theta=0}^{\theta=\pi} R^2 \rho_0 \sin \theta dr d\phi d\theta$$

First integrating with respect to the angle θ :

$$\begin{aligned} M_{\text{star}} &= \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} \left[-R^2 \rho_0 \cos \theta \right]_{\theta=0}^{\theta=\pi} dr d\phi \\ &= \int_{r=0}^{r=R} \int_{\phi=0}^{\phi=2\pi} 2R^2 \rho_0 dr d\phi \end{aligned}$$

then integrating with respect to the angle ϕ :

$$\begin{aligned} M_{\text{star}} &= \int_{r=0}^{r=R} \left[2R^2 \rho_0 \phi \right]_{\phi=0}^{\phi=2\pi} dr \\ &= \int_{r=0}^{r=R} 4\pi R^2 \rho_0 dr \end{aligned}$$

and finally integrating with respect to r :

$$\begin{aligned} M_{\text{star}} &= \left[4\pi R^2 \rho_0 r \right]_{r=0}^{r=R} \\ &= 4\pi \rho_0 R^3 \end{aligned}$$

Ex 5.1 (a) Using Equation 5.3, the distance travelled is found as

$$s_x = u_x t + \frac{1}{2} a_x t^2 = 0 + 0.5 \times (5.0 \text{ m s}^{-2}) \times (10 \text{ s})^2 = 250 \text{ m}.$$

(b) Using Equation 5.4, the speed at this time is found as $v_x = u_x + a_x t = 0 + (5.0 \text{ m s}^{-2}) \times (10 \text{ s}) = 50 \text{ m s}^{-1}$.

Ex 5.2 (a) The angular speed of the neutron star is

$$\begin{aligned} \omega &= \frac{2\pi}{P} \\ &= \frac{2\pi}{(10 \times 24 \times 3600 \text{ s})} \\ &= 7.27 \times 10^{-6} \text{ s}^{-1} \end{aligned}$$

So, the magnitude of the instantaneous velocity is

$$\begin{aligned} v &= r\omega \\ &= (3.7 \times 10^{10} \text{ m}) \times (7.27 \times 10^{-6} \text{ s}^{-1}) \\ &= 2.7 \times 10^5 \text{ m s}^{-1} \text{ or } 270 \text{ km s}^{-1} \end{aligned}$$

(b) The magnitude of the centripetal acceleration is

$$\begin{aligned} a &= r\omega^2 \\ &= (3.7 \times 10^{10} \text{ m}) \times (7.27 \times 10^{-6} \text{ s}^{-1})^2 \\ &= 2.0 \text{ m s}^{-2} \end{aligned}$$

Ex 5.3 Imagine that the astronaut in the space station is at the origin of frame of reference A, and spaceship X is at the origin of frame of reference B. Then frames of reference A and B are in standard configuration with $V = 3c/4$. The x -component of the velocity of spaceship Y in frame of reference A is $v_x = -3c/4$. The question then becomes: what is the x' -component of the velocity of spaceship Y in frame of reference B? From Equation 5.29,

$$\begin{aligned} v'_x &= \frac{v_x - V}{1 - \frac{Vv_x}{c^2}} \\ &= \frac{-(3c/4) - (3c/4)}{1 - (-9c^2/16c^2)} \\ &= \frac{-6c/4}{1 + (9/16)} \\ &= -\frac{6c}{4} \times \frac{16}{25} \\ &= -\frac{96c}{100} \end{aligned}$$

The negative sign indicates that spaceship X measures spaceship Y to be racing towards it at 96% of the speed of light.

Ex 5.4 The work done on the particle is equal to its change in kinetic energy. In this case

$$\begin{aligned} W &= 0.5 \times 10^{-6} \text{ kg} \times (10^2 - 5^2) \text{ m}^2 \text{ s}^{-2} \\ &= 3.75 \times 10^{-5} \text{ J or about } 40 \mu\text{J} \end{aligned}$$

Ex 5.5 Equation 5.33 states that $E_{\text{GR}} = -\frac{GmM}{r}$ and Equation 5.35 can be written in this case as $F = -\frac{dE_{\text{GR}}}{dr}$. So differentiating Equation 5.33 with respect to r gives

$$\begin{aligned}\frac{dE_{\text{GR}}}{dr} &= -GMm \frac{d}{dr}(r^{-1}) \\ &= -GMm \times (-r^{-2}) \\ &= \frac{GMm}{r^2}\end{aligned}$$

So the magnitude of the force of gravity is $-GMm/r^2$, where the minus sign indicates that the force of gravity acts in the *opposite* direction to that in which r increases (i.e. towards the centre of the body of mass M). This is clearly just Newton's law of universal gravitation.

