some sort of relationship between successive terms. For example, if the nth term of a series is given by

$$u_n = \frac{1}{2^n},$$

for $n = 1, 2, 3, \ldots, N$ then the sum of the first N terms will be

$$S_N = \sum_{n=1}^{N} u_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^N}. \tag{4.1}$$

It is clear that the sum of a finite number of terms is always finite, provided that each term is itself finite. It is often of practical interest, however, to consider the sum of a series with an infinite number of finite terms. The sum of an infinite number of terms is best defined by first considering the partial sum of the first N terms, S_N. If the value of the partial sum S_N tends to a finite limit, S, as N tends to infinity, then the series is said to converge and its sum is given by the limit S. In other words, the sum of an infinite series is given by

$$S = \lim_{N\to\infty} S_N,$$

provided the limit exists. For complex infinite series, if S_N approaches a limit $S = X + iY$ as $N \to \infty$, this means that $X_N \to X$ and $Y_N \to Y$ separately, i.e. the real and imaginary parts of the series are each convergent series with sums X and Y respectively.

However, not all infinite series have finite sums. As $N \to \infty$, the value of the partial sum S_N may diverge: it may approach $+\infty$ or $-\infty$, or oscillate finitely or infinitely. Moreover, for a series where each term depends on some variable, its convergence can depend on the value assumed by the variable. Whether an infinite series converges, diverges or oscillates has important implications when describing physical systems. Methods for determining whether a series converges are discussed in section 4.3.

4.2 Summation of series

It is often necessary to find the sum of a finite series or a convergent infinite series. We now describe arithmetic, geometric and arithmetico-geometric series, which are particularly common and for which the sums are easily found. Other methods that can sometimes be used to sum more complicated series are discussed below.

4.2.1 Arithmetic series

An *arithmetic series* has the characteristic that the difference between successive terms is constant. The sum of a general arithmetic series is written

$$S_N = a + (a+d) + (a+2d) + \cdots + [a + (N-1)d] = \sum_{n=0}^{N-1}(a+nd).$$

Rewriting the series in the opposite order and adding this term by term to the original expression for S_N, we find

$$S_N = \frac{N}{2}\,[a + a + (N-1)d] = \frac{N}{2}(\text{first term} + \text{last term}). \qquad (4.2)$$

If an infinite number of such terms are added the series will increase (or decrease) indefinitely; that is to say, it diverges.

►*Sum the integers between* 1 *and* 1000 *inclusive.*

This is an arithmetic series with $a = 1$, $d = 1$ and $N = 1000$. Therefore, using (4.2) we find

$$S_N = \frac{1000}{2}(1 + 1000) = 500500,$$

which can be checked directly only with considerable effort. ◄

4.2.2 Geometric series

Equation (4.1) is a particular example of a *geometric series*, which has the characteristic that the ratio of successive terms is a constant (one-half in this case). The sum of a geometric series is in general written

$$S_N = a + ar + ar^2 + \cdots + ar^{N-1} = \sum_{n=0}^{N-1} ar^n,$$

where a is a constant and r is the ratio of successive terms, the *common ratio*. The sum may be evaluated by considering S_N and rS_N:

$$S_N = a + ar + ar^2 + ar^3 + \cdots + ar^{N-1},$$
$$rS_N = ar + ar^2 + ar^3 + ar^4 + \cdots + ar^N.$$

If we now subtract the second equation from the first we obtain

$$(1-r)S_N = a - ar^N,$$

and hence

$$S_N = \frac{a(1-r^N)}{1-r}. \qquad (4.3)$$

For a series with an infinite number of terms and $|r| < 1$, we have $\lim_{N\to\infty} r^N = 0$, and the sum tends to the limit

$$S = \frac{a}{1-r}. \tag{4.4}$$

In (4.1), $r = \frac{1}{2}$, $a = \frac{1}{2}$, and so $S = 1$. For $|r| \geq 1$, however, the series either diverges or oscillates.

►*Consider a ball that drops from a height of* 27 m *and on each bounce retains only a third of its kinetic energy; thus after one bounce it will return to a height of* 9 m*, after two bounces to* 3 m*, and so on. Find the total distance travelled between the first bounce and the Mth bounce.*

The total distance travelled between the first bounce and the Mth bounce is given by the sum of $M-1$ terms:

$$S_{M-1} = 2\,(9+3+1+\cdots) = 2\sum_{m=0}^{M-2}\frac{9}{3^m}$$

for $M > 1$, where the factor 2 is included to allow for both the upward and the downward journey. Inside the parentheses we clearly have a geometric series with first term 9 and common ratio $1/3$ and hence the distance is given by (4.3), i.e.

$$S_{M-1} = 2 \times \frac{9\left[1-\left(\frac{1}{3}\right)^{M-1}\right]}{1-\frac{1}{3}} = 27\left[1-\left(\tfrac{1}{3}\right)^{M-1}\right],$$

where the number of terms N in (4.3) has been replaced by $M-1$. ◄

4.2.3 Arithmetico-geometric series

An arithmetico-geometric series, as its name suggests, is a combined arithmetic and geometric series. It has the general form

$$S_N = a + (a+d)r + (a+2d)r^2 + \cdots + [a+(N-1)d]\,r^{N-1} = \sum_{n=0}^{N-1}(a+nd)r^n,$$

and can be summed, in a similar way to a pure geometric series, by multiplying by r and subtracting the result from the original series to obtain

$$(1-r)S_N = a + rd + r^2d + \cdots + r^{N-1}d - [a+(N-1)d]\,r^N.$$

Using the expression for the sum of a geometric series (4.3) and rearranging, we find

$$S_N = \frac{a-[a+(N-1)d]\,r^N}{1-r} + \frac{rd(1-r^{N-1})}{(1-r)^2}.$$

For an infinite series with $|r| < 1$, $\lim_{N\to\infty} r^N = 0$ as in the previous subsection, and the sum tends to the limit

$$S = \frac{a}{1-r} + \frac{rd}{(1-r)^2}. \tag{4.5}$$

As for a geometric series, if $|r| \geq 1$ then the series either diverges or oscillates.

►*Sum the series*

$$S = 2 + \frac{5}{2} + \frac{8}{2^2} + \frac{11}{2^3} + \cdots .$$

This is an infinite arithmetico-geometric series with $a = 2$, $d = 3$ and $r = 1/2$. Therefore, from (4.5), we obtain $S = 10$. ◄

4.2.4 The difference method

The difference method is sometimes useful in summing series that are more complicated than the examples discussed above. Let us consider the general series

$$\sum_{n=1}^{N} u_n = u_1 + u_2 + \cdots + u_N.$$

If the terms of the series, u_n, can be expressed in the form

$$u_n = f(n) - f(n-1)$$

for some function $f(n)$ then its (partial) sum is given by

$$S_N = \sum_{n=1}^{N} u_n = f(N) - f(0).$$

This can be shown as follows. The sum is given by

$$S_N = u_1 + u_2 + \cdots + u_N$$

and since $u_n = f(n) - f(n-1)$, it may be rewritten

$$S_N = [f(1) - f(0)] + [f(2) - f(1)] + \cdots + [f(N) - f(N-1)].$$

By cancelling terms we see that

$$S_N = f(N) - f(0).$$

►*Evaluate the sum*

$$\sum_{n=1}^{N} \frac{1}{n(n+1)}.$$

Using partial fractions we find

$$u_n = -\left(\frac{1}{n+1} - \frac{1}{n}\right).$$

Hence $u_n = f(n) - f(n-1)$ with $f(n) = -1/(n+1)$, and so the sum is given by

$$S_N = f(N) - f(0) = -\frac{1}{N+1} + 1 = \frac{N}{N+1}. \blacktriangleleft$$

The difference method may be easily extended to evaluate sums in which each term can be expressed in the form

$$u_n = f(n) - f(n-m), \tag{4.6}$$

where m is an integer. By writing out the sum to N terms with each term expressed in this form, and cancelling terms in pairs as before, we find

$$S_N = \sum_{k=1}^{m} f(N-k+1) - \sum_{k=1}^{m} f(1-k).$$

►*Evaluate the sum*

$$\sum_{n=1}^{N} \frac{1}{n(n+2)}.$$

Using partial fractions we find

$$u_n = -\left[\frac{1}{2(n+2)} - \frac{1}{2n}\right].$$

Hence $u_n = f(n) - f(n-2)$ with $f(n) = -1/[2(n+2)]$, and so the sum is given by

$$S_N = f(N) + f(N-1) - f(0) - f(-1) = \frac{3}{4} - \frac{1}{2}\left(\frac{1}{N+2} + \frac{1}{N+1}\right). \blacktriangleleft$$

In fact the difference method is quite flexible and may be used to evaluate sums even when each term cannot be expressed as in (4.6). The method still relies, however, on being able to write u_n in terms of a single function such that most terms in the sum cancel, leaving only a few terms at the beginning and the end. This is best illustrated by an example.

►*Evaluate the sum*

$$\sum_{n=1}^{N} \frac{1}{n(n+1)(n+2)}.$$

Using partial fractions we find

$$u_n = \frac{1}{2(n+2)} - \frac{1}{n+1} + \frac{1}{2n}.$$

Hence $u_n = f(n) - 2f(n-1) + f(n-2)$ with $f(n) = 1/[2(n+2)]$. If we write out the sum, expressing each term u_n in this form, we find that most terms cancel and the sum is given by

$$S_N = f(N) - f(N-1) - f(0) + f(-1) = \frac{1}{4} + \frac{1}{2}\left(\frac{1}{N+2} - \frac{1}{N+1}\right). \blacktriangleleft$$

4.2.5 Series involving natural numbers

Series consisting of the natural numbers 1, 2, 3, …, or the square or cube of these numbers, occur frequently and deserve a special mention. Let us first consider the sum of the first N natural numbers,

$$S_N = 1 + 2 + 3 + \cdots + N = \sum_{n=1}^{N} n.$$

This is clearly an arithmetic series with first term $a = 1$ and common difference $d = 1$. Therefore, from (4.2), $S_N = \frac{1}{2}N(N+1)$.

Next, we consider the sum of the squares of the first N natural numbers:

$$S_N = 1^2 + 2^2 + 3^2 + \ldots + N^2 = \sum_{n=1}^{N} n^2,$$

which may be evaluated using the difference method. The nth term in the series is $u_n = n^2$, which we need to express in the form $f(n) - f(n-1)$ for some function $f(n)$. Consider the function

$$f(n) = n(n+1)(2n+1) \qquad \Rightarrow \qquad f(n-1) = (n-1)n(2n-1).$$

For this function $f(n) - f(n-1) = 6n^2$, and so we can write

$$u_n = \tfrac{1}{6}[f(n) - f(n-1)].$$

Therefore, by the difference method,

$$S_N = \tfrac{1}{6}[f(N) - f(0)] = \tfrac{1}{6}N(N+1)(2N+1).$$

Finally, we calculate the sum of the cubes of the first N natural numbers,

$$S_N = 1^3 + 2^3 + 3^3 + \cdots + N^3 = \sum_{n=1}^{N} n^3,$$

again using the difference method. Consider the function

$$f(n) = [n(n+1)]^2 \qquad \Rightarrow \qquad f(n-1) = [(n-1)n]^2,$$

for which $f(n) - f(n-1) = 4n^3$. Therefore we can write the general nth term of the series as

$$u_n = \tfrac{1}{4}[f(n) - f(n-1)],$$

and using the difference method we find

$$S_N = \tfrac{1}{4}[f(N) - f(0)] = \tfrac{1}{4}N^2(N+1)^2.$$

Note that this is the square of the sum of the natural numbers, i.e.

$$\sum_{n=1}^{N} n^3 = \left(\sum_{n=1}^{N} n\right)^2.$$

►*Sum the series*

$$\sum_{n=1}^{N}(n+1)(n+3).$$

The nth term in this series is

$$u_n = (n+1)(n+3) = n^2 + 4n + 3,$$

and therefore we can write

$$\begin{aligned}\sum_{n=1}^{N}(n+1)(n+3) &= \sum_{n=1}^{N}(n^2+4n+3)\\ &= \sum_{n=1}^{N} n^2 + 4\sum_{n=1}^{N} n + \sum_{n=1}^{N} 3\\ &= \tfrac{1}{6}N(N+1)(2N+1) + 4\times\tfrac{1}{2}N(N+1) + 3N\\ &= \tfrac{1}{6}N(2N^2+15N+31). \blacktriangleleft\end{aligned}$$

4.2.6 Transformation of series

A complicated series may sometimes be summed by transforming it into a familiar series for which we already know the sum, perhaps a geometric series or the Maclaurin expansion of a simple function (see subsection 4.6.3). Various techniques are useful, and deciding which one to use in any given case is a matter of experience. We now discuss a few of the more common methods.

The differentiation or integration of a series is often useful in transforming an apparently intractable series into a more familiar one. If we wish to differentiate or integrate a series that already depends on some variable then we may do so in a straightforward manner.

►*Sum the series*

$$S(x) = \frac{x^4}{3(0!)} + \frac{x^5}{4(1!)} + \frac{x^6}{5(2!)} + \cdots.$$

Dividing both sides by x we obtain

$$\frac{S(x)}{x} = \frac{x^3}{3(0!)} + \frac{x^4}{4(1!)} + \frac{x^5}{5(2!)} + \cdots,$$

which is easily differentiated to give

$$\frac{d}{dx}\left[\frac{S(x)}{x}\right] = \frac{x^2}{0!} + \frac{x^3}{1!} + \frac{x^4}{2!} + \frac{x^5}{3!} + \cdots.$$

Recalling the Maclaurin expansion of $\exp x$ given in subsection 4.6.3, we recognise that the RHS is equal to $x^2 \exp x$. Having done so, we can now integrate both sides to obtain

$$S(x)/x = \int x^2 \exp x \, dx.$$

Integrating the RHS by parts we find

$$S(x)/x = x^2 \exp x - 2x \exp x + 2 \exp x + c,$$

where the value of the constant of integration c can be fixed by the requirement that $S(x)/x = 0$ at $x = 0$. Thus we find that $c = -2$ and that the sum is given by

$$S(x) = x^3 \exp x - 2x^2 \exp x + 2x \exp x - 2x. \blacktriangleleft$$

Often, however, we require the sum of a series that does not depend on a variable. In this case, in order that we may differentiate or integrate the series, we define a function of some variable x such that the value of this function is equal to the sum of the series for some particular value of x (usually at $x = 1$).

►*Sum the series*

$$S = 1 + \frac{2}{2} + \frac{3}{2^2} + \frac{4}{2^3} + \cdots .$$

Let us begin by defining the function

$$f(x) = 1 + 2x + 3x^2 + 4x^3 + \cdots ,$$

so that the sum $S = f(1/2)$. Integrating this function we obtain

$$\int f(x)\, dx = x + x^2 + x^3 + \cdots ,$$

which we recognise as an infinite geometric series with first term $a = x$ and common ratio $r = x$. Therefore, from (4.4), we find that the sum of this series is $x/(1-x)$. In other words

$$\int f(x)\, dx = \frac{x}{1-x},$$

so that $f(x)$ is given by

$$f(x) = \frac{d}{dx}\left(\frac{x}{1-x}\right) = \frac{1}{(1-x)^2}.$$

The sum of the original series is therefore $S = f(1/2) = 4$. ◄

Aside from differentiation and integration, an appropriate substitution can sometimes transform a series into a more familiar form. In particular, series with terms that contain trigonometric functions can often be summed by the use of complex exponentials.

►*Sum the series*

$$S(\theta) = 1 + \cos\theta + \frac{\cos 2\theta}{2!} + \frac{\cos 3\theta}{3!} + \cdots .$$

Replacing the cosine terms with a complex exponential, we obtain

$$\begin{aligned} S(\theta) &= \mathrm{Re}\left\{ 1 + \exp i\theta + \frac{\exp 2i\theta}{2!} + \frac{\exp 3i\theta}{3!} + \cdots \right\} \\ &= \mathrm{Re}\left\{ 1 + \exp i\theta + \frac{(\exp i\theta)^2}{2!} + \frac{(\exp i\theta)^3}{3!} + \cdots \right\}. \end{aligned}$$

Again using the Maclaurin expansion of $\exp x$ given in subsection 4.6.3, we notice that

$$\begin{aligned} S(\theta) &= \text{Re}\,[\exp(\exp i\theta)] = \text{Re}\,[\exp(\cos\theta + i\sin\theta)] \\ &= \text{Re}\,\{[\exp(\cos\theta)][\exp(i\sin\theta)]\} = [\exp(\cos\theta)]\text{Re}\,[\exp(i\sin\theta)] \\ &= [\exp(\cos\theta)][\cos(\sin\theta)]. \blacktriangleleft \end{aligned}$$

4.3 Convergence of infinite series

Although the sums of some commonly occurring infinite series may be found, the sum of a general infinite series is usually difficult to calculate. Nevertheless, it is often useful to know whether the partial sum of such a series converges to a limit, even if the limit cannot be found explicitly. As mentioned at the end of section 4.1, if we allow N to tend to infinity, the partial sum

$$S_N = \sum_{n=1}^{N} u_n$$

of a series may tend to a definite limit (i.e. the sum S of the series), or increase or decrease without limit, or oscillate finitely or infinitely.

To investigate the convergence of any given series, it is useful to have available a number of tests and theorems of general applicability. We discuss them below; some we will merely state, since once they have been stated they become almost self-evident, but are no less useful for that.

4.3.1 Absolute and conditional convergence

Let us first consider some general points concerning the convergence, or otherwise, of an infinite series. In general an infinite series $\sum u_n$ can have complex terms, and in the special case of a real series the terms can be positive or negative. From any such series, however, we can always construct another series $\sum |u_n|$ in which each term is simply the modulus of the corresponding term in the original series. Then each term in the new series will be a positive real number.

If the series $\sum |u_n|$ converges then $\sum u_n$ also converges, and $\sum u_n$ is said to be *absolutely convergent*, i.e. the series formed by the absolute values is convergent. For an absolutely convergent series, the terms may be reordered without affecting the convergence of the series. However, if $\sum |u_n|$ diverges whilst $\sum u_n$ converges then $\sum u_n$ is said to be *conditionally convergent*. For a conditionally convergent series, rearranging the order of the terms can affect the behaviour of the sum and, hence, whether the series converges or diverges. In fact, a theorem due to Riemann shows that, by a suitable rearrangement, a conditionally convergent series may be made to converge to any arbitrary limit, or to diverge, or to oscillate finitely or infinitely! Of course, if the original series $\sum u_n$ consists only of positive real terms and converges then automatically it is absolutely convergent.

4.3.2 Convergence of a series containing only real positive terms

As discussed above, in order to test for the absolute convergence of a series $\sum u_n$, we first construct the corresponding series $\sum |u_n|$ that consists only of real positive terms. Therefore in this subsection we will restrict our attention to series of this type.

We discuss below some tests that may be used to investigate the convergence of such a series. Before doing so, however, we note the following *crucial consideration*. In all the tests for, or discussions of, the convergence of a series, it is not what happens in the first ten, or the first thousand, or the first million terms (or any other finite number of terms) that matters, but what happens *ultimately*.

Preliminary test

A necessary *but not sufficient* condition for a series of real positive terms $\sum u_n$ to be convergent is that the term u_n tends to zero as n tends to infinity, i.e. we require

$$\lim_{n\to\infty} u_n = 0.$$

If this condition is not satisfied then the series must diverge. Even if it is satisfied, however, the series may still diverge, and further testing is required.

Comparison test

The comparison test is the most basic test for convergence. Let us consider two series $\sum u_n$ and $\sum v_n$ and suppose that we *know* the latter to be convergent (by some earlier analysis, for example). Then, if each term u_n in the first series is less than or equal to the corresponding term v_n in the second series, for all n greater than some fixed number N that will vary from series to series, then the original series $\sum u_n$ is also convergent. In other words, if $\sum v_n$ is convergent and

$$u_n \leq v_n \qquad \text{for } n > N,$$

then $\sum u_n$ converges.

However, if $\sum v_n$ diverges and $u_n \geq v_n$ for all n greater than some fixed number then $\sum u_n$ diverges.

►*Determine whether the following series converges:*

$$\sum_{n=1}^{\infty} \frac{1}{n!+1} = \frac{1}{2} + \frac{1}{3} + \frac{1}{7} + \frac{1}{25} + \cdots. \tag{4.7}$$

Let us compare this series with the series

$$\sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots = 2 + \frac{1}{2!} + \frac{1}{3!} + \cdots, \tag{4.8}$$

which is merely the series obtained by setting $x = 1$ in the Maclaurin expansion of $\exp x$ (see subsection 4.6.3), i.e.

$$\exp(1) = e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots .$$

Clearly this second series is convergent, since it consists of only positive terms and has a finite sum. Thus, since each term u_n in the series (4.7) is less than the corresponding term $1/n!$ in (4.8), we conclude from the comparison test that (4.7) is also convergent. ◄

D'Alembert's ratio test

The ratio test determines whether a series converges by comparing the relative magnitude of successive terms. If we consider a series $\sum u_n$ and set

$$\rho = \lim_{n\to\infty} \left(\frac{u_{n+1}}{u_n} \right), \tag{4.9}$$

then if $\rho < 1$ the series is convergent; if $\rho > 1$ the series is divergent; if $\rho = 1$ then the behaviour of the series is undetermined by this test.

To prove this we observe that if the limit (4.9) is less than unity, i.e. $\rho < 1$ then we can find a value r in the range $\rho < r < 1$ and a value N such that

$$\frac{u_{n+1}}{u_n} < r,$$

for all $n > N$. Now the terms u_n of the series that follow u_N are

$$u_{N+1}, \qquad u_{N+2}, \qquad u_{N+3}, \qquad \ldots,$$

and each of these is less than the corresponding term of

$$ru_N, \qquad r^2u_N, \qquad r^3u_N, \qquad \ldots . \tag{4.10}$$

However, the terms of (4.10) are those of a geometric series with a common ratio r that is less than unity. This geometric series consequently converges and therefore, by the comparison test discussed above, so must the original series $\sum u_n$. An analogous argument may be used to prove the divergent case when $\rho > 1$.

►*Determine whether the following series converges:*

$$\sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots = 2 + \frac{1}{2!} + \frac{1}{3!} + \cdots .$$

As mentioned in the previous example, this series may be obtained by setting $x = 1$ in the Maclaurin expansion of $\exp x$, and hence we know already that it converges and has the sum $\exp(1) = e$. Nevertheless, we may use the ratio test to confirm that it converges.

Using (4.9), we have

$$\rho = \lim_{n\to\infty} \left[\frac{n!}{(n+1)!} \right] = \lim_{n\to\infty} \left(\frac{1}{n+1} \right) = 0 \tag{4.11}$$

and since $\rho < 1$, the series converges, as expected. ◄

Ratio comparison test

As its name suggests, the ratio comparison test is a combination of the ratio and comparison tests. Let us consider the two series $\sum u_n$ and $\sum v_n$ and assume that we know the latter to be convergent. It may be shown that if

$$\frac{u_{n+1}}{u_n} \leq \frac{v_{n+1}}{v_n}$$

for all n greater than some fixed value N then $\sum u_n$ is also convergent.

Similarly, if

$$\frac{u_{n+1}}{u_n} \geq \frac{v_{n+1}}{v_n}$$

for all sufficiently large n, and $\sum v_n$ diverges then $\sum u_n$ also diverges.

►*Determine whether the following series converges:*

$$\sum_{n=1}^{\infty} \frac{1}{(n!)^2} = 1 + \frac{1}{2^2} + \frac{1}{6^2} + \cdots .$$

In this case the ratio of successive terms, as n tends to infinity, is given by

$$R = \lim_{n\to\infty} \left[\frac{n!}{(n+1)!}\right]^2 = \lim_{n\to\infty} \left(\frac{1}{n+1}\right)^2,$$

which is less than the ratio seen in (4.11). Hence, by the ratio comparison test, the series converges. (It is clear that this series could also be found to be convergent using the ratio test.) ◄

Quotient test

The quotient test may also be considered as a combination of the ratio and comparison tests. Let us again consider the two series $\sum u_n$ and $\sum v_n$, and define ρ as the limit

$$\rho = \lim_{n\to\infty} \left(\frac{u_n}{v_n}\right). \tag{4.12}$$

Then, it can be shown that:

(i) if $\rho \neq 0$ but is finite then $\sum u_n$ and $\sum v_n$ either both converge or both diverge;

(ii) if $\rho = 0$ and $\sum v_n$ converges then $\sum u_n$ converges;

(iii) if $\rho = \infty$ and $\sum v_n$ diverges then $\sum u_n$ diverges.

►*Given that the series $\sum_{n=1}^{\infty} 1/n$ diverges, determine whether the following series converges:*

$$\sum_{n=1}^{\infty} \frac{4n^2 - n - 3}{n^3 + 2n}. \tag{4.13}$$

If we set $u_n = (4n^2 - n - 3)/(n^3 + 2n)$ and $v_n = 1/n$ then the limit (4.12) becomes

$$\rho = \lim_{n\to\infty} \left[\frac{(4n^2 - n - 3)/(n^3 + 2n)}{1/n} \right] = \lim_{n\to\infty} \left[\frac{4n^3 - n^2 - 3n}{n^3 + 2n} \right] = 4.$$

Since ρ is finite but non-zero and $\sum v_n$ diverges, from (i) above $\sum u_n$ must also diverge. ◄

Integral test

The integral test is an extremely powerful means of investigating the convergence of a series $\sum u_n$. Suppose that there exists a function $f(x)$ which monotonically decreases for x greater than some fixed value x_0 and for which $f(n) = u_n$, i.e. the value of the function at integer values of x is equal to the corresponding term in the series under investigation. Then it can be shown that, if the limit of the integral

$$\lim_{N\to\infty} \int^{N} f(x)\,dx$$

exists, the series $\sum u_n$ is convergent. Otherwise the series diverges. Note that the integral defined here has no lower limit; the test is sometimes stated with a lower limit, equal to unity, for the integral, but this can lead to unnecessary difficulties.

►*Determine whether the following series converges:*

$$\sum_{n=1}^{\infty} \frac{1}{(n - 3/2)^2} = 4 + 4 + \frac{4}{9} + \frac{4}{25} + \cdots .$$

Let us consider the function $f(x) = (x - 3/2)^{-2}$. Clearly $f(n) = u_n$ and $f(x)$ monotonically decreases for $x > 3/2$. Applying the integral test, we consider

$$\lim_{N\to\infty} \int^{N} \frac{1}{(x - 3/2)^2}\,dx = \lim_{N\to\infty} \left(\frac{-1}{N - 3/2} \right) = 0.$$

Since the limit exists the series converges. Note, however, that if we had included a lower limit, equal to unity, in the integral then we would have run into problems, since the integrand diverges at $x = 3/2$. ◄

The integral test is also useful for examining the convergence of the Riemann zeta series. This is a special series that occurs regularly and is of the form

$$\sum_{n=1}^{\infty} \frac{1}{n^p}.$$

It converges for $p > 1$ and diverges if $p \le 1$. These convergence criteria may be derived as follows.

Using the integral test, we consider

$$\lim_{N\to\infty}\int^{N}\frac{1}{x^p}dx=\lim_{N\to\infty}\left(\frac{N^{1-p}}{1-p}\right),$$

and it is obvious that the limit tends to zero for $p > 1$ and to ∞ for $p \leq 1$.

Cauchy's root test

Cauchy's root test may be useful in testing for convergence, especially if the nth terms of the series contains an nth power. If we define the limit

$$\rho=\lim_{n\to\infty}(u_n)^{1/n},$$

then it may be proved that the series $\sum u_n$ converges if $\rho < 1$. If $\rho > 1$ then the series diverges. Its behaviour is undetermined if $\rho = 1$.

►*Determine whether the following series converges:*

$$\sum_{n=1}^{\infty}\left(\frac{1}{n}\right)^n=1+\frac{1}{4}+\frac{1}{27}+\cdots.$$

Using Cauchy's root test, we find

$$\rho=\lim_{n\to\infty}\left(\frac{1}{n}\right)=0,$$

and hence the series converges. ◄

Grouping terms

We now consider the Riemann zeta series, mentioned above, with an alternative proof of its convergence that uses the method of grouping terms. In general there are better ways of determining convergence, but the grouping method may be used if it is not immediately obvious how to approach a problem by a better method.

First consider the case where $p > 1$, and group the terms in the series as follows:

$$S_N=\frac{1}{1^p}+\left(\frac{1}{2^p}+\frac{1}{3^p}\right)+\left(\frac{1}{4^p}+\cdots+\frac{1}{7^p}\right)+\cdots.$$

Now we can see that each bracket of this series is less than each term of the geometric series

$$S_N=\frac{1}{1^p}+\frac{2}{2^p}+\frac{4}{4^p}+\cdots.$$

This geometric series has common ratio $r=\left(\frac{1}{2}\right)^{p-1}$; since $p > 1$, it follows that $r < 1$ and that the geometric series converges. Then the comparison test shows that the Riemann zeta series also converges for $p > 1$.

The divergence of the Riemann zeta series for $p \leq 1$ can be seen by first considering the case $p = 1$. The series is

$$S_N = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots,$$

which does *not* converge, as may be seen by bracketing the terms of the series in groups in the following way:

$$S_N = \sum_{n=1}^{N} u_n = 1 + \left(\frac{1}{2}\right) + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \cdots.$$

The sum of the terms in each bracket is $\geq \frac{1}{2}$ and, since as many such groupings can be made as we wish, it is clear that S_N increases indefinitely as N is increased.

Now returning to the case of the Riemann zeta series for $p < 1$, we note that each term in the series is greater than the corresponding one in the series for which $p = 1$. In other words $1/n^p > 1/n$ for $n > 1$, $p < 1$. The comparison test then shows us that the Riemann zeta series will diverge for all $p \leq 1$.

4.3.3 Alternating series test

The tests discussed in the last subsection have been concerned with determining whether the series of real positive terms $\sum |u_n|$ converges, and so whether $\sum u_n$ is absolutely convergent. Nevertheless, it is sometimes useful to consider whether a series is merely convergent rather than absolutely convergent. This is especially true for series containing an infinite number of both positive and negative terms. In particular, we will consider the convergence of series in which the positive and negative terms alternate, i.e. an *alternating series.*

An alternating series can be written as

$$\sum_{n=1}^{\infty} (-1)^{n+1} u_n = u_1 - u_2 + u_3 - u_4 + u_5 - \cdots,$$

with all $u_n \geq 0$. Such a series can be shown to converge provided (i) $u_n \to 0$ as $n \to \infty$ and (ii) $u_n < u_{n-1}$ for all $n > N$ for some finite N. If these conditions are not met then the series oscillates.

To prove this, suppose for definiteness that N is odd and consider the series starting at u_N. The sum of its first $2m$ terms is

$$S_{2m} = (u_N - u_{N+1}) + (u_{N+2} - u_{N+3}) + \cdots + (u_{N+2m-2} - u_{N+2m-1}).$$

By condition (ii) above, all the parentheses are positive, and so S_{2m} increases as m increases. We can also write, however,

$$S_{2m} = u_N - (u_{N+1} - u_{N+2}) - \cdots - (u_{N+2m-3} - u_{N+2m-2}) - u_{N+2m-1},$$

and since each parenthesis is positive, we must have $S_{2m} < u_N$. Thus, since S_{2m}

is always less than u_N for all m and $u_n \to 0$ as $n \to \infty$, the alternating series converges. It is clear that an analogous proof can be constructed in the case where N is even.

►*Determine whether the following series converges:*

$$\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \cdots .$$

This alternating series clearly satisfies conditions (i) and (ii) above and hence converges. However, as shown above by the method of grouping terms, the corresponding series with all positive terms is divergent. ◄

4.4 Operations with series

Simple operations with series are fairly intuitive, and we discuss them here only for completeness. The following points apply to both finite and infinite series unless otherwise stated.

(i) If $\sum u_n = S$ then $\sum ku_n = kS$ where k is any constant.

(ii) If $\sum u_n = S$ and $\sum v_n = T$ then $\sum(u_n + v_n) = S + T$.

(iii) If $\sum u_n = S$ then $a + \sum u_n = a + S$. A simple extension of this trivial result shows that the removal or insertion of a finite number of terms anywhere in a series does not affect its convergence.

(iv) If the infinite series $\sum u_n$ and $\sum v_n$ are both absolutely convergent then the series $\sum w_n$, where

$$w_n = u_1 v_n + u_2 v_{n-1} + \cdots + u_n v_1,$$

is also absolutely convergent. The series $\sum w_n$ is called the *Cauchy product* of the two original series. Furthermore, if $\sum u_n$ converges to the sum S and $\sum v_n$ converges to the sum T then $\sum w_n$ converges to the sum ST.

(v) It is not true in general that term-by-term differentiation or integration of a series will result in a new series with the same convergence properties.

4.5 Power series

A power series has the form

$$P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots ,$$

where a_0, a_1, a_2, a_3 etc. are constants. Such series regularly occur in physics and engineering and are useful because, for $|x| < 1$, the later terms in the series may become very small and be discarded. For example the series

$$P(x) = 1 + x + x^2 + x^3 + \cdots ,$$

although in principle infinitely long, in practice may be simplified if x happens to have a value small compared with unity. To see this note that $P(x)$ for $x = 0.1$ has the following values: 1, if just one term is taken into account; 1.1, for two terms; 1.11, for three terms; 1.111, for four terms, etc. If the quantity that it represents can only be measured with an accuracy of two decimal places, then all but the first three terms may be ignored, i.e. when $x = 0.1$ or less

$$P(x) = 1 + x + x^2 + \mathrm{O}(x^3) \approx 1 + x + x^2.$$

This sort of approximation is often used to simplify equations into manageable forms. It may seem imprecise at first but is perfectly acceptable insofar as it matches the experimental accuracy that can be achieved.

The symbols O and $\approx$ used above need some further explanation. They are used to compare the behaviour of two functions when a variable upon which both functions depend tends to a particular limit, usually zero or infinity (and obvious from the context). For two functions $f(x)$ and $g(x)$, with g positive, the formal *definitions* of the above symbols are as follows:

(i) If there exists a constant k such that $|f| \leq kg$ as the limit is approached then $f = \mathrm{O}(g)$.

(ii) If as the limit of x is approached f/g tends to a limit l, where $l \neq 0$, then $f \approx lg$. The statement $f \approx g$ means that the ratio of the two sides tends to unity.

4.5.1 Convergence of power series

The convergence or otherwise of power series is a crucial consideration in practical terms. For example, if we are to use a power series as an approximation, it is clearly important that it tends to the precise answer as more and more terms of the approximation are taken. Consider the general power series

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots .$$

Using d'Alembert's ratio test (see subsection 4.3.2), we see that $P(x)$ converges absolutely if

$$\rho = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} x \right| = |x| \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| < 1.$$

Thus the convergence of $P(x)$ depends upon the value of x, i.e. there is, in general, a range of values of x for which $P(x)$ converges, an *interval of convergence*. Note that at the limits of this range $\rho = 1$, and so the series may converge or diverge. The convergence of the series at the end-points may be determined by substituting these values of x into the power series $P(x)$ and testing the resulting series using any applicable method (discussed in section 4.3).

►*Determine the range of values of x for which the following power series converges:*

$$P(x) = 1 + 2x + 4x^2 + 8x^3 + \cdots .$$

By using the interval-of-convergence method discussed above,

$$\rho = \lim_{n\to\infty} \left| \frac{2^{n+1}}{2^n} x \right| = |2x|,$$

and hence the power series will converge for $|x| < 1/2$. Examining the end-points of the interval separately, we find

$$\begin{aligned} P(1/2) &= 1 + 1 + 1 + \cdots , \\ P(-1/2) &= 1 - 1 + 1 - \cdots . \end{aligned}$$

Obviously $P(1/2)$ diverges, while $P(-1/2)$ oscillates. Therefore $P(x)$ is not convergent at either end-point of the region but is convergent for $-1 < x < 1$. ◄

The convergence of power series may be extended to the case where the parameter z is complex. For the power series

$$P(z) = a_0 + a_1 z + a_2 z^2 + \cdots ,$$

we find that $P(z)$ converges if

$$\rho = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} z \right| = |z| \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| < 1.$$

We therefore have a range in $|z|$ for which $P(z)$ converges, i.e. $P(z)$ converges for values of z lying within a circle in the Argand diagram (in this case centred on the origin of the Argand diagram). The radius of the circle is called the *radius of convergence*: if z lies inside the circle, the series will converge whereas if z lies outside the circle, the series will diverge; if, though, z lies on the circle then the convergence must be tested using another method. Clearly the radius of convergence R is given by $1/R = \lim_{n\to\infty} |a_{n+1}/a_n|$.

►*Determine the range of values of z for which the following complex power series converges:*

$$P(z) = 1 - \frac{z}{2} + \frac{z^2}{4} - \frac{z^3}{8} + \cdots .$$

We find that $\rho = |z/2|$, which shows that $P(z)$ converges for $|z| < 2$. Therefore the circle of convergence in the Argand diagram is centred on the origin and has a radius $R = 2$. On this circle we must test the convergence by substituting the value of z into $P(z)$ and considering the resulting series. On the circle of convergence we can write $z = 2 \exp i\theta$. Substituting this into $P(z)$, we obtain

$$\begin{aligned} P(z) &= 1 - \frac{2 \exp i\theta}{2} + \frac{4 \exp 2i\theta}{4} - \cdots \\ &= 1 - \exp i\theta + [\exp i\theta]^2 - \cdots , \end{aligned}$$

which is a complex infinite geometric series with first term $a = 1$ and common ratio

$r = -\exp i\theta$. Therefore, on the the circle of convergence we have

$$P(z) = \frac{1}{1 + \exp i\theta}.$$

Unless $\theta = \pi$ this is a finite complex number, and so $P(z)$ converges at all points on the circle $|z| = 2$ except at $\theta = \pi$ (i.e. $z = -2$), where it diverges. Note that $P(z)$ is just the binomial expansion of $(1 + z/2)^{-1}$, for which it is obvious that $z = -2$ is a singular point. In general, for power series expansions of complex functions about a given point in the complex plane, the circle of convergence extends as far as the nearest singular point. This is discussed further in chapter 24. ◀

Note that the centre of the circle of convergence does not necessarily lie at the origin. For example, applying the ratio test to the complex power series

$$P(z) = 1 + \frac{z-1}{2} + \frac{(z-1)^2}{4} + \frac{(z-1)^3}{8} + \cdots,$$

we find that for it to converge we require $|(z-1)/2| < 1$. Thus the series converges for z lying within a circle of radius 2 centred on the point $(1, 0)$ in the Argand diagram.

4.5.2 Operations with power series

The following rules are useful when manipulating power series; they apply to power series in a real or complex variable.

(i) If two power series $P(x)$ and $Q(x)$ have regions of convergence that overlap to some extent then the series produced by taking the sum, the difference or the product of $P(x)$ and $Q(x)$ converges in the common region.