Ex 5.6 If the relativistic translational kinetic energy of a particle is equal to its mass energy, then Equation 5.40 becomes

$$\begin{aligned}E_{\text{KE}} &= \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} - mc^2 \\ mc^2 &= \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} - mc^2 \\ 2mc^2 &= \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} \\ \sqrt{1 - \frac{v^2}{c^2}} &= \frac{1}{2} \\ 1 - \frac{v^2}{c^2} &= \frac{1}{4} \\ \frac{v^2}{c^2} &= \frac{3}{4} \\ v &= \frac{\sqrt{3}c}{2}\end{aligned}$$

So the speed of the particle is $2.6 \times 10^8 \text{ m s}^{-1}$.

Ex 5.7 (a) The moment of inertia of an annulus will be greater than that of a uniform disc of the same mass, because most of the mass will be further away from the central axis.

(b) Since the magnitude of angular momentum is $L = I\omega$ and the rotational kinetic energy is $E_{\text{rot}} = \frac{1}{2}I\omega^2$, but ω is the same for both the disc and the annulus, then the annulus will have the greater angular momentum and rotational kinetic energy.

Ex 5.8 (a) The average density is simply the number of atoms per unit volume multiplied by the mass per atom, i.e.

$$\begin{aligned}\rho &= (10^{16} \text{ atoms cm}^{-3}) \times (1.67 \times 10^{-27} \text{ kg}) \\ &= 1.67 \times 10^{-11} \text{ kg cm}^{-3}\end{aligned}$$

(b) The average translational kinetic energy of the molecules is

$$\begin{aligned}\langle E_{\text{KE}} \rangle &= 3kT/2 \\ &= 1.5 \times (1.38 \times 10^{-23} \text{ J K}^{-1}) \times 5800 \text{ K} \\ &= 1.2 \times 10^{-19} \text{ J}\end{aligned}$$

This is equivalent to 0.75 eV.

(c) The pressure in the gas is found from

$$\begin{aligned}P &= NkT/V \\ &= (10^{16}) \times (1.38 \times 10^{-23} \text{ J K}^{-1}) \times 5800 \text{ K} / (10^{-6} \text{ m}^3) \\ &= 800 \text{ Pa}\end{aligned}$$

where we have used the fact that $1 \text{ cm}^3 = (10^{-2} \text{ m})^3 = 10^{-6} \text{ m}^3$ in order to get the value of V .

This is less than 1% of normal atmospheric pressure at sea-level on Earth.

Ex 5.9 The transitions which give rise to lines of the Balmer series all have a lowest energy level corresponding to $n = 2$, or $E_2 = -3.40 \text{ eV}$. The first four lines of the Balmer series involve transitions with a highest energy level of $n = 3, 4, 5$ and 6 or $E_n = -1.51 \text{ eV}, -0.85 \text{ eV}, -0.54 \text{ eV}$ and -0.38 eV respectively. So the energies of the photons corresponding to each of these transitions are

$$\begin{aligned}E_3 - E_2 &= (-1.51 + 3.40) \text{ eV} = 1.89 \text{ eV} \\ E_4 - E_2 &= (-0.85 + 3.40) \text{ eV} = 2.55 \text{ eV} \\ E_5 - E_2 &= (-0.54 + 3.40) \text{ eV} = 2.86 \text{ eV} \\ E_6 - E_2 &= (-0.38 + 3.40) \text{ eV} = 3.02 \text{ eV}\end{aligned}$$

Ex 5.10 Using $\Delta E \Delta t \geq \hbar/2$ since $\Delta t \simeq 10^{-8} \text{ s}$, $\Delta E \geq 5.5 \times 10^{-27} \text{ J}$. This short, but finite, time implies an indeterminacy in the energy of the excited atomic states. Such an indeterminacy in energy will result in a corresponding indeterminacy in the frequency and wavelength of spectral lines. (Comment: The intrinsic width, or spread, $\Delta\lambda$, of the wavelength of spectral lines due to this effect, is called the ‘natural width’ of the spectral lines.)

Ex 5.11

(a) Since the total energy of the confined particle is $9\hbar^2/(8m_e D^2)$, the wavefunction describing its behaviour must be characterized by the number $n = 3$. Hence, there are three half-wavelengths of the probability wave between the confining walls.