(ii) If two power series $P(x)$ and $Q(x)$ converge for all values of x then one series may be substituted into the other to give a third series, which also converges for all values of x. For example, consider the power series expansions of $\sin x$ and e^x given below in subsection 4.6.3,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$
$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots,$$

both of which converge for all values of x. Substituting the series for $\sin x$ into that for e^x we obtain

$$e^{\sin x} = 1 + x + \frac{x^2}{2!} - \frac{3x^4}{4!} - \frac{8x^5}{5!} + \cdots,$$

which also converges for all values of x.

If, however, either of the power series $P(x)$ and $Q(x)$ has only a limited region of convergence, or if they both do so, then further care must be taken when substituting one series into the other. For example, suppose $Q(x)$ converges for all x, but $P(x)$ only converges for x within a finite range. We may substitute

$Q(x)$ into $P(x)$ to obtain $P(Q(x))$, but we must be careful since the value of $Q(x)$ may lie outside the region of convergence for $P(x)$, with the consequence that the resulting series $P(Q(x))$ does not converge.

(iii) If a power series $P(x)$ converges for a particular range of x then the series obtained by differentiating every term and the series obtained by integrating every term also converge in this range.

This is easily seen for the power series

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots ,$$

which converges if $|x| < \lim_{n\to\infty} |a_n/a_{n+1}| \equiv k$. The series obtained by differentiating $P(x)$ with respect to x is given by

$$\frac{dP}{dx} = a_1 + 2a_2 x + 3a_3 x^2 + \cdots$$

and converges if

$$|x| < \lim_{n\to\infty} \left| \frac{na_n}{(n+1)a_{n+1}} \right| = k.$$

Similarly the series obtained by integrating $P(x)$ term by term,

$$\int P(x)\, dx = a_0 x + \frac{a_1 x^2}{2} + \frac{a_2 x^3}{3} + \cdots ,$$

converges if

$$|x| < \lim_{n\to\infty} \left| \frac{(n+2)a_n}{(n+1)a_{n+1}} \right| = k.$$

So, series resulting from differentiation or integration have the same interval of convergence as the original series. However, even if the original series converges at either end-point of the interval, it is not necessarily the case that the new series will do so. The new series must be tested separately at the end-points in order to determine whether it converges there. Note that although power series may be integrated or differentiated without altering their interval of convergence, this is not true for series in general.

It is also worth noting that differentiating or integrating a power series term by term within its interval of convergence is equivalent to differentiating or integrating the function it represents. For example, consider the power series expansion of $\sin x$,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots , \tag{4.14}$$

which converges for all values of x. If we differentiate term by term, the series becomes

$$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots ,$$

which is the series expansion of $\cos x$, as we expect.

4.6 Taylor series

Taylor's theorem provides a way of expressing a function as a power series in x, known as a *Taylor series*, but it can be applied only to those functions that are continuous and differentiable within the x-range of interest.

4.6.1 Taylor's theorem

Suppose that we have a function $f(x)$ that we wish to express as a power series in $x - a$ about the point $x = a$. We shall assume that, in a given x-range, $f(x)$ is a continuous, single-valued function of x having continuous derivatives with respect to x, denoted by $f'(x)$, $f''(x)$ and so on, up to and including $f^{(n-1)}(x)$. We shall also assume that $f^{(n)}(x)$ exists in this range.

From the equation following (2.31) we may write

$$\int_a^{a+h} f'(x)\,dx = f(a+h) - f(a),$$

where a, $a + h$ are neighbouring values of x. Rearranging this equation, we may express the value of the function at $x = a + h$ in terms of its value at a by

$$f(a+h) = f(a) + \int_a^{a+h} f'(x)\,dx. \tag{4.15}$$

A *first approximation* for $f(a+h)$ may be obtained by substituting $f'(a)$ for $f'(x)$ in (4.15), to obtain

$$f(a+h) \approx f(a) + hf'(a).$$

This approximation is shown graphically in figure 4.1. We may write this first approximation in terms of x and a as

$$f(x) \approx f(a) + (x-a)f'(a),$$

and, in a similar way,

$$f'(x) \approx f'(a) + (x-a)f''(a),$$
$$f''(x) \approx f''(a) + (x-a)f'''(a),$$

and so on. Substituting for $f'(x)$ in (4.15), we obtain the *second approximation*:

$$\begin{aligned} f(a+h) &\approx f(a) + \int_a^{a+h} [f'(a) + (x-a)f''(a)]\ dx \\ &\approx f(a) + hf'(a) + \frac{h^2}{2}f''(a). \end{aligned}$$

We may repeat this procedure as often as we like (so long as the derivatives of $f(x)$ exist) to obtain higher-order approximations to $f(a+h)$; we find the

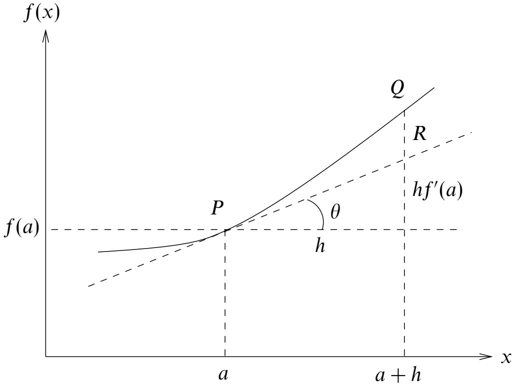


Figure 4.1 The first-order Taylor series approximation to a function $f(x)$. The slope of the function at P, i.e. $\tan\theta$, equals $f'(a)$. Thus the value of the function at Q, $f(a+h)$, is approximated by the ordinate of R, $f(a)+hf'(a)$.

$(n-1)$th-order approximation[§] to be

$$f(a+h) \approx f(a) + hf'(a) + \frac{h^2}{2!}f''(a) + \cdots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(a). \qquad (4.16)$$

As might have been anticipated, the error associated with approximating $f(a+h)$ by this $(n-1)$th-order power series is of the order of the next term in the series. This error or *remainder* can be shown to be given by

$$R_n(h) = \frac{h^n}{n!}f^{(n)}(\xi),$$

for some ξ that lies in the range $[a, a+h]$. Taylor's theorem then states that we may write the *equality*

$$f(a+h) = f(a) + hf'(a) + \frac{h^2}{2!}f''(a) + \cdots + \frac{h^{(n-1)}}{(n-1)!}f^{(n-1)}(a) + R_n(h). \qquad (4.17)$$

The theorem may also be written in a form suitable for finding $f(x)$ given the value of the function and its relevant derivatives at $x = a$, by substituting

[§] The order of the approximation is simply the highest power of h in the series. Note, though, that the $(n-1)$th-order approximation contains n terms.

$x = a + h$ in the above expression. It then reads

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^{n-1}}{(n-1)!}f^{(n-1)}(a) + R_n(x), \tag{4.18}$$

where the remainder now takes the form

$$R_n(x) = \frac{(x-a)^n}{n!}f^{(n)}(\xi),$$

and ξ lies in the range $[a, x]$. Each of the formulae (4.17), (4.18) gives us the *Taylor expansion* of the function about the point $x = a$. A special case occurs when $a = 0$. Such Taylor expansions, about $x = 0$, are called *Maclaurin series*.

Taylor's theorem is also valid without significant modification for functions of a complex variable (see chapter 24). The extension of Taylor's theorem to functions of more than one variable is given in chapter 5.

For a function to be expressible as an infinite power series we require it to be infinitely differentiable and the remainder term R_n to tend to zero as n tends to infinity, i.e. $\lim_{n\to\infty} R_n = 0$. In this case the infinite power series will represent the function within the interval of convergence of the series.

►*Expand* $f(x) = \sin x$ *as a Maclaurin series, i.e. about* $x = 0$.

We must first verify that $\sin x$ may indeed be represented by an infinite power series. It is easily shown that the nth derivative of $f(x)$ is given by

$$f^{(n)}(x) = \sin\left(x + \frac{n\pi}{2}\right).$$

Therefore the remainder after expanding $f(x)$ as an $(n-1)$th-order polynomial about $x = 0$ is given by

$$R_n(x) = \frac{x^n}{n!}\sin\left(\xi + \frac{n\pi}{2}\right),$$

where ξ lies in the range $[0, x]$. Since the modulus of the sine term is always less than or equal to unity, we can write $|R_n(x)| < |x^n|/n!$. For any particular value of x, say $x = c$, $R_n(c) \to 0$ as $n \to \infty$. Hence $\lim_{n\to\infty} R_n(x) = 0$, and so $\sin x$ can be represented by an infinite Maclaurin series.

Evaluating the function and its derivatives at $x = 0$ we obtain

$$\begin{aligned} f(0) &= \sin 0 = 0, \\ f'(0) &= \sin(\pi/2) = 1, \\ f''(0) &= \sin \pi = 0, \\ f'''(0) &= \sin(3\pi/2) = -1, \end{aligned}$$

and so on. Therefore, the Maclaurin series expansion of $\sin x$ is given by

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots .$$

Note that, as expected, since $\sin x$ is an odd function, its power series expansion contains only odd powers of x. ◄

We may follow a similar procedure to obtain a Taylor series about an arbitrary point $x = a$.

►*Expand* $f(x) = \cos x$ *as a Taylor series about* $x = \pi/3$.

As in the above example, it is easily shown that the nth derivative of $f(x)$ is given by

$$f^{(n)}(x) = \cos\left(x + \frac{n\pi}{2}\right).$$

Therefore the remainder after expanding $f(x)$ as an $(n-1)$th-order polynomial about $x = \pi/3$ is given by

$$R_n(x) = \frac{(x-\pi/3)^n}{n!}\cos\left(\xi + \frac{n\pi}{2}\right),$$

where ξ lies in the range $[\pi/3, x]$. The modulus of the cosine term is always less than or equal to unity, and so $|R_n(x)| < |(x-\pi/3)^n|/n!$. As in the previous example, $\lim_{n\to\infty} R_n(x) = 0$ for any particular value of x, and so $\cos x$ can be represented by an infinite Taylor series about $x = \pi/3$.

Evaluating the function and its derivatives at $x = \pi/3$ we obtain

$$\begin{aligned} f(\pi/3) &= \cos(\pi/3) = 1/2, \\ f'(\pi/3) &= \cos(5\pi/6) = -\sqrt{3}/2, \\ f''(\pi/3) &= \cos(4\pi/3) = -1/2, \end{aligned}$$

and so on. Thus the Taylor series expansion of $\cos x$ about $x = \pi/3$ is given by

$$\cos x = \frac{1}{2} - \frac{\sqrt{3}}{2}\left(x - \pi/3\right) - \frac{1}{2}\frac{\left(x-\pi/3\right)^2}{2!} + \cdots. ◄$$

4.6.2 Approximation errors in Taylor series

In the previous subsection we saw how to represent a function $f(x)$ by an infinite power series, which is exactly equal to $f(x)$ for all x within the interval of convergence of the series. However, in physical problems we usually do not want to have to sum an infinite number of terms, but prefer to use only a finite number of terms in the Taylor series to *approximate* the function in some given range of x. In this case it is desirable to know what is the maximum possible error associated with the approximation.

As given in (4.18), a function $f(x)$ can be represented by a finite $(n-1)$th-order power series together with a remainder term such that

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^{n-1}}{(n-1)!}f^{(n-1)}(a) + R_n(x),$$

where

$$R_n(x) = \frac{(x-a)^n}{n!}f^{(n)}(\xi)$$

and ξ lies in the range $[a, x]$. $R_n(x)$ is the remainder term, and represents the error in approximating $f(x)$ by the above $(n-1)$th-order power series. Since the exact value of ξ that satisfies the expression for $R_n(x)$ is not known, an upper limit on the error may be found by differentiating $R_n(x)$ with respect to ξ and equating the derivative to zero in the usual way for finding maxima.

►*Expand* $f(x) = \cos x$ *as a Taylor series about* $x = 0$ *and find the error associated with using the approximation to evaluate* $\cos(0.5)$ *if only the first two non-vanishing terms are taken. (Note that the Taylor expansions of trigonometric functions are only valid for angles measured in radians.)*

Evaluating the function and its derivatives at $x = 0$, we find

$$\begin{aligned} f(0) &= \cos 0 = 1, \\ f'(0) &= -\sin 0 = 0, \\ f''(0) &= -\cos 0 = -1, \\ f'''(0) &= \sin 0 = 0. \end{aligned}$$

So, for small $|x|$, we find from (4.18)

$$\cos x \approx 1 - \frac{x^2}{2}.$$

Note that since $\cos x$ is an even function, its power series expansion contains only even powers of x. Therefore, in order to estimate the error in this approximation, we must consider the term in x^4, which is the next in the series. The required derivative is $f^{(4)}(x)$ and this is (by chance) equal to $\cos x$. Thus, adding in the remainder term $R_4(x)$, we find

$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4!}\cos\xi,$$

where ξ lies in the range $[0, x]$. Thus, the maximum possible error is $x^4/4!$, since $\cos\xi$ cannot exceed unity. If $x = 0.5$, taking just the first two terms yields $\cos(0.5) \approx 0.875$ with a predicted error of less than 0.002 60. In fact $\cos(0.5) = 0.877\,58$ to 5 decimal places. Thus, to this accuracy, the true error is 0.002 58, an error of about 0.3%. ◄

4.6.3 Standard Maclaurin series

It is often useful to have a readily available table of Maclaurin series for standard elementary functions, and therefore these are listed below.

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \quad \text{for } -\infty < x < \infty, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \quad \text{for } -\infty < x < \infty, \\ \tan^{-1} x &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \quad \text{for } -1 < x < 1, \\ e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \quad \text{for } -\infty < x < \infty, \\ \ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \quad \text{for } -1 < x \le 1, \\ (1+x)^n &= 1 + nx + n(n-1)\frac{x^2}{2!} + n(n-1)(n-2)\frac{x^3}{3!} + \cdots \quad \text{for } -\infty < x < \infty. \end{aligned}$$

These can all be derived by straightforward application of Taylor's theorem to the expansion of a function about $x = 0$.

4.7 Evaluation of limits

The idea of the limit of a function $f(x)$ as x approaches a value a is fairly intuitive, though a strict definition exists and is stated below. In many cases the limit of the function as x approaches a will be simply the value $f(a)$, but sometimes this is not so. Firstly, the function may be undefined at $x = a$, as, for example, when

$$f(x) = \frac{\sin x}{x},$$

which takes the value $0/0$ at $x = 0$. However, the limit as x approaches zero does exist and can be evaluated as unity using l'Hôpital's rule below. Another possibility is that even if $f(x)$ is defined at $x = a$ its value may not be equal to the limiting value $\lim_{x\to a} f(x)$. This can occur for a discontinuous function at a point of discontinuity. The strict definition of a limit is that *if* $\lim_{x\to a} f(x) = l$ *then for any number* ϵ *however small, it must be possible to find a number* η *such that* $|f(x)-l| < \epsilon$ *whenever* $|x-a| < \eta$. In other words, as x becomes arbitrarily close to a, $f(x)$ becomes arbitrarily close to its limit, l. To remove any ambiguity, it should be stated that, in general, the number η will depend on both ϵ and the form of $f(x)$.

The following observations are often useful in finding the limit of a function.

(i) A limit may be $\pm\infty$. For example as $x \to 0$, $1/x^2 \to \infty$.

(ii) A limit may be approached from below or above and the value may be different in each case. For example consider the function $f(x) = \tan x$. As x tends to $\pi/2$ from below $f(x) \to \infty$, but if the limit is approached from above then $f(x) \to -\infty$. Another way of writing this is

$$\lim_{x\to\frac{\pi}{2}^-} \tan x = \infty, \qquad \lim_{x\to\frac{\pi}{2}^+} \tan x = -\infty.$$

(iii) It may ease the evaluation of limits if the function under consideration is split into a sum, product or quotient. Provided that in each case a limit exists, the rules for evaluating such limits are as follows.

(a) $\lim_{x\to a} \{f(x) + g(x)\} = \lim_{x\to a} f(x) + \lim_{x\to a} g(x)$.

(b) $\lim_{x\to a} \{f(x)g(x)\} = \lim_{x\to a} f(x) \lim_{x\to a} g(x)$.

(c) $\lim_{x\to a} \dfrac{f(x)}{g(x)} = \dfrac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$, provided that the numerator and denominator are not both equal to zero or infinity.

Examples of cases (a)–(c) are discussed below.

►*Evaluate the limits*

$$\lim_{x\to 1}(x^2+2x^3), \qquad \lim_{x\to 0}(x\cos x), \qquad \lim_{x\to\pi/2}\frac{\sin x}{x}.$$

Using (a) above,

$$\lim_{x\to 1}(x^2+2x^3)=\lim_{x\to 1}x^2+\lim_{x\to 1}2x^3=3.$$

Using (b),

$$\lim_{x\to 0}(x\cos x)=\lim_{x\to 0}x\lim_{x\to 0}\cos x=0\times 1=0.$$

Using (c),

$$\lim_{x\to\pi/2}\frac{\sin x}{x}=\frac{\lim_{x\to\pi/2}\sin x}{\lim_{x\to\pi/2}x}=\frac{1}{\pi/2}=\frac{2}{\pi}. \blacktriangleleft$$

(iv) Limits of functions of x that contain exponents that themselves depend on x can often be found by taking logarithms.

►*Evaluate the limit*

$$\lim_{x\to\infty}\left(1-\frac{a^2}{x^2}\right)^{x^2}.$$

Let us define

$$y=\left(1-\frac{a^2}{x^2}\right)^{x^2}$$

and consider the logarithm of the required limit, i.e.

$$\lim_{x\to\infty}\ln y=\lim_{x\to\infty}\left[x^2\ln\left(1-\frac{a^2}{x^2}\right)\right].$$

Using the Maclaurin series for $\ln(1+x)$ given in subsection 4.6.3, we can expand the logarithm as a series and obtain

$$\lim_{x\to\infty}\ln y=\lim_{x\to\infty}\left[x^2\left(-\frac{a^2}{x^2}-\frac{a^4}{2x^4}+\cdots\right)\right]=-a^2.$$

Therefore, since $\lim_{x\to\infty}\ln y=-a^2$ it follows that $\lim_{x\to\infty}y=\exp(-a^2)$. ◄

(v) L'Hôpital's rule may be used; it is an extension of (iii)(c) above. In cases where both numerator and denominator are zero or both are infinite, further consideration of the limit must follow. Let us first consider $\lim_{x\to a}f(x)/g(x)$, where $f(a)=g(a)=0$. Expanding the numerator and denominator as Taylor series we obtain

$$\frac{f(x)}{g(x)}=\frac{f(a)+(x-a)f'(a)+[(x-a)^2/2!]f''(a)+\cdots}{g(a)+(x-a)g'(a)+[(x-a)^2/2!]g''(a)+\cdots}.$$

However, $f(a)=g(a)=0$ so

$$\frac{f(x)}{g(x)}=\frac{f'(a)+[(x-a)/2!]f''(a)+\cdots}{g'(a)+[(x-a)/2!]g''(a)+\cdots}.$$

Therefore we find

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\frac{f'(a)}{g'(a)},$$

provided $f'(a)$ and $g'(a)$ are not themselves both equal to zero. If, however, $f'(a)$ and $g'(a)$ *are* both zero then the same process can be applied to the ratio $f'(x)/g'(x)$ to yield

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\frac{f''(a)}{g''(a)},$$

provided that at least one of $f''(a)$ and $g''(a)$ is non-zero. If the original limit does exist then it can be found by repeating the process as many times as is necessary for the ratio of corresponding nth derivatives not to be of the indeterminate form $0/0$, i.e.

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\frac{f^{(n)}(a)}{g^{(n)}(a)}.$$

►*Evaluate the limit*

$$\lim_{x\to 0}\frac{\sin x}{x}.$$

We first note that if $x=0$, both numerator and denominator are zero. Thus we apply l'Hôpital's rule: differentiating, we obtain

$$\lim_{x\to 0}(\sin x/x)=\lim_{x\to 0}(\cos x/1)=1. ◄$$

So far we have only considered the case where $f(a)=g(a)=0$. For the case where $f(a)=g(a)=\infty$ we may still apply l'Hôpital's rule by writing

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{1/g(x)}{1/f(x)},$$

which is now of the form $0/0$ at $x=a$. Note also that l'Hôpital's rule is still valid for finding limits as $x\to\infty$, i.e. when $a=\infty$. This is easily shown by letting $y=1/x$ as follows:

$$\begin{aligned}\lim_{x\to\infty}\frac{f(x)}{g(x)}&=\lim_{y\to 0}\frac{f(1/y)}{g(1/y)}\\&=\lim_{y\to 0}\frac{-f'(1/y)/y^2}{-g'(1/y)/y^2}\\&=\lim_{y\to 0}\frac{f'(1/y)}{g'(1/y)}\\&=\lim_{x\to\infty}\frac{f'(x)}{g'(x)}.\end{aligned}$$

Summary of methods for evaluating limits

To find the limit of a continuous function $f(x)$ at a point $x = a$, simply substitute the value a into the function noting that $\frac{0}{\infty} = 0$ and that $\frac{\infty}{0} = \infty$. The only difficulty occurs when either of the expressions $\frac{0}{0}$ or $\frac{\infty}{\infty}$ results. In this case differentiate top and bottom and try again. Continue differentiating until the top and bottom limits are no longer both zero or both infinity. If the undetermined form $0 \times \infty$ occurs then it can always be rewritten as $\frac{0}{0}$ or $\frac{\infty}{\infty}$.

4.8 Exercises

4.1 Sum the even numbers between 1000 and 2000 inclusive.

4.2 If you invest £1000 on the first day of each year, and interest is paid at 5% on your balance at the end of each year, how much money do you have after 25 years?

4.3 How does the convergence of the series

$$\sum_{n=r}^{\infty} \frac{(n-r)!}{n!}$$

depend on the integer r?

4.4 Show that for testing the convergence of the series

$$x + y + x^2 + y^2 + x^3 + y^3 + \cdots,$$

where $0 < x < y < 1$, the D'Alembert ratio test fails but the Cauchy root test is successful.

4.5 Find the sum S_N of the first N terms of the following series, and hence determine whether the series are convergent, divergent or oscillatory:

$$\text{(a)} \sum_{n=1}^{\infty} \ln\left(\frac{n+1}{n}\right), \qquad \text{(b)} \sum_{n=0}^{\infty} (-2)^n, \qquad \text{(c)} \sum_{n=1}^{\infty} \frac{(-1)^{n+1} n}{3^n}.$$

4.6 By grouping and rearranging terms of the absolutely convergent series

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2},$$

show that

$$S_{\text{o}} = \sum_{n \text{ odd}}^{\infty} \frac{1}{n^2} = \frac{3S}{4}.$$

4.7 Use the difference method to sum the series

$$\sum_{n=2}^{N} \frac{2n-1}{2n^2(n-1)^2}.$$

4.8 The $N+1$ complex numbers ω_m are given by $\omega_m = \exp(2\pi i m/N)$, for $m = 0, 1, 2, \ldots, N$.

(a) Evaluate the following:

$$\text{(i)} \sum_{m=0}^{N} \omega_m, \quad \text{(ii)} \sum_{m=0}^{N} \omega_m^2, \quad \text{(iii)} \sum_{m=0}^{N} \omega_m x^m.$$

(b) Use these results to evaluate:

$$\text{(i)} \sum_{m=0}^{N} \left[\cos\left(\frac{2\pi m}{N}\right) - \cos\left(\frac{4\pi m}{N}\right)\right], \quad \text{(ii)} \sum_{m=0}^{3} 2^m \sin\left(\frac{2\pi m}{3}\right).$$

4.9 Prove that

$$\cos\theta + \cos(\theta+\alpha) + \cdots + \cos(\theta + n\alpha) = \frac{\sin\frac{1}{2}(n+1)\alpha}{\sin\frac{1}{2}\alpha}\cos(\theta + \tfrac{1}{2}n\alpha).$$

4.10 Determine whether the following series converge (θ and p are positive real numbers):

$$\text{(a)} \sum_{n=1}^{\infty} \frac{2\sin n\theta}{n(n+1)}, \quad \text{(b)} \sum_{n=1}^{\infty} \frac{2}{n^2}, \quad \text{(c)} \sum_{n=1}^{\infty} \frac{1}{2n^{1/2}},$$

$$\text{(d)} \sum_{n=2}^{\infty} \frac{(-1)^n (n^2+1)^{1/2}}{n \ln n}, \quad \text{(e)} \sum_{n=1}^{\infty} \frac{n^p}{n!}.$$

4.11 Find the real values of x for which the following series are convergent:

$$\text{(a)} \sum_{n=1}^{\infty} \frac{x^n}{n+1}, \quad \text{(b)} \sum_{n=1}^{\infty} (\sin x)^n, \quad \text{(c)} \sum_{n=1}^{\infty} n^x,$$

$$\text{(d)} \sum_{n=1}^{\infty} e^{nx}, \quad \text{(e)} \sum_{n=2}^{\infty} (\ln n)^x.$$

4.12 Determine whether the following series are convergent:

$$\text{(a)} \sum_{n=1}^{\infty} \frac{n^{1/2}}{(n+1)^{1/2}}, \quad \text{(b)} \sum_{n=1}^{\infty} \frac{n^2}{n!}, \quad \text{(c)} \sum_{n=1}^{\infty} \frac{(\ln n)^n}{n^{n/2}}, \quad \text{(d)} \sum_{n=1}^{\infty} \frac{n^n}{n!}.$$

4.13 Determine whether the following series are absolutely convergent, convergent or oscillatory:

$$\text{(a)} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^{5/2}}, \quad \text{(b)} \sum_{n=1}^{\infty} \frac{(-1)^n(2n+1)}{n}, \quad \text{(c)} \sum_{n=0}^{\infty} \frac{(-1)^n |x|^n}{n!},$$

$$\text{(d)} \sum_{n=0}^{\infty} \frac{(-1)^n}{n^2+3n+2}, \quad \text{(e)} \sum_{n=1}^{\infty} \frac{(-1)^n 2^n}{n^{1/2}}.$$

4.14 Obtain the positive values of x for which the following series converges:

$$\sum_{n=1}^{\infty} \frac{x^{n/2} e^{-n}}{n}.$$

4.15 Prove that

$$\sum_{n=2}^{\infty} \ln\left[\frac{n^r + (-1)^n}{n^r}\right]$$

is absolutely convergent for $r = 2$, but only conditionally convergent for $r = 1$.

4.16 An extension to the proof of the integral test (subsection 4.3.2) shows that, if $f(x)$ is positive, continuous and monotonically decreasing, for $x \geq 1$, and the series $f(1) + f(2) + \cdots$ is convergent, then its sum does not exceed $f(1) + L$, where L is the integral

$$\int_1^{\infty} f(x)\, dx.$$

Use this result to show that the sum $\zeta(p)$ of the Riemann zeta series $\sum n^{-p}$, with $p > 1$, is not greater than $p/(p-1)$.

4.17 Demonstrate that rearranging the order of its terms can make a conditionally convergent series converge to a different limit by considering the series $\sum(-1)^{n+1}n^{-1} = \ln 2 = 0.693$. Rearrange the series as

$$S = \tfrac{1}{1} + \tfrac{1}{3} - \tfrac{1}{2} + \tfrac{1}{5} + \tfrac{1}{7} - \tfrac{1}{4} + \tfrac{1}{9} + \tfrac{1}{11} - \tfrac{1}{6} + \tfrac{1}{13} + \cdots$$

and group each set of three successive terms. Show that the series can then be written

$$\sum_{m=1}^{\infty} \frac{8m-3}{2m(4m-3)(4m-1)},$$

which is convergent (by comparison with $\sum n^{-2}$) and contains only positive terms. Evaluate the first of these and hence deduce that S is not equal to $\ln 2$.

4.18 Illustrate result (iv) of section 4.4, concerning Cauchy products, by considering the double summation

$$S = \sum_{n=1}^{\infty} \sum_{r=1}^{n} \frac{1}{r^2(n+1-r)^3}.$$

By examining the points in the nr-plane over which the double summation is to be carried out, show that S can be written as

$$S = \sum_{n=r}^{\infty} \sum_{r=1}^{\infty} \frac{1}{r^2(n+1-r)^3}.$$

Deduce that $S \leq 3$.

4.19 A Fabry–Pérot interferometer consists of two parallel heavily silvered glass plates; light enters normally to the plates, and undergoes repeated reflections between them, with a small transmitted fraction emerging at each reflection. Find the intensity of the emerging wave, $|B|^2$, where

$$B = A(1-r)\sum_{n=0}^{\infty} r^n e^{in\phi},$$

with r and ϕ real.

4.20 Identify the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^{2n}}{(2n-1)!},$$

and then, by integration and differentiation, deduce the values S of the following series:

$$\text{(a)} \sum_{n=1}^{\infty} \frac{(-1)^{n+1} n^2}{(2n)!}, \qquad \text{(b)} \sum_{n=1}^{\infty} \frac{(-1)^{n+1} n}{(2n+1)!},$$
$$\text{(c)} \sum_{n=1}^{\infty} \frac{(-1)^{n+1} n \pi^{2n}}{4^n (2n-1)!}, \qquad \text{(d)} \sum_{n=0}^{\infty} \frac{(-1)^n (n+1)}{(2n)!}.$$

4.21 Starting from the Maclaurin series for $\cos x$, show that

$$(\cos x)^{-2} = 1 + x^2 + \frac{2x^4}{3} + \cdots .$$

Deduce the first three terms in the Maclaurin series for $\tan x$.

4.22 Find the Maclaurin series for:

$$\text{(a)} \;\; \ln\left(\frac{1+x}{1-x}\right), \qquad \text{(b)} \;\; (x^2+4)^{-1}, \qquad \text{(c)} \;\; \sin^2 x.$$

4.23 Writing the nth derivative of $f(x) = \sinh^{-1} x$ as

$$f^{(n)}(x) = \frac{P_n(x)}{(1+x^2)^{n-1/2}},$$

where $P_n(x)$ is a polynomial (of order $n-1$), show that the $P_n(x)$ satisfy the recurrence relation

$$P_{n+1}(x) = (1+x^2)P_n'(x) - (2n-1)xP_n(x).$$

Hence generate the coefficients necessary to express $\sinh^{-1} x$ as a Maclaurin series up to terms in x^5.

4.24 Find the first three non-zero terms in the Maclaurin series for the following functions:

$$\begin{array}{lll} \text{(a) } (x^2+9)^{-1/2}, & \text{(b) } \ln[(2+x)^3], & \text{(c) } \exp(\sin x), \\ \text{(d) } \ln(\cos x), & \text{(e) } \exp[-(x-a)^{-2}], & \text{(f) } \tan^{-1} x. \end{array}$$

4.25 By using the logarithmic series, prove that if a and b are positive and nearly equal then

$$\ln \frac{a}{b} \simeq \frac{2(a-b)}{a+b}.$$

Show that the error in this approximation is about $2(a-b)^3/[3(a+b)^3]$.

4.26 Determine whether the following functions $f(x)$ are (i) continuous, and (ii) differentiable at $x = 0$:

(a) $f(x) = \exp(-|x|)$;
(b) $f(x) = (1-\cos x)/x^2$ for $x \neq 0$, $f(0) = \frac{1}{2}$;
(c) $f(x) = x \sin(1/x)$ for $x \neq 0$, $f(0) = 0$;
(d) $f(x) = [4 - x^2]$, where $[y]$ denotes the integer part of y.

4.27 Find the limit as $x \to 0$ of $[\sqrt{1+x^m} - \sqrt{1-x^m}]/x^n$, in which m and n are positive integers.

4.28 Evaluate the following limits:

$$\text{(a)} \lim_{x\to 0} \frac{\sin 3x}{\sinh x}, \qquad \text{(b)} \lim_{x\to 0} \frac{\tan x - \tanh x}{\sinh x - x},$$

$$\text{(c)} \lim_{x\to 0} \frac{\tan x - x}{\cos x - 1}, \qquad \text{(d)} \lim_{x\to 0} \left(\frac{\operatorname{cosec} x}{x^3} - \frac{\sinh x}{x^5}\right).$$

4.29 Find the limits of the following functions:

(a) $\dfrac{x^3 + x^2 - 5x - 2}{2x^3 - 7x^2 + 4x + 4}$, as $x \to 0$, $x \to \infty$ and $x \to 2$;

(b) $\dfrac{\sin x - x \cosh x}{\sinh x - x}$, as $x \to 0$;

(c) $\displaystyle\int_x^{\pi/2} \left(\frac{y \cos y - \sin y}{y^2}\right) dy$, as $x \to 0$.

4.30 Use Taylor expansions to three terms to find approximations to (a) $\sqrt[4]{17}$, and (b) $\sqrt[3]{26}$.

4.31 Using a first-order Taylor expansion about $x = x_0$, show that a better approximation than x_0 to the solution of the equation

$$f(x) = \sin x + \tan x = 2$$

is given by $x = x_0 + \delta$, where

$$\delta = \frac{2 - f(x_0)}{\cos x_0 + \sec^2 x_0}.$$

(a) Use this procedure twice to find the solution of $f(x) = 2$ to six significant figures, given that it is close to $x = 0.9$.

(b) Use the result in (a) to deduce, to the same degree of accuracy, one solution of the quartic equation

$$y^4 - 4y^3 + 4y^2 + 4y - 4 = 0.$$

4.32 Evaluate

$$\lim_{x\to 0} \left[\frac{1}{x^3}\left(\operatorname{cosec} x - \frac{1}{x} - \frac{x}{6}\right)\right].$$

4.33 In quantum theory, a system of oscillators, each of fundamental frequency ν and interacting at temperature T, has an average energy $\bar{E}$ given by

$$\bar{E} = \frac{\sum_{n=0}^{\infty} nh\nu e^{-nx}}{\sum_{n=0}^{\infty} e^{-nx}},$$

where $x = h\nu/kT$, h and k being the Planck and Boltzmann constants, respectively. Prove that both series converge, evaluate their sums, and show that at high temperatures $\bar{E} \approx kT$, whilst at low temperatures $\bar{E} \approx h\nu \exp(-h\nu/kT)$.

4.34 In a very simple model of a crystal, point-like atomic ions are regularly spaced along an infinite one-dimensional row with spacing R. Alternate ions carry equal and opposite charges $\pm e$. The potential energy of the ith ion in the electric field due to another ion, the jth, is

$$\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}},$$

where q_i, q_j are the charges on the ions and r_{ij} is the distance between them.

Write down a series giving the total contribution V_i of the ith ion to the overall potential energy. Show that the series converges, and, if V_i is written as

$$V_i = \frac{\alpha e^2}{4\pi\epsilon_0 R},$$

find a closed-form expression for α, the Madelung constant for this (unrealistic) lattice.

4.35 One of the factors contributing to the high relative permittivity of water to static electric fields is the permanent electric dipole moment, p, of the water molecule. In an external field E the dipoles tend to line up with the field, but they do not do so completely because of thermal agitation corresponding to the temperature, T, of the water. A classical (non-quantum) calculation using the Boltzmann distribution shows that the average polarisability per molecule, α, is given by

$$\alpha = \frac{p}{E}(\coth x - x^{-1}),$$

where $x = pE/(kT)$ and k is the Boltzmann constant.

At ordinary temperatures, even with high field strengths (10^4 V m^{-1} or more), $x \ll 1$. By making suitable series expansions of the hyperbolic functions involved, show that $\alpha = p^2/(3kT)$ to an accuracy of about one part in $15x^{-2}$.

4.36 In quantum theory, a certain method (the Born approximation) gives the (so-called) amplitude $f(\theta)$ for the scattering of a particle of mass m through an angle θ by a uniform potential well of depth V_0 and radius b (i.e. the potential energy of the particle is $-V_0$ within a sphere of radius b and zero elsewhere) as

$$f(\theta) = \frac{2mV_0}{\hbar^2 K^3}(\sin Kb - Kb \cos Kb).$$

Here $\hbar$ is the Planck constant divided by 2π, the energy of the particle is $\hbar^2 k^2/(2m)$ and K is $2k\sin(\theta/2)$.

Use l'Hôpital's rule to evaluate the amplitude at low energies, i.e. when k and hence K tend to zero, and so determine the low-energy total cross-section.

[Note: the differential cross-section is given by $|f(\theta)|^2$ and the total cross-section by the integral of this over all solid angles, i.e. $2\pi \int_0^\pi |f(\theta)|^2 \sin\theta \, d\theta$.]

4.9 Hints and answers

4.1 Write as $2(\sum_{n=1}^{1000} n - \sum_{n=1}^{499} n) = 751\,500$.

4.3 Divergent for $r \le 1$; convergent for $r \ge 2$.

4.5 (a) $\ln(N+1)$, divergent; (b) $\frac{1}{3}[1-(-2)^n]$, oscillates infinitely; (c) Add $\frac{1}{3}S_N$ to the S_N series; $\frac{3}{16}[1-(-3)^{-N}] + \frac{3}{4}N(-3)^{-N-1}$, convergent to $\frac{3}{16}$.

4.7 Write the nth term as the difference between two consecutive values of a partial-fraction function of n. The sum equals $\frac{1}{2}(1-N^{-2})$.

4.9 Sum the geometric series with rth term $\exp[i(\theta + r\alpha)]$. Its real part is

$$\{\cos\theta - \cos[(n+1)\alpha + \theta] - \cos(\theta - \alpha) + \cos(\theta + n\alpha)\}/4\sin^2(\alpha/2),$$

which can be reduced to the given answer.

4.11 (a) $-1 \le x < 1$; (b) all x except $x = (2n \pm 1)\pi/2$; (c) $x < -1$; (d) $x < 0$; (e) always divergent. Clearly divergent for $x > -1$. For $-X = x < -1$, consider

$$\sum_{k=1}^{\infty} \sum_{n=M_{k-1}+1}^{M_k} \frac{1}{(\ln M_k)^X},$$

where $\ln M_k = k$ and note that $M_k - M_{k-1} = e^{-1}(e-1)M_k$; hence show that the series diverges.

4.13 (a) Absolutely convergent, compare with exercise 4.10(b). (b) Oscillates finitely. (c) Absolutely convergent for all x. (d) Absolutely convergent; use partial fractions. (e) Oscillates infinitely.

4.15 Divide the series into two series, n odd and n even. For $r = 2$ both are absolutely convergent, by comparison with $\sum n^{-2}$. For $r = 1$ neither series is convergent, by comparison with $\sum n^{-1}$. However, the sum of the two is convergent, by the alternating sign test or by showing that the terms cancel in pairs.

4.17 The first term has value 0.833 and all other terms are positive.

4.19 $|A|^2(1-r)^2/(1+r^2-2r\cos\phi)$.

4.21 Use the binomial expansion and collect terms up to x^4. Integrate both sides of the displayed equation. $\tan x = x + x^3/3 + 2x^5/15 + \cdots$.

4.23 For example, $P_5(x) = 24x^4 - 72x^2 + 9$. $\sinh^{-1} x = x - x^3/6 + 3x^5/40 - \cdots$.

4.25 Set $a = D + \delta$ and $b = D - \delta$ and use the expansion for $\ln(1 \pm \delta/D)$.

4.27 The limit is 0 for $m > n$, 1 for $m = n$, and ∞ for $m < n$.

4.29 (a) $-\frac{1}{2}$, $\frac{1}{2}$, ∞; (b) -4; (c) $-1 + 2/\pi$.