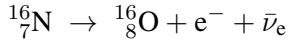
(b) The electron is most likely to be detected where $|\psi_{n=3}|^2$ is a maximum. There are three such maxima located at $x = 0$ and at $x = \pm D/3$.

(c) When an electron makes a transition between two energy levels, a single photon is ejected with an energy equal to the spacing of the two levels. So, in this case, a single photon would be ejected with energy

$$E = \frac{9\hbar^2}{8m_e D^2} - \frac{\hbar^2}{8m_e D^2} = \frac{\hbar^2}{m_e D^2}$$

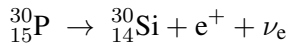
Ex 5.12 Since there are four more α -decays, the total decrease in mass number is $(4 \times 4) = 16$, and the total decrease in atomic number is $(4 \times 2) = 8$. So the resultant nucleus has $A = 230 - 16 = 214$ and $Z = 90 - 8 = 82$. The element with atomic number 82 is lead, as noted in the question, so the resulting nucleus is the lead isotope ${}_{82}^{214}\text{Pb}$.

Ex 5.13 The nitrogen isotope will undergo β^- -decay as follows:



Since the resulting nucleus has an atomic number of eight, this is an isotope of oxygen.

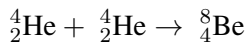
Ex 5.14 The phosphorus isotope will undergo β^+ -decay as follows:



Since the resulting nucleus has an atomic number of fourteen, this is an isotope of silicon.

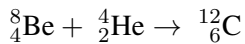
Ex 5.15 In the process of γ -decay, the number of protons and neutrons in the nucleus remains unchanged. So the atomic number and mass number of the barium nucleus after the γ -decay are the same as they were before, namely 56 and 137 respectively.

Ex 5.16 For the first stage of the reaction:



The energy deficit on the right-hand side is $3.7274 \text{ GeV} + 3.7274 \text{ GeV} - 7.4549 \text{ GeV} = -0.0001 \text{ GeV}$ or -0.1 MeV . Since the energy of the products is more than that of the reactants, this reaction is not energetically favoured.

The second stage of the reaction involves:



The energy surplus on the right-hand side is $3.7274 \text{ GeV} + 7.4549 \text{ GeV} - 11.1749 \text{ GeV} = +0.0074 \text{ GeV}$ or $+7.4 \text{ MeV}$. Since the energy of the products is less than that of the reactants, this reaction is energetically favoured.

The net energy released in the triple-alpha process is therefore $-0.1 \text{ MeV} + 7.4 \text{ MeV} = +7.3 \text{ MeV}$ per carbon-12 nucleus produced.

Ex 5.17 (a) The cyclotron radius of the electron is given by $r = mv/qB$ which in this case gives

$$\begin{aligned} r &= \frac{9.1 \times 10^{-31} \text{ kg} \times 3.0 \times 10^6 \text{ m s}^{-1}}{1.6 \times 10^{-19} \text{ C} \times 3.0 \times 10^8 \text{ T}} \\ &= 5.7 \times 10^{-14} \text{ m} \end{aligned}$$

(b) The period for the electron to complete one orbit is given by the circumference of the orbit divided by the electron's speed, i.e. $P = 2\pi r/v$, so the frequency at which the electron orbits is simply $f = 1/P = v/2\pi r$. However, since we know that $r = mv/qB$, the cyclotron frequency becomes

$f = (v/2\pi) \times (qB/mv) = qB/2\pi m$, which depends only on the magnetic field strength, and is independent of the electron's speed.

$$\begin{aligned} f &= qB/2\pi m \\ &= (1.6 \times 10^{-19} \text{ C} \times 3.0 \times 10^8 \text{ T}) / (2\pi \times 9.1 \times 10^{-31} \text{ kg}) \\ &= 8.4 \times 10^{18} \text{ Hz} \end{aligned}$$

Ex 5.18 (a) From Equation 5.92, the longest wavelength spectral line in each series corresponds to the line with the lowest photon energy (since $E_{\text{ph}} \propto 1/\lambda$).

The lowest energy line in the Lyman series corresponds to a transition between the $n = 1$ energy level and the $n = 2$ energy level. The energy of a photon corresponding to such a transition is therefore $(-3.40 + 13.60) \text{ eV} = 10.20 \text{ eV}$. The wavelength corresponding to this photon energy is

$$\begin{aligned} \lambda &= \frac{hc}{E_{\text{ph}}} \\ &= \frac{(4.14 \times 10^{-15} \text{ eV Hz}^{-1}) \times (3.00 \times 10^8 \text{ m s}^{-1})}{10.20 \text{ eV}} \\ &= 1.22 \times 10^{-7} \text{ m or } 122 \text{ nm} \end{aligned}$$

From Figure 5.25 this is in the ultraviolet part of the electromagnetic spectrum.