4.31 (a) First approximation 0.886 452; second approximation 0.886 287. (b) Set $y = \sin x$ and re-express $f(x) = 2$ as a polynomial equation. $y = \sin(0.886\,287) = 0.774\,730$.

4.33 If $S(x) = \sum_{n=0}^{\infty} e^{-nx}$ evaluate $S(x)$ and consider $dS(x)/dx$.
$E = h\nu\,[\exp(h\nu/kT) - 1]^{-1}$.

4.35 The series expansion is $\dfrac{px}{E}\left(\dfrac{1}{3} - \dfrac{x^2}{45} + \cdots\right)$.

5

Partial differentiation

In chapter 2, we discussed functions f of only one variable x, which were usually written $f(x)$. Certain constants and parameters may also have appeared in the definition of f, e.g. $f(x) = ax + 2$ contains the constant 2 and the parameter a, but only x was considered as a variable and only the derivatives $f^{(n)}(x) = d^n f/dx^n$ were defined.

However, we may equally well consider functions that depend on more than one variable, e.g. the function $f(x, y) = x^2 + 3xy$, which depends on the two variables x and y. For any pair of values x, y, the function $f(x, y)$ has a well-defined value, e.g. $f(2, 3) = 22$. This notion can clearly be extended to functions dependent on more than two variables. For the n-variable case, we write $f(x_1, x_2, \ldots, x_n)$ for a function that depends on the variables $x_1, x_2, \ldots, x_n$. When $n = 2$, x_1 and x_2 correspond to the variables x and y used above.

Functions of one variable, like $f(x)$, can be represented by a graph on a plane sheet of paper, and it is apparent that functions of two variables can, with little effort, be represented by a surface in three-dimensional space. Thus, we may also picture $f(x, y)$ as describing the variation of height with position in a mountainous landscape. Functions of many variables, however, are usually very difficult to visualise and so the preliminary discussion in this chapter will concentrate on functions of just two variables.

5.1 Definition of the partial derivative

It is clear that a function $f(x, y)$ of two variables will have a gradient in all directions in the xy-plane. A general expression for this rate of change can be found and will be discussed in the next section. However, we first consider the simpler case of finding the rate of change of $f(x, y)$ in the positive x- and y-directions. These rates of change are called the *partial derivatives* with respect to x and y respectively, and they are extremely important in a wide range of physical applications.

For a function of two variables $f(x, y)$ we may define the derivative with respect to x, for example, by saying that it is that for a one-variable function when y is held fixed and treated as a constant. To signify that a derivative is with respect to x, but at the same time to recognize that a derivative with respect to y also exists, the former is denoted by $\partial f/\partial x$ and is the *partial derivative of $f(x, y)$ with respect to x*. Similarly, the partial derivative of f with respect to y is denoted by $\partial f/\partial y$.

To define formally the partial derivative of $f(x, y)$ with respect to x, we have

$$\frac{\partial f}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}, \tag{5.1}$$

provided that the limit exists. This is much the same as for the derivative of a one-variable function. The other partial derivative of $f(x, y)$ is similarly defined as a limit (provided it exists):

$$\frac{\partial f}{\partial y} = \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}. \tag{5.2}$$

It is common practice in connection with partial derivatives of functions involving more than one variable to indicate those variables that are held constant by writing them as subscripts to the derivative symbol. Thus, the partial derivatives defined in (5.1) and (5.2) would be written respectively as

$$\left(\frac{\partial f}{\partial x}\right)_y \qquad \text{and} \qquad \left(\frac{\partial f}{\partial y}\right)_x.$$

In this form, the subscript shows explicitly which variable is to be kept constant. A more compact notation for these partial derivatives is f_x and f_y. However, it is extremely important when using partial derivatives to remember which variables are being held constant and it is wise to write out the partial derivative in explicit form if there is any possibility of confusion.

The extension of the definitions (5.1), (5.2) to the general n-variable case is straightforward and can be written formally as

$$\frac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_i} = \lim_{\Delta x_i \to 0} \frac{[f(x_1, x_2, \ldots, x_i + \Delta x_i, \ldots, x_n) - f(x_1, x_2, \ldots, x_i, \ldots, x_n)]}{\Delta x_i},$$

provided that the limit exists.

Just as for one-variable functions, second (and higher) partial derivatives may be defined in a similar way. For a two-variable function $f(x, y)$ they are

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2} = f_{xx}, \qquad \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2} = f_{yy},$$

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x \partial y} = f_{xy}, \qquad \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y \partial x} = f_{yx}.$$

Only three of the second derivatives are independent since the relation

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x},$$

is always obeyed, provided that the second partial derivatives are continuous at the point in question. This relation often proves useful as a labour-saving device when evaluating second partial derivatives. It can also be shown that for a function of n variables, $f(x_1, x_2, \ldots, x_n)$, under the same conditions,

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

►*Find the first and second partial derivatives of the function*

$$f(x, y) = 2x^3y^2 + y^3.$$

The first partial derivatives are

$$\frac{\partial f}{\partial x} = 6x^2y^2, \qquad \frac{\partial f}{\partial y} = 4x^3y + 3y^2,$$

and the second partial derivatives are

$$\frac{\partial^2 f}{\partial x^2} = 12xy^2, \qquad \frac{\partial^2 f}{\partial y^2} = 4x^3 + 6y, \qquad \frac{\partial^2 f}{\partial x \partial y} = 12x^2y, \qquad \frac{\partial^2 f}{\partial y \partial x} = 12x^2y,$$

the last two being equal, as expected. ◄

5.2 The total differential and total derivative

Having defined the (first) partial derivatives of a function $f(x, y)$, which give the rate of change of f along the positive x- and y-axes, we consider next the rate of change of $f(x, y)$ in an arbitrary direction. Suppose that we make simultaneous small changes Δx in x and Δy in y and that, as a result, f changes to $f + \Delta f$. Then we must have

$$\begin{aligned}
\Delta f &= f(x + \Delta x, y + \Delta y) - f(x, y) \\
&= f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y) + f(x, y + \Delta y) - f(x, y) \\
&= \left[\frac{f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)}{\Delta x}\right] \Delta x + \left[\frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}\right] \Delta y.
\end{aligned} \tag{5.3}$$

In the last line we note that the quantities in brackets are very similar to those involved in the definitions of partial derivatives (5.1), (5.2). For them to be strictly equal to the partial derivatives, Δx and Δy would need to be infinitesimally small. But even for finite (but not too large) Δx and Δy the approximate formula

$$\Delta f \approx \frac{\partial f(x, y)}{\partial x} \Delta x + \frac{\partial f(x, y)}{\partial y} \Delta y, \tag{5.4}$$

can be obtained. It will be noticed that the first bracket in (5.3) actually approximates to $\partial f(x, y + \Delta y)/\partial x$ but that this has been replaced by $\partial f(x, y)/\partial x$ in (5.4). This approximation clearly has the same degree of validity as that which replaces the bracket by the partial derivative.

How valid an approximation (5.4) is to (5.3) depends not only on how small Δx and Δy are but also on the magnitudes of higher partial derivatives; this is discussed further in section 5.7 in the context of Taylor series for functions of more than one variable. Nevertheless, letting the small changes Δx and Δy in (5.4) become infinitesimal, we can define the *total differential* df of the function $f(x, y)$, without any approximation, as

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy. \tag{5.5}$$

Equation (5.5) can be extended to the case of a function of n variables, $f(x_1, x_2, \ldots, x_n)$;

$$df = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \cdots + \frac{\partial f}{\partial x_n}dx_n. \tag{5.6}$$

►*Find the total differential of the function* $f(x, y) = y\exp(x + y)$.

Evaluating the first partial derivatives, we find

$$\frac{\partial f}{\partial x} = y\exp(x + y), \quad \frac{\partial f}{\partial y} = \exp(x + y) + y\exp(x + y).$$

Applying (5.5), we then find that the total differential is given by

$$df = [y\exp(x + y)]dx + [(1 + y)\exp(x + y)]dy. \blacktriangleleft$$

In some situations, despite the fact that several variables x_i, $i = 1, 2, \ldots, n$, appear to be involved, effectively only one of them is. This occurs if there are subsidiary relationships constraining all the x_i to have values dependent on the value of one of them, say x_1. These relationships may be represented by equations that are typically of the form

$$x_i = x_i(x_1), \qquad i = 2, 3, \ldots, n. \tag{5.7}$$

In principle f can then be expressed as a function of x_1 alone by substituting from (5.7) for $x_2, x_3, \ldots, x_n$, and then the *total derivative* (or simply the derivative) of f with respect to x_1 is obtained by ordinary differentiation.

Alternatively, (5.6) can be used to give

$$\frac{df}{dx_1} = \frac{\partial f}{\partial x_1} + \left(\frac{\partial f}{\partial x_2}\right)\frac{dx_2}{dx_1} + \cdots + \left(\frac{\partial f}{\partial x_n}\right)\frac{dx_n}{dx_1}. \tag{5.8}$$

It should be noted that the LHS of this equation is the total derivative df/dx_1, whilst the partial derivative $\partial f/\partial x_1$ forms only a part of the RHS. In evaluating this partial derivative account must be taken only of *explicit* appearances of x_1 in the function f, and *no* allowance must be made for the knowledge that changing x_1 necessarily changes $x_2, x_3, \ldots, x_n$. The contribution from these latter changes is precisely that of the remaining terms on the RHS of (5.8). Naturally, what has been shown using x_1 in the above argument applies equally well to any other of the x_i, with the appropriate consequent changes.

►*Find the total derivative of $f(x,y) = x^2 + 3xy$ with respect to x, given that $y = \sin^{-1} x$.*

We can see immediately that

$$\frac{\partial f}{\partial x} = 2x + 3y, \qquad \frac{\partial f}{\partial y} = 3x, \qquad \frac{dy}{dx} = \frac{1}{(1-x^2)^{1/2}}$$

and so, using (5.8) with $x_1 = x$ and $x_2 = y$,

$$\begin{aligned}\frac{df}{dx} &= 2x + 3y + 3x\frac{1}{(1-x^2)^{1/2}}\\ &= 2x + 3\sin^{-1} x + \frac{3x}{(1-x^2)^{1/2}}.\end{aligned}$$

Obviously the same expression would have resulted if we had substituted for y from the start, but the above method often produces results with reduced calculation, particularly in more complicated examples. ◄

5.3 Exact and inexact differentials

In the last section we discussed how to find the total differential of a function, i.e. its infinitesimal change in an arbitrary direction, in terms of its gradients $\partial f/\partial x$ and $\partial f/\partial y$ in the x- and y- directions (see (5.5)). Sometimes, however, we wish to reverse the process and find the function f that differentiates to give a known differential. Usually, finding such functions relies on inspection and experience.

As an example, it is easy to see that the function whose differential is $df = x\,dy + y\,dx$ is simply $f(x,y) = xy + c$, where c is a constant. Differentials such as this, which integrate directly, are called *exact differentials*, whereas those that do not are *inexact differentials*. For example, $x\,dy + 3y\,dx$ is not the straightforward differential of any function (see below). Inexact differentials can be made exact, however, by multiplying through by a suitable function called an integrating factor. This is discussed further in subsection 14.2.3.

►*Show that the differential $x\,dy + 3y\,dx$ is inexact.*

On the one hand, if we integrate with respect to x we conclude that $f(x,y) = 3xy + g(y)$, where $g(y)$ is any function of y. On the other hand, if we integrate with respect to y we conclude that $f(x,y) = xy + h(x)$ where $h(x)$ is any function of x. These conclusions are inconsistent for any and every choice of $g(y)$ and $h(x)$, and therefore the differential is inexact. ◄

It is naturally of interest to investigate which properties of a differential make it exact. Consider the general differential containing two variables,

$$df = A(x, y)\, dx + B(x, y)\, dy.$$

We see that

$$\frac{\partial f}{\partial x} = A(x, y), \qquad \frac{\partial f}{\partial y} = B(x, y)$$

and, using the property $f_{xy} = f_{yx}$, we therefore require

$$\frac{\partial A}{\partial y} = \frac{\partial B}{\partial x}. \tag{5.9}$$

This is in fact both a necessary and a sufficient condition for the differential to be exact.

►*Using (5.9) show that* $x\, dy + 3y\, dx$ *is inexact.*

In the above notation, $A(x, y) = 3y$ and $B(x, y) = x$ and so

$$\frac{\partial A}{\partial y} = 3, \qquad \frac{\partial B}{\partial x} = 1.$$

As these are not equal it follows that the differential is inexact. ◄

Determining whether a differential containing many variable $x_1, x_2, \ldots, x_n$ is exact is a simple extension of the above. A differential containing many variables can be written in general as

$$df = \sum_{i=1}^{n} g_i(x_1, x_2, \ldots, x_n)\, dx_i$$

and will be exact if

$$\frac{\partial g_i}{\partial x_j} = \frac{\partial g_j}{\partial x_i} \qquad \text{for all pairs } i, j. \tag{5.10}$$

There will be $\frac{1}{2}n(n-1)$ such relationships to be satisfied.

►*Show that*

$$(y + z)\, dx + x\, dy + x\, dz$$

is an exact differential.

In this case, $g_1(x, y, z) = y + z$, $g_2(x, y, z) = x$, $g_3(x, y, z) = x$ and hence $\partial g_1/\partial y = 1 = \partial g_2/\partial x$, $\partial g_3/\partial x = 1 = \partial g_1/\partial z$, $\partial g_2/\partial z = 0 = \partial g_3/\partial y$; therefore, from (5.10), the differential is exact. As mentioned above, it is sometimes possible to show that a differential is exact simply by finding by inspection the function from which it originates. In this example, it can be seen easily that $f(x, y, z) = x(y + z) + c$. ◄

5.4 Useful theorems of partial differentiation

So far our discussion has centred on a function $f(x, y)$ dependent on two variables, x and y. Equally, however, we could have expressed x as a function of f and y, or y as a function of f and x. To emphasise the point that all the variables are of equal standing, we now replace f by z. This does not imply that x, y and z are coordinate positions (though they might be). Since x is a function of y and z, it follows that

$$dx = \left(\frac{\partial x}{\partial y}\right)_z dy + \left(\frac{\partial x}{\partial z}\right)_y dz \tag{5.11}$$

and similarly, since $y = y(x, z)$,

$$dy = \left(\frac{\partial y}{\partial x}\right)_z dx + \left(\frac{\partial y}{\partial z}\right)_x dz. \tag{5.12}$$

We may now substitute (5.12) into (5.11) to obtain

$$dx = \left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial x}\right)_z dx + \left[\left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial z}\right)_x + \left(\frac{\partial x}{\partial z}\right)_y\right] dz. \tag{5.13}$$

Now if we hold z constant, so that $dz = 0$, we obtain the *reciprocity relation*

$$\left(\frac{\partial x}{\partial y}\right)_z = \left(\frac{\partial y}{\partial x}\right)_z^{-1},$$

which holds provided both partial derivatives exist and neither is equal to zero. Note, further, that this relationship only holds when the variable being kept constant, in this case z, is the same on both sides of the equation.

Alternatively we can put $dx = 0$ in (5.13). Then the contents of the square brackets also equal zero, and we obtain the *cyclic relation*

$$\left(\frac{\partial y}{\partial z}\right)_x \left(\frac{\partial z}{\partial x}\right)_y \left(\frac{\partial x}{\partial y}\right)_z = -1,$$

which holds unless any of the derivatives vanish. In deriving this result we have used the reciprocity relation to replace $(\partial x/\partial z)_y^{-1}$ by $(\partial z/\partial x)_y$.

5.5 The chain rule

So far we have discussed the differentiation of a function $f(x, y)$ with respect to its variables x and y. We now consider the case where x and y are themselves functions of another variable, say u. If we wish to find the derivative df/du, we could simply substitute in $f(x, y)$ the expressions for $x(u)$ and $y(u)$ and then differentiate the resulting function of u. Such substitution will quickly give the desired answer in simple cases, but in more complicated examples it is easier to make use of the total differentials described in the previous section.

From equation (5.5) the total differential of $f(x, y)$ is given by

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy,$$

but we now note that by using the formal device of dividing through by du this immediately implies

$$\frac{df}{du} = \frac{\partial f}{\partial x}\frac{dx}{du} + \frac{\partial f}{\partial y}\frac{dy}{du}, \tag{5.14}$$

which is called the *chain rule* for partial differentiation. This expression provides a direct method for calculating the total derivative of f with respect to u and is particularly useful when an equation is expressed in a parametric form.

►*Given that $x(u) = 1 + au$ and $y(u) = bu^3$, find the rate of change of $f(x, y) = xe^{-y}$ with respect to u.*

As discussed above, this problem could be addressed by substituting for x and y to obtain f as a function only of u and then differentiating with respect to u. However, using (5.14) directly we obtain

$$\frac{df}{du} = (e^{-y})a + (-xe^{-y})3bu^2,$$

which on substituting for x and y gives

$$\frac{df}{du} = e^{-bu^3}(a - 3bu^2 - 3bau^3). \blacktriangleleft$$

Equation (5.14) is an example of the chain rule for a function of two variables each of which depends on a single variable. The chain rule may be extended to functions of many variables, each of which is itself a function of a variable u, i.e. $f(x_1, x_2, x_3, \ldots, x_n)$, with $x_i = x_i(u)$. In this case the chain rule gives

$$\frac{df}{du} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{dx_i}{du} = \frac{\partial f}{\partial x_1}\frac{dx_1}{du} + \frac{\partial f}{\partial x_2}\frac{dx_2}{du} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{du}. \tag{5.15}$$

5.6 Change of variables

It is sometimes necessary or desirable to make a change of variables during the course of an analysis, and consequently to have to change an equation expressed in one set of variables into an equation using another set. The same situation arises if a function f depends on one set of variables x_i, so that $f = f(x_1, x_2, \ldots, x_n)$ but the x_i are themselves functions of a further set of variables u_j and given by the equations

$$x_i = x_i(u_1, u_2, \ldots, u_m). \tag{5.16}$$

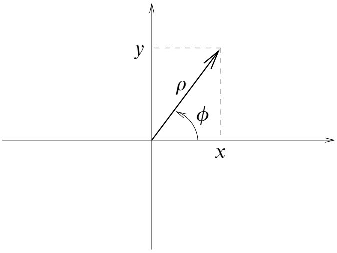


Figure 5.1 The relationship between Cartesian and plane polar coordinates.

For each different value of i, x_i will be a different function of the u_j. In this case the chain rule (5.15) becomes

$$\frac{\partial f}{\partial u_j} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{\partial x_i}{\partial u_j}, \qquad j = 1, 2, \ldots, m, \tag{5.17}$$

and is said to express a *change of variables*. In general the number of variables in each set need not be equal, i.e. m need not equal n, but if both the x_i and the u_i are sets of independent variables then $m = n$.

►*Plane polar coordinates, ρ and ϕ, and Cartesian coordinates, x and y, are related by the expressions*

$$x = \rho\cos\phi, \qquad y = \rho\sin\phi,$$

as can be seen from figure 5.1. An arbitrary function $f(x, y)$ can be re-expressed as a function $g(\rho, \phi)$. Transform the expression

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

into one in ρ and ϕ.

We first note that $\rho^2 = x^2 + y^2$, $\phi = \tan^{-1}(y/x)$. We can now write down the four partial derivatives

$$\frac{\partial \rho}{\partial x} = \frac{x}{(x^2+y^2)^{1/2}} = \cos\phi, \quad \frac{\partial \phi}{\partial x} = \frac{-(y/x^2)}{1+(y/x)^2} = -\frac{\sin\phi}{\rho},$$

$$\frac{\partial \rho}{\partial y} = \frac{y}{(x^2+y^2)^{1/2}} = \sin\phi, \quad \frac{\partial \phi}{\partial y} = \frac{1/x}{1+(y/x)^2} = \frac{\cos\phi}{\rho}.$$

Thus, from (5.17), we may write

$$\frac{\partial}{\partial x} = \cos\phi\frac{\partial}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial}{\partial\phi}, \qquad \frac{\partial}{\partial y} = \sin\phi\frac{\partial}{\partial\rho} + \frac{\cos\phi}{\rho}\frac{\partial}{\partial\phi}.$$

Now it is only a matter of writing

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}\right)f \\ &= \left(\cos\phi\frac{\partial}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial}{\partial\phi}\right)\left(\cos\phi\frac{\partial}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial}{\partial\phi}\right)g \\ &= \left(\cos\phi\frac{\partial}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial}{\partial\phi}\right)\left(\cos\phi\frac{\partial g}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial g}{\partial\phi}\right) \\ &= \cos^2\phi\frac{\partial^2 g}{\partial\rho^2} + \frac{2\cos\phi\sin\phi}{\rho^2}\frac{\partial g}{\partial\phi} - \frac{2\cos\phi\sin\phi}{\rho}\frac{\partial^2 g}{\partial\phi\partial\rho} \\ &\quad + \frac{\sin^2\phi}{\rho}\frac{\partial g}{\partial\rho} + \frac{\sin^2\phi}{\rho^2}\frac{\partial^2 g}{\partial\phi^2}\end{aligned}$$

and a similar expression for $\partial^2 f/\partial y^2$,

$$\begin{aligned}\frac{\partial^2 f}{\partial y^2} &= \left(\sin\phi\frac{\partial}{\partial\rho} + \frac{\cos\phi}{\rho}\frac{\partial}{\partial\phi}\right)\left(\sin\phi\frac{\partial}{\partial\rho} + \frac{\cos\phi}{\rho}\frac{\partial}{\partial\phi}\right)g \\ &= \sin^2\phi\frac{\partial^2 g}{\partial\rho^2} - \frac{2\cos\phi\sin\phi}{\rho^2}\frac{\partial g}{\partial\phi} + \frac{2\cos\phi\sin\phi}{\rho}\frac{\partial^2 g}{\partial\phi\partial\rho} \\ &\quad + \frac{\cos^2\phi}{\rho}\frac{\partial g}{\partial\rho} + \frac{\cos^2\phi}{\rho^2}\frac{\partial^2 g}{\partial\phi^2}.\end{aligned}$$

When these two expressions are added together the change of variables is complete and we obtain

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 g}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial g}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2 g}{\partial\phi^2}. \blacktriangleleft$$

5.7 Taylor's theorem for many-variable functions

We have already introduced Taylor's theorem for a function $f(x)$ of one variable, in section 4.6. In an analogous way, the Taylor expansion of a function $f(x, y)$ of two variables is given by

$$\begin{aligned}f(x, y) &= f(x_0, y_0) + \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y \\ &\quad + \frac{1}{2!}\left[\frac{\partial^2 f}{\partial x^2}(\Delta x)^2 + 2\frac{\partial^2 f}{\partial x\partial y}\Delta x\Delta y + \frac{\partial^2 f}{\partial y^2}(\Delta y)^2\right] + \cdots, \qquad (5.18)\end{aligned}$$

where $\Delta x = x - x_0$ and $\Delta y = y - y_0$, and all the derivatives are to be evaluated at (x_0, y_0).

►*Find the Taylor expansion, up to quadratic terms in* $x-2$ *and* $y-3$, *of* $f(x,y) = y \exp xy$ *about the point* $x=2$, $y=3$.

We first evaluate the required partial derivatives of the function, i.e.

$$\frac{\partial f}{\partial x} = y^2 \exp xy, \qquad \frac{\partial f}{\partial y} = \exp xy + xy \exp xy,$$

$$\frac{\partial^2 f}{\partial x^2} = y^3 \exp xy, \qquad \frac{\partial^2 f}{\partial y^2} = 2x \exp xy + x^2 y \exp xy,$$

$$\frac{\partial^2 f}{\partial x \partial y} = 2y \exp xy + xy^2 \exp xy.$$

Using (5.18), the Taylor expansion of a two-variable function, we find

$$f(x,y) \approx e^6 \Big\{ 3 + 9(x-2) + 7(y-3) \\ + (2!)^{-1} \left[27(x-2)^2 + 48(x-2)(y-3) + 16(y-3)^2 \right] \Big\}. \blacktriangleleft$$

It will be noticed that the terms in (5.18) containing first derivatives can be written as

$$\frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y = \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right) f(x,y),$$

where both sides of this relation should be evaluated at the point (x_0, y_0). Similarly the terms in (5.18) containing second derivatives can be written as

$$\frac{1}{2!}\left[\frac{\partial^2 f}{\partial x^2}(\Delta x)^2 + 2\frac{\partial^2 f}{\partial x \partial y}\Delta x \Delta y + \frac{\partial^2 f}{\partial y^2}(\Delta y)^2 \right] = \frac{1}{2!}\left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right)^2 f(x,y), \tag{5.19}$$

where it is understood that the partial derivatives resulting from squaring the expression in parentheses act only on $f(x,y)$ and its derivatives, and not on Δx or Δy; again both sides of (5.19) should be evaluated at (x_0, y_0). It can be shown that the higher-order terms of the Taylor expansion of $f(x,y)$ can be written in an analogous way, and that we may write the full Taylor series as

$$f(x,y) = \sum_{n=0}^{\infty} \frac{1}{n!} \left[\left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y} \right)^n f(x,y) \right]_{x_0, y_0}$$

where, as indicated, all the terms on the RHS are to be evaluated at (x_0, y_0).

The most general form of Taylor's theorem, for a function $f(x_1, x_2, \ldots, x_n)$ of n variables, is a simple extension of the above. Although it is not necessary to do so, we may think of the x_i as coordinates in n-dimensional space and write the function as $f(\mathbf{x})$, where $\mathbf{x}$ is a vector from the origin to $(x_1, x_2, \ldots, x_n)$. Taylor's theorem then becomes

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \sum_i \frac{\partial f}{\partial x_i}\Delta x_i + \frac{1}{2!}\sum_i \sum_j \frac{\partial^2 f}{\partial x_i \partial x_j}\Delta x_i \Delta x_j + \cdots, \qquad (5.20)$$

where $\Delta x_i = x_i - x_{i_0}$ and the partial derivatives are evaluated at $(x_{1_0}, x_{2_0}, \ldots, x_{n_0})$. For completeness, we note that in this case the full Taylor series can be written in the form

$$f(\mathbf{x}) = \sum_{n=0}^{\infty} \frac{1}{n!}\left[(\Delta\mathbf{x}\cdot\nabla)^n f(\mathbf{x})\right]_{\mathbf{x}=\mathbf{x}_0},$$

where ∇ is the vector differential operator del, to be discussed in chapter 10.

5.8 Stationary values of many-variable functions

The idea of the *stationary points* of a function of just one variable has already been discussed in subsection 2.1.8. We recall that the function $f(x)$ has a stationary point at $x = x_0$ if its gradient df/dx is zero at that point. A function may have any number of stationary points, and their nature, i.e. whether they are maxima, minima or stationary points of inflection, is determined by the value of the second derivative at the point. A stationary point is

(i) a minimum if $d^2f/dx^2 > 0$;
(ii) a maximum if $d^2f/dx^2 < 0$;
(iii) a stationary point of inflection if $d^2f/dx^2 = 0$ and changes sign through the point.

We now consider the stationary points of functions of more than one variable; we will see that partial differential analysis is ideally suited to the determination of the position and nature of such points. It is helpful to consider first the case of a function of just two variables but, even in this case, the general situation is more complex than that for a function of one variable, as can be seen from figure 5.2.

This figure shows part of a three-dimensional model of a function $f(x, y)$. At positions P and B there are a peak and a bowl respectively or, more mathematically, a local maximum and a local minimum. At position S the gradient in any direction is zero but the situation is complicated, since a section parallel to the plane $x = 0$ would show a maximum, but one parallel to the plane $y = 0$ would show a minimum. A point such as S is known as a *saddle point*. The orientation of the ‘saddle’ in the xy-plane is irrelevant; it is as shown in the figure solely for ease of discussion. For any saddle point the function increases in some directions away from the point but decreases in other directions.

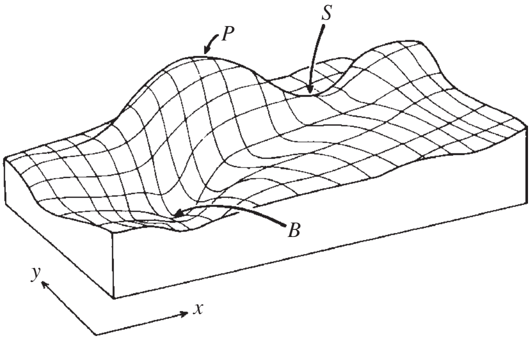


Figure 5.2 Stationary points of a function of two variables. A minimum occurs at B, a maximum at P and a saddle point at S.

For functions of two variables, such as the one shown, it should be clear that a necessary condition for a stationary point (maximum, minimum or saddle point) to occur is that

$$\frac{\partial f}{\partial x} = 0 \qquad \text{and} \qquad \frac{\partial f}{\partial y} = 0. \tag{5.21}$$

The vanishing of the partial derivatives in directions parallel to the axes is enough to ensure that the partial derivative in any arbitrary direction is also zero. The latter can be considered as the superposition of two contributions, one along each axis; since both contributions are zero, so is the partial derivative in the arbitrary direction. This may be made more precise by considering the total differential

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy.$$

Using (5.21) we see that although the infinitesimal changes dx and dy can be chosen independently the change in the value of the infinitesimal function df is always zero at a stationary point.

We now turn our attention to determining the nature of a stationary point of a function of two variables, i.e. whether it is a maximum, a minimum or a saddle point. By analogy with the one-variable case we see that $\partial^2 f/\partial x^2$ and $\partial^2 f/\partial y^2$ must both be positive for a minimum and both be negative for a maximum. However these are not sufficient conditions since they could also be obeyed at complicated saddle points. What is important for a minimum (or maximum) is that the second partial derivative must be positive (or negative) in *all* directions, not just in the x- and y- directions.

To establish just what constitutes sufficient conditions we first note that, since f is a function of two variables and $\partial f/\partial x = \partial f/\partial y = 0$, a Taylor expansion of the type (5.18) about the stationary point yields

$$f(x,y) - f(x_0,y_0) \approx \frac{1}{2!}\left[(\Delta x)^2 f_{xx} + 2\Delta x \Delta y f_{xy} + (\Delta y)^2 f_{yy}\right],$$

where $\Delta x = x - x_0$ and $\Delta y = y - y_0$ and where the partial derivatives have been written in more compact notation. Rearranging the contents of the bracket as the weighted sum of two squares, we find

$$f(x,y) - f(x_0,y_0) \approx \frac{1}{2}\left[f_{xx}\left(\Delta x + \frac{f_{xy}\Delta y}{f_{xx}}\right)^2 + (\Delta y)^2\left(f_{yy} - \frac{f_{xy}^2}{f_{xx}}\right)\right]. \tag{5.22}$$

For a minimum, we require (5.22) to be positive for all Δx and Δy, and hence $f_{xx} > 0$ and $f_{yy} - (f_{xy}^2/f_{xx}) > 0$. Given the first constraint, the second can be written $f_{xx}f_{yy} > f_{xy}^2$. Similarly for a maximum we require (5.22) to be negative, and hence $f_{xx} < 0$ and $f_{xx}f_{yy} > f_{xy}^2$. For minima and maxima, symmetry requires that f_{yy} obeys the same criteria as f_{xx}. When (5.22) is negative (or zero) for some values of Δx and Δy but positive (or zero) for others, we have a saddle point. In this case $f_{xx}f_{yy} < f_{xy}^2$. In summary, all stationary points have $f_x = f_y = 0$ and they may be classified further as

(i) minima if both f_{xx} and f_{yy} are positive *and* $f_{xy}^2 < f_{xx}f_{yy}$,

(ii) maxima if both f_{xx} and f_{yy} are negative *and* $f_{xy}^2 < f_{xx}f_{yy}$,

(iii) saddle points if f_{xx} and f_{yy} have opposite signs *or* $f_{xy}^2 > f_{xx}f_{yy}$.

Note, however, that if $f_{xy}^2 = f_{xx}f_{yy}$ then $f(x,y) - f(x_0,y_0)$ can be written in one of the four forms

$$\pm\frac{1}{2}\left(\Delta x|f_{xx}|^{1/2} \pm \Delta y|f_{yy}|^{1/2}\right)^2.$$

For some choice of the ratio $\Delta y/\Delta x$ this expression has zero value, showing that, for a displacement from the stationary point in this particular direction, $f(x_0 + \Delta x, y_0 + \Delta y)$ does not differ from $f(x_0,y_0)$ to second order in Δx and Δy; in such situations further investigation is required. In particular, if f_{xx}, f_{yy} and f_{xy} are all zero then the Taylor expansion has to be taken to a higher order. As examples, such extended investigations would show that the function $f(x,y) = x^4 + y^4$ has a minimum at the origin but that $g(x,y) = x^4 + y^3$ has a saddle point there.

►*Show that the function $f(x, y) = x^3 \exp(-x^2 - y^2)$ has a maximum at the point $(\sqrt{3/2}, 0)$, a minimum at $(-\sqrt{3/2}, 0)$ and a stationary point at the origin whose nature cannot be determined by the above procedures.*

Setting the first two partial derivatives to zero to locate the stationary points, we find

$$\frac{\partial f}{\partial x} = (3x^2 - 2x^4)\exp(-x^2 - y^2) = 0, \tag{5.23}$$

$$\frac{\partial f}{\partial y} = -2yx^3 \exp(-x^2 - y^2) = 0. \tag{5.24}$$

For (5.24) to be satisfied we require $x = 0$ or $y = 0$ and for (5.23) to be satisfied we require $x = 0$ or $x = \pm\sqrt{3/2}$. Hence the stationary points are at $(0, 0)$, $(\sqrt{3/2}, 0)$ and $(-\sqrt{3/2}, 0)$. We now find the second partial derivatives:

$$\begin{aligned} f_{xx} &= (4x^5 - 14x^3 + 6x)\exp(-x^2 - y^2), \\ f_{yy} &= x^3(4y^2 - 2)\exp(-x^2 - y^2), \\ f_{xy} &= 2x^2y(2x^2 - 3)\exp(-x^2 - y^2). \end{aligned}$$

We then substitute the pairs of values of x and y for each stationary point and find that at $(0, 0)$

$$f_{xx} = 0, \qquad f_{yy} = 0, \qquad f_{xy} = 0$$

and at $(\pm\sqrt{3/2}, 0)$

$$f_{xx} = \mp 6\sqrt{3/2}\ \exp(-3/2), \qquad f_{yy} = \mp 3\sqrt{3/2}\ \exp(-3/2), \qquad f_{xy} = 0.$$

Hence, applying criteria (i)–(iii) above, we find that $(0, 0)$ is an undetermined stationary point, $(\sqrt{3/2}, 0)$ is a maximum and $(-\sqrt{3/2}, 0)$ is a minimum. The function is shown in figure 5.3. ◄

Determining the nature of stationary points for functions of a general number of variables is considerably more difficult and requires a knowledge of the eigenvectors and eigenvalues of matrices. Although these are not discussed until chapter 8, we present the analysis here for completeness. The remainder of this section can therefore be omitted on a first reading.

For a function of n real variables, $f(x_1, x_2, \ldots, x_n)$, we require that, at all stationary points,

$$\frac{\partial f}{\partial x_i} = 0 \qquad \text{for all } x_i.$$

In order to determine the nature of a stationary point, we must expand the function as a Taylor series about the point. Recalling the Taylor expansion (5.20) for a function of n variables, we see that

$$\Delta f = f(\mathbf{x}) - f(\mathbf{x}_0) \approx \frac{1}{2}\sum_i \sum_j \frac{\partial^2 f}{\partial x_i \partial x_j}\Delta x_i \Delta x_j. \tag{5.25}$$

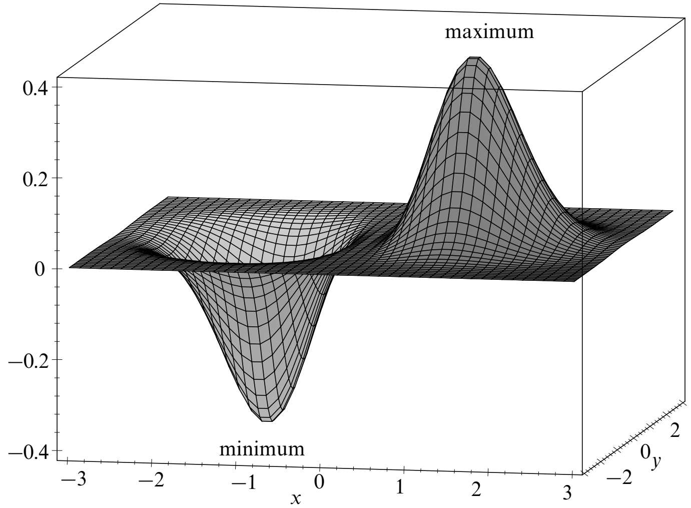


Figure 5.3 The function $f(x, y) = x^3 \exp(-x^2 - y^2)$.

If we define the matrix M to have elements given by

$$M_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j},$$

then we can rewrite (5.25) as

$$\Delta f = \tfrac{1}{2} \Delta \mathsf{x}^{\mathrm{T}} \mathsf{M} \Delta \mathsf{x}, \qquad (5.26)$$

where $\Delta \mathsf{x}$ is the column vector with the Δx_i as its components and $\Delta \mathsf{x}^{\mathrm{T}}$ is its transpose. Since M is real and symmetric it has n real eigenvalues λ_r and n orthogonal eigenvectors e_r, which after suitable normalisation satisfy

$$\mathsf{M}\mathsf{e}_r = \lambda_r \mathsf{e}_r, \qquad \mathsf{e}_r^{\mathrm{T}} \mathsf{e}_s = \delta_{rs},$$

where the *Kronecker delta*, written δ_{rs}, equals unity for $r = s$ and equals zero otherwise. These eigenvectors form a basis set for the n-dimensional space and we can therefore expand $\Delta \mathsf{x}$ in terms of them, obtaining

$$\Delta \mathsf{x} = \sum_r a_r \mathsf{e}_r,$$

where the a_r are coefficients dependent upon $\Delta\mathsf{x}$. Substituting this into (5.26), we find

$$\Delta f = \tfrac{1}{2}\Delta\mathsf{x}^{\mathrm{T}}\mathsf{M}\Delta\mathsf{x} = \tfrac{1}{2}\sum_r \lambda_r a_r^2.$$

Now, for the stationary point to be a minimum, we require $\Delta f = \frac{1}{2}\sum_r \lambda_r a_r^2 > 0$ for all sets of values of the a_r, and therefore all the eigenvalues of M to be greater than zero. Conversely, for a maximum we require $\Delta f = \frac{1}{2}\sum_r \lambda_r a_r^2 < 0$, and therefore all the eigenvalues of M to be less than zero. If the eigenvalues have mixed signs, then we have a saddle point. Note that the test may fail if some or all of the eigenvalues are equal to zero and all the non-zero ones have the same sign.