(b) The lowest energy line in the Paschen series corresponds to a transition between the $n = 3$ energy level and the $n = 4$ energy level. The energy of a photon corresponding to such a transition is therefore $(-0.85 + 1.51) \text{ eV} = 0.66 \text{ eV}$. The wavelength corresponding to this photon energy is

$$\begin{aligned} \lambda &= \frac{hc}{E_{\text{ph}}} \\ &= \frac{(4.14 \times 10^{-15} \text{ eV Hz}^{-1}) \times (3.00 \times 10^8 \text{ m s}^{-1})}{0.66 \text{ eV}} \\ &= 1.88 \times 10^{-6} \text{ m or } 1.88 \mu\text{m} \end{aligned}$$

From Figure 5.25 this is in the infrared part of the electromagnetic spectrum.

Ex 5.19 From Equations 5.95 and 5.96

$$\begin{aligned} \nu B_{\nu}(T) &= \nu \times \left(\frac{2h\nu^3}{c^2} \right) \times \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} \\ &= \left(\frac{2h\nu^4}{c^2} \right) \times \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1} \end{aligned}$$

and

$$\begin{aligned} \lambda B_{\lambda}(T) &= \lambda \times \left(\frac{2hc^2}{\lambda^5} \right) \times \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1} \\ &= \left(\frac{2hc^2}{\lambda^4} \right) \times \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1} \end{aligned}$$

Now we simply replace ν in the first equation by c/λ , to give

$$\begin{aligned}\nu B_\nu(T) &= \left(\frac{2hc^4}{\lambda^4 c^2}\right) \times \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1} \\ &= \left(\frac{2hc^2}{\lambda^4}\right) \times \frac{1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1}\end{aligned}$$

which is identical to the second equation as required.

Ex 5.20 (a) B_ν reaches a maximum at a wavelength given by

$$\begin{aligned}\lambda_{\max} &= 5.1 \times 10^{-3} \text{ m K}/10^8 \text{ K} \\ &= 5.1 \times 10^{-11} \text{ m or } 0.051 \text{ nm}\end{aligned}$$

This corresponds to a photon energy of

$$\begin{aligned}E_{\text{ph}} &= hc/\lambda \\ &= (6.63 \times 10^{-34} \text{ J s}) \times (3.00 \times 10^8 \text{ m s}^{-1}) / (5.1 \times 10^{-11} \text{ m}) \\ &= 3.90 \times 10^{-15} \text{ J} \\ &= (3.90 \times 10^{-15} / 1.60 \times 10^{-19}) \text{ eV} \\ &= 2.44 \times 10^4 \text{ eV} \\ &= 24 \text{ keV (to 2 s.f.)}\end{aligned}$$

(b) B_λ reaches a maximum at a wavelength given by

$$\begin{aligned}\lambda_{\max} &= 2.9 \times 10^{-3} \text{ m K}/10^8 \text{ K} \\ &= 2.9 \times 10^{-11} \text{ m or } 0.029 \text{ nm}\end{aligned}$$

This corresponds to a photon energy of

$$\begin{aligned}E_{\text{ph}} &= hc/\lambda \\ &= (6.63 \times 10^{-34} \text{ J s}) \times (3.00 \times 10^8 \text{ m s}^{-1}) / (2.9 \times 10^{-11} \text{ m}) \\ &= 6.86 \times 10^{-15} \text{ J} \\ &= (6.86 \times 10^{-15} / 1.60 \times 10^{-19}) \text{ eV} \\ &= 4.29 \times 10^4 \text{ eV} \\ &= 43 \text{ keV (to 2 s.f.)}\end{aligned}$$

(c) The mean photon energy of the spectrum is given by

$$\begin{aligned}\langle E_{\text{ph}} \rangle &= 2.7kT \\ &= 2.7 \times (1.38 \times 10^{-23} \text{ J K}^{-1}) \times 10^8 \text{ K} \\ &= 3.73 \times 10^{-15} \text{ J} \\ &= (3.73 \times 10^{-15} / 1.60 \times 10^{-19}) \text{ eV} \\ &= 2.33 \times 10^4 \text{ eV} \\ &= 23 \text{ keV (to 2 s.f.)}\end{aligned}$$

(d) Photon energies of a few tens of keV, or equivalently, wavelengths of a few hundredths of a nanometre, correspond to the X-ray part of the electromagnetic spectrum.