►*Derive the conditions for maxima, minima and saddle points for a function of two real variables, using the above analysis.*

For a two-variable function the matrix M is given by

$$\mathsf{M} = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix}.$$

Therefore its eigenvalues satisfy the equation

$$\begin{vmatrix} f_{xx} - \lambda & f_{xy} \\ f_{xy} & f_{yy} - \lambda \end{vmatrix} = 0.$$

Hence

$$(f_{xx} - \lambda)(f_{yy} - \lambda) - f_{xy}^2 = 0$$

$$\Rightarrow \qquad f_{xx}f_{yy} - (f_{xx} + f_{yy})\lambda + \lambda^2 - f_{xy}^2 = 0$$

$$\Rightarrow \qquad 2\lambda = (f_{xx} + f_{yy}) \pm \sqrt{(f_{xx} + f_{yy})^2 - 4(f_{xx}f_{yy} - f_{xy}^2)},$$

which by rearrangement of the terms under the square root gives

$$2\lambda = (f_{xx} + f_{yy}) \pm \sqrt{(f_{xx} - f_{yy})^2 + 4f_{xy}^2}.$$

Now, that M is real and symmetric implies that its eigenvalues are real, and so for both eigenvalues to be positive (corresponding to a minimum), we require f_{xx} and f_{yy} positive and also

$$f_{xx} + f_{yy} > \sqrt{(f_{xx} + f_{yy})^2 - 4(f_{xx}f_{yy} - f_{xy}^2)},$$

$$\Rightarrow \qquad f_{xx}f_{yy} - f_{xy}^2 > 0.$$

A similar procedure will find the criteria for maxima and saddle points. ◄

5.9 Stationary values under constraints

In the previous section we looked at the problem of finding stationary values of a function of two or more variables when all the variables may be independently varied. However, it is often the case in physical problems that not all the variables used to describe a situation are in fact independent, i.e. some relationship between the variables must be satisfied. For example, if we walk through a hilly landscape and we are constrained to walk along a path, we will never reach the highest peak on the landscape unless the path happens to take us to it. Nevertheless, we can still find the highest point that we have reached during our journey.

We first discuss the case of a function of just two variables. Let us consider finding the maximum value of the differentiable function $f(x, y)$ subject to the constraint $g(x, y) = c$, where c is a constant. In the above analogy, $f(x, y)$ might represent the height of the land above sea-level in some hilly region, whilst $g(x, y) = c$ is the equation of the path along which we walk.

We could, of course, use the constraint $g(x, y) = c$ to substitute for x or y in $f(x, y)$, thereby obtaining a new function of only one variable whose stationary points could be found using the methods discussed in subsection 2.1.8. However, such a procedure can involve a lot of algebra and becomes very tedious for functions of more than two variables. A more direct method for solving such problems is the *method of Lagrange undetermined multipliers*, which we now discuss.

To maximise f we require

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy = 0.$$

If dx and dy were independent, we could conclude $f_x = 0 = f_y$. However, here they are not independent, but constrained because g is constant:

$$dg = \frac{\partial g}{\partial x}dx + \frac{\partial g}{\partial y}dy = 0.$$

Multiplying dg by an as yet unknown number λ and adding it to df we obtain

$$d(f + \lambda g) = \left(\frac{\partial f}{\partial x} + \lambda\frac{\partial g}{\partial x}\right)dx + \left(\frac{\partial f}{\partial y} + \lambda\frac{\partial g}{\partial y}\right)dy = 0,$$

where λ is called a *Lagrange undetermined multiplier*. In this equation dx and dy are to be independent and arbitrary; we must therefore choose λ such that

$$\frac{\partial f}{\partial x} + \lambda\frac{\partial g}{\partial x} = 0, \tag{5.27}$$

$$\frac{\partial f}{\partial y} + \lambda\frac{\partial g}{\partial y} = 0. \tag{5.28}$$

These equations, together with the constraint $g(x, y) = c$, are sufficient to find the three unknowns, i.e. λ and the values of x and y at the stationary point.

►*The temperature of a point (x, y) on a unit circle is given by $T(x, y) = 1 + xy$. Find the temperature of the two hottest points on the circle.*

We need to maximise $T(x, y)$ subject to the constraint $x^2 + y^2 = 1$. Applying (5.27) and (5.28), we obtain

$$y + 2\lambda x = 0, \tag{5.29}$$

$$x + 2\lambda y = 0. \tag{5.30}$$

These results, together with the original constraint $x^2 + y^2 = 1$, provide three simultaneous equations that may be solved for λ, x and y.

From (5.29) and (5.30) we find $\lambda = \pm 1/2$, which in turn implies that $y = \mp x$. Remembering that $x^2 + y^2 = 1$, we find that

$$\begin{aligned} y = x \quad &\Rightarrow \quad x = \pm\frac{1}{\sqrt{2}}, \qquad y = \pm\frac{1}{\sqrt{2}} \\ y = -x \quad &\Rightarrow \quad x = \mp\frac{1}{\sqrt{2}}, \qquad y = \pm\frac{1}{\sqrt{2}}. \end{aligned}$$

We have not yet determined which of these stationary points are maxima and which are minima. In this simple case, we need only substitute the four pairs of x- and y- values into $T(x, y) = 1 + xy$ to find that the maximum temperature on the unit circle is $T_{\max} = 3/2$ at the points $y = x = \pm 1/\sqrt{2}$. ◄

The method of Lagrange multipliers can be used to find the stationary points of functions of more than two variables, subject to several constraints, provided that the number of constraints is smaller than the number of variables. For example, if we wish to find the stationary points of $f(x, y, z)$ subject to the constraints $g(x, y, z) = c_1$ and $h(x, y, z) = c_2$, where c_1 and c_2 are constants, then we proceed as above, obtaining

$$\begin{aligned} \frac{\partial}{\partial x}(f + \lambda g + \mu h) = \frac{\partial f}{\partial x} + \lambda\frac{\partial g}{\partial x} + \mu\frac{\partial h}{\partial x} &= 0, \\ \frac{\partial}{\partial y}(f + \lambda g + \mu h) = \frac{\partial f}{\partial y} + \lambda\frac{\partial g}{\partial y} + \mu\frac{\partial h}{\partial y} &= 0, \\ \frac{\partial}{\partial z}(f + \lambda g + \mu h) = \frac{\partial f}{\partial z} + \lambda\frac{\partial g}{\partial z} + \mu\frac{\partial h}{\partial z} &= 0. \end{aligned} \tag{5.31}$$

We may now solve these three equations, together with the two constraints, to give λ, μ, x, y and z.

►*Find the stationary points of* $f(x, y, z) = x^3 + y^3 + z^3$ *subject to the following constraints:*

(i) $g(x, y, z) = x^2 + y^2 + z^2 = 1$;
(ii) $g(x, y, z) = x^2 + y^2 + z^2 = 1$ *and* $h(x, y, z) = x + y + z = 0$.

Case (i). Since there is only one constraint in this case, we need only introduce a single Lagrange multiplier to obtain

$$\begin{aligned} \frac{\partial}{\partial x}(f + \lambda g) &= 3x^2 + 2\lambda x = 0, \\ \frac{\partial}{\partial y}(f + \lambda g) &= 3y^2 + 2\lambda y = 0, \\ \frac{\partial}{\partial z}(f + \lambda g) &= 3z^2 + 2\lambda z = 0. \end{aligned} \tag{5.32}$$

These equations are highly symmetrical and clearly have the solution $x = y = z = -2\lambda/3$. Using the constraint $x^2 + y^2 + z^2 = 1$ we find $\lambda = \pm\sqrt{3}/2$ and so stationary points occur at

$$x = y = z = \pm\frac{1}{\sqrt{3}}. \tag{5.33}$$

In solving the three equations (5.32) in this way, however, we have implicitly assumed that x, y and z are non-zero. However, it is clear from (5.32) that any of these values can equal zero, with the exception of the case $x = y = z = 0$ since this is prohibited by the constraint $x^2 + y^2 + z^2 = 1$. We must consider the other cases separately.

If $x = 0$, for example, we require

$$\begin{aligned} 3y^2 + 2\lambda y &= 0, \\ 3z^2 + 2\lambda z &= 0, \\ y^2 + z^2 &= 1. \end{aligned}$$

Clearly, we require $\lambda \neq 0$, otherwise these equations are inconsistent. If neither y nor z is zero we find $y = -2\lambda/3 = z$ and from the third equation we require $y = z = \pm 1/\sqrt{2}$. If $y = 0$, however, then $z = \pm 1$ and, similarly, if $z = 0$ then $y = \pm 1$. Thus the stationary points having $x = 0$ are $(0, 0, \pm 1)$, $(0, \pm 1, 0)$ and $(0, \pm 1/\sqrt{2}, \pm 1/\sqrt{2})$. A similar procedure can be followed for the cases $y = 0$ and $z = 0$ respectively and, in addition to those already obtained, we find the stationary points $(\pm 1, 0, 0)$, $(\pm 1/\sqrt{2}, 0, \pm 1/\sqrt{2})$ and $(\pm 1/\sqrt{2}, \pm 1/\sqrt{2}, 0)$.

Case (ii). We now have two constraints and must therefore introduce two Lagrange multipliers to obtain (cf. (5.31))

$$\frac{\partial}{\partial x}(f + \lambda g + \mu h) = 3x^2 + 2\lambda x + \mu = 0, \tag{5.34}$$

$$\frac{\partial}{\partial y}(f + \lambda g + \mu h) = 3y^2 + 2\lambda y + \mu = 0, \tag{5.35}$$

$$\frac{\partial}{\partial z}(f + \lambda g + \mu h) = 3z^2 + 2\lambda z + \mu = 0. \tag{5.36}$$

These equations are again highly symmetrical and the simplest way to proceed is to subtract (5.35) from (5.34) to obtain

$$\begin{aligned} &3(x^2 - y^2) + 2\lambda(x - y) = 0 \\ \Rightarrow \quad &3(x + y)(x - y) + 2\lambda(x - y) = 0. \end{aligned} \tag{5.37}$$

This equation is clearly satisfied if $x = y$; then, from the second constraint, $x + y + z = 0$, we find $z = -2x$. Substituting these values into the first constraint, $x^2 + y^2 + z^2 = 1$, we obtain

$$x = \pm\frac{1}{\sqrt{6}}, \qquad y = \pm\frac{1}{\sqrt{6}}, \qquad z = \mp\frac{2}{\sqrt{6}}. \tag{5.38}$$

Because of the high degree of symmetry amongst the equations (5.34)–(5.36), we may obtain by inspection two further relations analogous to (5.37), one containing the variables y, z and the other the variables x, z. Assuming $y = z$ in the first relation and $x = z$ in the second, we find the stationary points

$$x = \pm\frac{1}{\sqrt{6}}, \qquad y = \mp\frac{2}{\sqrt{6}}, \qquad z = \pm\frac{1}{\sqrt{6}} \tag{5.39}$$

and

$$x = \mp\frac{2}{\sqrt{6}}, \qquad y = \pm\frac{1}{\sqrt{6}}, \qquad z = \pm\frac{1}{\sqrt{6}}. \tag{5.40}$$

We note that in finding the stationary points (5.38)–(5.40) we did not need to evaluate the Lagrange multipliers λ and μ explicitly. This is not always the case, however, and in some problems it may be simpler to begin by finding the values of these multipliers.

Returning to (5.37) we must now consider the case where $x \neq y$; then we find

$$3(x + y) + 2\lambda = 0. \tag{5.41}$$

However, in obtaining the stationary points (5.39), (5.40), we did *not* assume $x = y$ but only required $y = z$ and $x = z$ respectively. It is clear that $x \neq y$ at these stationary points, and it can be shown that they do indeed satisfy (5.41). Similarly, several stationary points for which $x \neq z$ or $y \neq z$ have already been found.

Thus we need to consider further only two cases, $x = y = z$, and x, y and z are all different. The first is clearly prohibited by the constraint $x + y + z = 0$. For the second case, (5.41) must be satisfied, together with the analogous equations containing y, z and x, z respectively, i.e.

$$3(x + y) + 2\lambda = 0,$$
$$3(y + z) + 2\lambda = 0,$$
$$3(x + z) + 2\lambda = 0.$$

Adding these three equations together and using the constraint $x+y+z = 0$ we find $\lambda = 0$. However, for $\lambda = 0$ the equations are inconsistent for non-zero x, y and z. Therefore all the stationary points have already been found and are given by (5.38)–(5.40). ◄

The method may be extended to functions of any number n of variables subject to any smaller number m of constraints. This means that effectively there are $n - m$ independent variables and, as mentioned above, we could solve by substitution and then by the methods of the previous section. However, for large n this becomes cumbersome and the use of Lagrange undetermined multipliers is a useful simplification.

►*A system contains a very large number N of particles, each of which can be in any of R energy levels with a corresponding energy E_i, $i = 1, 2, \ldots, R$. The number of particles in the ith level is n_i and the total energy of the system is a constant, E. Find the distribution of particles amongst the energy levels that maximises the expression*

$$P = \frac{N!}{n_1!n_2!\cdots n_R!},$$

subject to the constraints that both the number of particles and the total energy remain constant, i.e.

$$g = N - \sum_{i=1}^{R} n_i = 0 \qquad \text{and} \qquad h = E - \sum_{i=1}^{R} n_i E_i = 0.$$

The way in which we proceed is as follows. In order to maximise P, we must minimise its denominator (since the numerator is fixed). Minimising the denominator is the same as minimising the logarithm of the denominator, i.e.

$$f = \ln(n_1!n_2!\cdots n_R!) = \ln(n_1!) + \ln(n_2!) + \cdots + \ln(n_R!).$$

Using Stirling's approximation, $\ln(n!) \approx n \ln n - n$, we find that

$$\begin{aligned} f &= n_1 \ln n_1 + n_2 \ln n_2 + \cdots + n_R \ln n_R - (n_1 + n_2 + \cdots + n_R) \\ &= \left(\sum_{i=1}^{R} n_i \ln n_i \right) - N. \end{aligned}$$

It has been assumed here that, for the desired distribution, all the n_i are large. Thus, we now have a function f subject to two constraints, $g = 0$ and $h = 0$, and we can apply the Lagrange method, obtaining (cf. (5.31))

$$\begin{aligned} \frac{\partial f}{\partial n_1} + \lambda \frac{\partial g}{\partial n_1} + \mu \frac{\partial h}{\partial n_1} &= 0, \\ \frac{\partial f}{\partial n_2} + \lambda \frac{\partial g}{\partial n_2} + \mu \frac{\partial h}{\partial n_2} &= 0, \\ &\vdots \\ \frac{\partial f}{\partial n_R} + \lambda \frac{\partial g}{\partial n_R} + \mu \frac{\partial h}{\partial n_R} &= 0. \end{aligned}$$

Since all these equations are alike, we consider the general case

$$\frac{\partial f}{\partial n_k} + \lambda \frac{\partial g}{\partial n_k} + \mu \frac{\partial h}{\partial n_k} = 0,$$

for $k = 1, 2, \ldots, R$. Substituting the functions f, g and h into this relation we find

$$\frac{n_k}{n_k} + \ln n_k + \lambda(-1) + \mu(-E_k) = 0,$$

which can be rearranged to give

$$\ln n_k = \mu E_k + \lambda - 1,$$

and hence

$$n_k = C \exp \mu E_k.$$

We now have the general form for the distribution of particles amongst energy levels, but in order to determine the two constants μ, C we recall that

$$\sum_{k=1}^{R} C \exp \mu E_k = N$$

and

$$\sum_{k=1}^{R} CE_k \exp \mu E_k = E.$$

This is known as the Boltzmann distribution and is a well-known result from statistical mechanics. ◀

5.10 Envelopes

As noted at the start of this chapter, many of the functions with which physicists, chemists and engineers have to deal contain, in addition to constants and one or more variables, quantities that are normally considered as parameters of the system under study. Such parameters may, for example, represent the capacitance of a capacitor, the length of a rod, or the mass of a particle – quantities that are normally taken as fixed for any particular physical set-up. The corresponding variables may well be time, currents, charges, positions and velocities. However, the parameters *could* be varied and in this section we study the effects of doing so; in particular we study how the form of dependence of one variable on another, typically $y = y(x)$, is affected when the value of a parameter is changed in a smooth and continuous way. In effect, we are making the parameter into an additional variable.

As a particular parameter, which we denote by α, is varied over its permitted range, the shape of the plot of y against x will change, usually, but not always, in a smooth and continuous way. For example, if the muzzle speed v of a shell fired from a gun is increased through a range of values then its height–distance trajectories will be a series of curves with a common starting point that are essentially just magnified copies of the original; furthermore the curves do not cross each other. However, if the muzzle speed is kept constant but θ, the angle of elevation of the gun, is increased through a series of values, the corresponding trajectories do not vary in a monotonic way. When θ has been increased beyond 45° the trajectories then do cross some of the trajectories corresponding to $\theta < 45°$. The trajectories for $\theta > 45°$ all lie within a curve that touches each individual trajectory at one point. Such a curve is called the *envelope* to the set of trajectory solutions; it is to the study of such envelopes that this section is devoted.

For our general discussion of envelopes we will consider an equation of the form $f = f(x, y, \alpha) = 0$. A function of three Cartesian variables, $f = f(x, y, \alpha)$, is defined at all points in $xy\alpha$-space, whereas $f = f(x, y, \alpha) = 0$ is a *surface* in this space. A plane of constant α, which is parallel to the xy-plane, cuts such

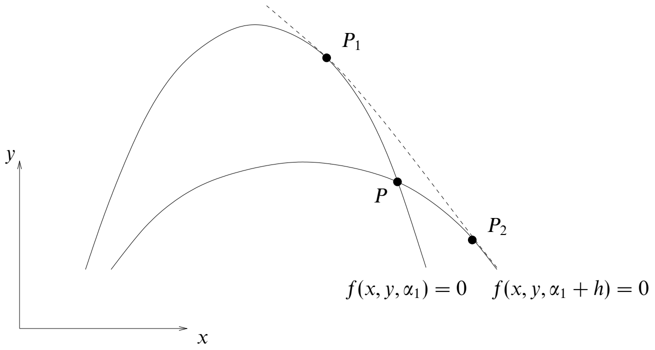


Figure 5.4 Two neighbouring curves in the xy-plane of the family $f(x, y, \alpha) = 0$ intersecting at P. For fixed α_1, the point P_1 is the limiting position of P as $h \to 0$. As α_1 is varied, P_1 delineates the envelope of the family (broken line).

a surface in a curve. Thus different values of the parameter α correspond to different curves, which can be plotted in the xy-plane. We now investigate how the *envelope equation* for such a family of curves is obtained.

5.10.1 Envelope equations

Suppose $f(x, y, \alpha_1) = 0$ and $f(x, y, \alpha_1 + h) = 0$ are two neighbouring curves of a family for which the parameter α differs by a small amount h. Let them intersect at the point P with coordinates x, y, as shown in figure 5.4. Then the envelope, indicated by the broken line in the figure, touches $f(x, y, \alpha_1) = 0$ at the point P_1, which is defined as the limiting position of P when α_1 is fixed but $h \to 0$. The full envelope is the curve traced out by P_1 as α_1 changes to generate successive members of the family of curves. Of course, for any finite h, $f(x, y, \alpha_1 + h) = 0$ is one of these curves and the envelope touches it at the point P_2.

We are now going to apply Rolle's theorem, see subsection 2.1.10, with the parameter α as the independent variable and x and y fixed as constants. In this context, the two curves in figure 5.4 can be thought of as the projections onto the xy-plane of the planar curves in which the *surface* $f = f(x,\ y,\ \alpha) = 0$ meets the planes $\alpha = \alpha_1$ and $\alpha = \alpha_1 + h$.

Along the normal to the page that passes through P, as α changes from α_1 to $\alpha_1 + h$ the value of $f = f(x,\ y,\ \alpha)$ will depart from zero, because the normal meets the surface $f = f(x,\ y,\ \alpha) = 0$ only at $\alpha = \alpha_1$ and at $\alpha = \alpha_1 + h$. However, at these end points the values of $f = f(x,\ y,\ \alpha)$ will both be zero, and therefore equal. This allows us to apply Rolle's theorem and so to conclude that for some θ in the range $0 \leq \theta \leq 1$ the partial derivative $\partial f(x,\ y,\ \alpha_1 + \theta h)/\partial\alpha$ is zero. When

h is made arbitrarily small, so that $P \to P_1$, the three defining equations reduce to two, which define the envelope point P_1:

$$f(x, y, \alpha_1) = 0 \qquad \text{and} \qquad \frac{\partial f(x, y, \alpha_1)}{\partial \alpha} = 0. \tag{5.42}$$

In (5.42) both the function and the gradient are evaluated at $\alpha = \alpha_1$. The equation of the envelope $g(x, y) = 0$ is found by eliminating α_1 between the two equations.

As a simple example we will now solve the problem which when posed mathematically reads 'calculate the envelope appropriate to the family of straight lines in the xy-plane whose points of intersection with the coordinate axes are a fixed distance apart'. In more ordinary language, the problem is about a ladder leaning against a wall.

►*A ladder of length L stands on level ground and can be leaned at any angle against a vertical wall. Find the equation of the curve bounding the vertical area below the ladder.*

We take the ground and the wall as the x- and y-axes respectively. If the foot of the ladder is a from the foot of the wall and the top is b above the ground then the straight-line equation of the ladder is

$$\frac{x}{a} + \frac{y}{b} = 1,$$

where a and b are connected by $a^2 + b^2 = L^2$. Expressed in standard form with only one independent parameter, a, the equation becomes

$$f(x, y, a) = \frac{x}{a} + \frac{y}{(L^2 - a^2)^{1/2}} - 1 = 0. \tag{5.43}$$

Now, differentiating (5.43) with respect to a and setting the derivative $\partial f/\partial a$ equal to zero gives

$$-\frac{x}{a^2} + \frac{ay}{(L^2 - a^2)^{3/2}} = 0;$$

from which it follows that

$$a = \frac{Lx^{1/3}}{(x^{2/3} + y^{2/3})^{1/2}} \quad \text{and} \quad (L^2 - a^2)^{1/2} = \frac{Ly^{1/3}}{(x^{2/3} + y^{2/3})^{1/2}}.$$

Eliminating a by substituting these values into (5.43) gives, for the equation of the envelope of all possible positions on the ladder,

$$x^{2/3} + y^{2/3} = L^{2/3}.$$

This is the equation of an astroid (mentioned in exercise 2.19), and, together with the wall and the ground, marks the boundary of the vertical area below the ladder. ◄

Other examples, drawn from both geometry and and the physical sciences, are considered in the exercises at the end of this chapter. The shell trajectory problem discussed earlier in this section is solved there, but in the guise of a question about the water bell of an ornamental fountain.

5.11 Thermodynamic relations

Thermodynamic relations provide a useful set of physical examples of partial differentiation. The relations we will derive are called *Maxwell's thermodynamic relations*. They express relationships between four thermodynamic quantities describing a unit mass of a substance. The quantities are the pressure P, the volume V, the thermodynamic temperature T and the entropy S of the substance. These four quantities are not independent; any two of them can be varied independently, but the other two are then determined.

The first law of thermodynamics may be expressed as

$$dU = T\,dS - P\,dV, \tag{5.44}$$

where U is the internal energy of the substance. Essentially this is a conservation of energy equation, but we shall concern ourselves, not with the physics, but rather with the use of partial differentials to relate the four basic quantities discussed above. The method involves writing a total differential, dU say, in terms of the differentials of two variables, say X and Y, thus

$$dU = \left(\frac{\partial U}{\partial X}\right)_Y dX + \left(\frac{\partial U}{\partial Y}\right)_X dY, \tag{5.45}$$

and then using the relationship

$$\frac{\partial^2 U}{\partial X \partial Y} = \frac{\partial^2 U}{\partial Y \partial X}$$

to obtain the required Maxwell relation. The variables X and Y are to be chosen from P, V, T and S.

►*Show that* $(\partial T/\partial V)_S = -(\partial P/\partial S)_V$.

Here the two variables that have to be held constant, in turn, happen to be those whose differentials appear on the RHS of (5.44). And so, taking X as S and Y as V in (5.45), we have

$$T\,dS - P\,dV = dU = \left(\frac{\partial U}{\partial S}\right)_V dS + \left(\frac{\partial U}{\partial V}\right)_S dV,$$

and find directly that

$$\left(\frac{\partial U}{\partial S}\right)_V = T \qquad \text{and} \qquad \left(\frac{\partial U}{\partial V}\right)_S = -P.$$

Differentiating the first expression with respect to V and the second with respect to S, and using

$$\frac{\partial^2 U}{\partial V \partial S} = \frac{\partial^2 U}{\partial S \partial V},$$

we find the Maxwell relation

$$\left(\frac{\partial T}{\partial V}\right)_S = -\left(\frac{\partial P}{\partial S}\right)_V. \blacktriangleleft$$

►*Show that* $(\partial S/\partial V)_T = (\partial P/\partial T)_V$.

Applying (5.45) to dS, with independent variables V and T, we find

$$dU = T\,dS - P\,dV = T\left[\left(\frac{\partial S}{\partial V}\right)_T dV + \left(\frac{\partial S}{\partial T}\right)_V dT\right] - P dV.$$

Similarly applying (5.45) to dU, we find

$$dU = \left(\frac{\partial U}{\partial V}\right)_T dV + \left(\frac{\partial U}{\partial T}\right)_V dT.$$

Thus, equating partial derivatives,

$$\left(\frac{\partial U}{\partial V}\right)_T = T\left(\frac{\partial S}{\partial V}\right)_T - P \quad \text{and} \quad \left(\frac{\partial U}{\partial T}\right)_V = T\left(\frac{\partial S}{\partial T}\right)_V.$$

But, since

$$\frac{\partial^2 U}{\partial T \partial V} = \frac{\partial^2 U}{\partial V \partial T}, \qquad \text{i.e.} \qquad \frac{\partial}{\partial T}\left(\frac{\partial U}{\partial V}\right)_T = \frac{\partial}{\partial V}\left(\frac{\partial U}{\partial T}\right)_V,$$

it follows that

$$\left(\frac{\partial S}{\partial V}\right)_T + T\frac{\partial^2 S}{\partial T \partial V} - \left(\frac{\partial P}{\partial T}\right)_V = \frac{\partial}{\partial V}\left[T\left(\frac{\partial S}{\partial T}\right)_V\right]_T = T\frac{\partial^2 S}{\partial V \partial T}.$$

Thus finally we get the Maxwell relation

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial P}{\partial T}\right)_V. \blacktriangleleft$$

The above derivation is rather cumbersome, however, and a useful trick that can simplify the working is to define a new function, called a *potential*. The internal energy U discussed above is one example of a potential but three others are commonly defined and they are described below.

►*Show that* $(\partial S/\partial V)_T = (\partial P/\partial T)_V$ *by considering the potential* $U - ST$.

We first consider the differential $d(U - ST)$. From (5.5), we obtain

$$d(U - ST) = dU - SdT - TdS = -SdT - PdV$$

when use is made of (5.44). We rewrite $U - ST$ as F for convenience of notation; F is called the *Helmholtz potential*. Thus

$$dF = -SdT - PdV,$$

and it follows that

$$\left(\frac{\partial F}{\partial T}\right)_V = -S \quad \text{and} \quad \left(\frac{\partial F}{\partial V}\right)_T = -P.$$

Using these results together with

$$\frac{\partial^2 F}{\partial T \partial V} = \frac{\partial^2 F}{\partial V \partial T},$$

we can see immediately that

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial P}{\partial T}\right)_V,$$

which is the same Maxwell relation as before. ◄

Although the Helmholtz potential has other uses, in this context it has simply provided a means for a quick derivation of the Maxwell relation. The other Maxwell relations can be derived similarly by using two other potentials, the *enthalpy*, $H = U + PV$, and the *Gibbs free energy*, $G = U + PV - ST$ (see exercise 5.25).

5.12 Differentiation of integrals

We conclude this chapter with a discussion of the differentiation of integrals. Let us consider the indefinite integral (cf. equation (2.30))

$$F(x,t) = \int f(x,t)\,dt,$$

from which it follows immediately that

$$\frac{\partial F(x,t)}{\partial t} = f(x,t).$$

Assuming that the second partial derivatives of $F(x,t)$ are continuous, we have

$$\frac{\partial^2 F(x,t)}{\partial t \partial x} = \frac{\partial^2 F(x,t)}{\partial x \partial t},$$

and so we can write

$$\frac{\partial}{\partial t}\left[\frac{\partial F(x,t)}{\partial x}\right] = \frac{\partial}{\partial x}\left[\frac{\partial F(x,t)}{\partial t}\right] = \frac{\partial f(x,t)}{\partial x}.$$

Integrating this equation with respect to t then gives

$$\frac{\partial F(x,t)}{\partial x} = \int \frac{\partial f(x,t)}{\partial x}\,dt. \tag{5.46}$$

Now consider the definite integral

$$\begin{aligned} I(x) &= \int_{t=u}^{t=v} f(x,t)\,dt \\ &= F(x,v) - F(x,u), \end{aligned}$$

where u and v are constants. Differentiating this integral with respect to x, and using (5.46), we see that

$$\begin{aligned} \frac{dI(x)}{dx} &= \frac{\partial F(x,v)}{\partial x} - \frac{\partial F(x,u)}{\partial x} \\ &= \int^{v} \frac{\partial f(x,t)}{\partial x} dt - \int^{u} \frac{\partial f(x,t)}{\partial x} dt \\ &= \int_{u}^{v} \frac{\partial f(x,t)}{\partial x} dt. \end{aligned}$$

This is *Leibnitz' rule* for differentiating integrals, and basically it states that for constant limits of integration the order of integration and differentiation can be reversed.

In the more general case where the limits of the integral are themselves functions of x, it follows immediately that

$$\begin{aligned} I(x) &= \int_{t=u(x)}^{t=v(x)} f(x,t)\,dt \\ &= F(x,v(x)) - F(x,u(x)), \end{aligned}$$

which yields the partial derivatives

$$\frac{\partial I}{\partial v} = f(x,v(x)), \qquad \frac{\partial I}{\partial u} = -f(x,u(x)).$$

Consequently

$$\begin{aligned} \frac{dI}{dx} &= \left(\frac{\partial I}{\partial v}\right)\frac{dv}{dx} + \left(\frac{\partial I}{\partial u}\right)\frac{du}{dx} + \frac{\partial I}{\partial x} \\ &= f(x,v(x))\frac{dv}{dx} - f(x,u(x))\frac{du}{dx} + \frac{\partial}{\partial x}\int_{u(x)}^{v(x)} f(x,t)dt \\ &= f(x,v(x))\frac{dv}{dx} - f(x,u(x))\frac{du}{dx} + \int_{u(x)}^{v(x)} \frac{\partial f(x,t)}{\partial x}dt, \end{aligned} \tag{5.47}$$

where the partial derivative with respect to x in the last term has been taken inside the integral sign using (5.46). This procedure is valid because $u(x)$ and $v(x)$ are being held constant in this term.

►*Find the derivative with respect to x of the integral*

$$I(x) = \int_x^{x^2} \frac{\sin xt}{t}\,dt.$$

Applying (5.47), we see that

$$\begin{aligned} \frac{dI}{dx} &= \frac{\sin x^3}{x^2}(2x) - \frac{\sin x^2}{x}(1) + \int_x^{x^2} \frac{t\cos xt}{t}dt \\ &= \frac{2\sin x^3}{x} - \frac{\sin x^2}{x} + \left[\frac{\sin xt}{x}\right]_x^{x^2} \\ &= 3\frac{\sin x^3}{x} - 2\frac{\sin x^2}{x} \\ &= \frac{1}{x}(3\sin x^3 - 2\sin x^2). \blacktriangleleft \end{aligned}$$

5.13 Exercises

5.1 Using the appropriate properties of ordinary derivatives, perform the following.

(a) Find all the first partial derivatives of the following functions $f(x, y)$: (i) x^2y, (ii) $x^2 + y^2 + 4$, (iii) $\sin(x/y)$, (iv) $\tan^{-1}(y/x)$, (v) $r(x, y, z) = (x^2 + y^2 + z^2)^{1/2}$.
(b) For (i), (ii) and (v), find $\partial^2 f/\partial x^2$, $\partial^2 f/\partial y^2$ and $\partial^2 f/\partial x \partial y$.
(c) For (iv) verify that $\partial^2 f/\partial x \partial y = \partial^2 f/\partial y \partial x$.

5.2 Determine which of the following are exact differentials:

(a) $(3x + 2)y\,dx + x(x + 1)\,dy$;
(b) $y \tan x\,dx + x \tan y\,dy$;
(c) $y^2(\ln x + 1)\,dx + 2xy \ln x\,dy$;
(d) $y^2(\ln x + 1)\,dy + 2xy \ln x\,dx$;
(e) $[x/(x^2 + y^2)]\,dy - [y/(x^2 + y^2)]\,dx$.

5.3 Show that the differential

$$df = x^2\,dy - (y^2 + xy)\,dx$$

is not exact, but that $dg = (xy^2)^{-1}df$ is exact.

5.4 Show that

$$df = y(1 + x - x^2)\,dx + x(x + 1)\,dy$$

is not an exact differential.

Find the differential equation that a function $g(x)$ must satisfy if $d\phi = g(x)df$ is to be an exact differential. Verify that $g(x) = e^{-x}$ is a solution of this equation and deduce the form of $\phi(x, y)$.

5.5 The equation $3y = z^3 + 3xz$ defines z implicitly as a function of x and y. Evaluate all three second partial derivatives of z with respect to x and/or y. Verify that z is a solution of

$$x\frac{\partial^2 z}{\partial y^2} + \frac{\partial^2 z}{\partial x^2} = 0.$$

5.6 A possible equation of state for a gas takes the form

$$PV = RT \exp\left(-\frac{\alpha}{VRT}\right),$$

in which α and R are constants. Calculate expressions for

$$\left(\frac{\partial P}{\partial V}\right)_T, \qquad \left(\frac{\partial V}{\partial T}\right)_P, \qquad \left(\frac{\partial T}{\partial P}\right)_V,$$

and show that their product is -1, as stated in section 5.4.

5.7 The function $G(t)$ is defined by

$$G(t) = F(x, y) = x^2 + y^2 + 3xy,$$

where $x(t) = at^2$ and $y(t) = 2at$. Use the chain rule to find the values of (x, y) at which $G(t)$ has stationary values as a function of t. Do any of them correspond to the stationary points of $F(x, y)$ as a function of x and y?

5.8 In the xy-plane, new coordinates s and t are defined by

$$s = \tfrac{1}{2}(x + y), \qquad t = \tfrac{1}{2}(x - y).$$

Transform the equation

$$\frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial y^2} = 0$$

into the new coordinates and deduce that its general solution can be written

$$\phi(x, y) = f(x + y) + g(x - y),$$

where $f(u)$ and $g(v)$ are arbitrary functions of u and v, respectively.

5.9 The function $f(x, y)$ satisfies the differential equation

$$y\frac{\partial f}{\partial x} + x\frac{\partial f}{\partial y} = 0.$$

By changing to new variables $u = x^2 - y^2$ and $v = 2xy$, show that f is, in fact, a function of $x^2 - y^2$ only.

5.10 If $x = e^u \cos\theta$ and $y = e^u \sin\theta$, show that

$$\frac{\partial^2 \phi}{\partial u^2} + \frac{\partial^2 \phi}{\partial \theta^2} = (x^2 + y^2)\left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\right),$$

where $f(x, y) = \phi(u, \theta)$.

5.11 Find and evaluate the maxima, minima and saddle points of the function

$$f(x, y) = xy(x^2 + y^2 - 1).$$

5.12 Show that

$$f(x, y) = x^3 - 12xy + 48x + by^2, \qquad b \neq 0,$$

has two, one, or zero stationary points, according to whether $|b|$ is less than, equal to, or greater than 3.

5.13 Locate the stationary points of the function

$$f(x, y) = (x^2 - 2y^2)\exp[-(x^2 + y^2)/a^2],$$

where a is a non-zero constant.

Sketch the function along the x- and y-axes and hence identify the nature and values of the stationary points.

5.14 Find the stationary points of the function

$$f(x, y) = x^3 + xy^2 - 12x - y^2$$

and identify their natures.

5.15 Find the stationary values of

$$f(x, y) = 4x^2 + 4y^2 + x^4 - 6x^2y^2 + y^4$$

and classify them as maxima, minima or saddle points. Make a rough sketch of the contours of f in the quarter plane $x, y \geq 0$.

5.16 The temperature of a point (x, y, z) on the unit sphere is given by

$$T(x, y, z) = 1 + xy + yz.$$

By using the method of Lagrange multipliers, find the temperature of the hottest point on the sphere.

5.17 A rectangular parallelepiped has all eight vertices on the ellipsoid

$$x^2 + 3y^2 + 3z^2 = 1.$$

Using the symmetry of the parallelepiped about each of the planes $x = 0$, $y = 0$, $z = 0$, write down the surface area of the parallelepiped in terms of the coordinates of the vertex that lies in the octant $x, y, z \geq 0$. Hence find the maximum value of the surface area of such a parallelepiped.

5.18 Two horizontal corridors, $0 \leq x \leq a$ with $y \geq 0$, and $0 \leq y \leq b$ with $x \geq 0$, meet at right angles. Find the length L of the longest ladder (considered as a stick) that may be carried horizontally around the corner.

5.19 A barn is to be constructed with a uniform cross-sectional area A throughout its length. The cross-section is to be a rectangle of wall height h (fixed) and width w, surmounted by an isosceles triangular roof that makes an angle θ with the horizontal. The cost of construction is α per unit height of wall and β per unit (slope) length of roof. Show that, irrespective of the values of α and β, to minimise costs w should be chosen to satisfy the equation

$$w^4 = 16A(A - wh),$$

and θ made such that $2\tan 2\theta = w/h$.

5.20 Show that the envelope of all concentric ellipses that have their axes along the x- and y-coordinate axes, and that have the sum of their semi-axes equal to a constant L, is the same curve (an astroid) as that found in the worked example in section 5.10.

5.21 Find the area of the region covered by points on the lines

$$\frac{x}{a} + \frac{y}{b} = 1,$$

where the sum of any line's intercepts on the coordinate axes is fixed and equal to c.

5.22 Prove that the envelope of the circles whose diameters are those chords of a given circle that pass through a fixed point on its circumference, is the cardioid

$$r = a(1 + \cos\theta).$$

Here a is the radius of the given circle and (r, θ) are the polar coordinates of the envelope. Take as the system parameter the angle ϕ between a chord and the polar axis from which θ is measured.

5.23 A water feature contains a spray head at water level at the centre of a round basin. The head is in the form of a small hemisphere perforated by many evenly distributed small holes, through which water spurts out at the same speed, v_0, in all directions.

(a) What is the shape of the 'water bell' so formed?

(b) What must be the minimum diameter of the bowl if no water is to be lost?

5.24 In order to make a focussing mirror that concentrates parallel axial rays to one spot (or conversely forms a parallel beam from a point source), a parabolic shape should be adopted. If a mirror that is part of a circular cylinder or sphere were used, the light would be spread out along a curve. This curve is known as a *caustic* and is the envelope of the rays reflected from the mirror. Denoting by θ the angle which a typical incident axial ray makes with the normal to the mirror at the place where it is reflected, the geometry of reflection (the angle of incidence equals the angle of reflection) is shown in figure 5.5.

Show that a parametric specification of the caustic is

$$x = R\cos\theta\left(\tfrac{1}{2} + \sin^2\theta\right), \qquad y = R\sin^3\theta,$$

where R is the radius of curvature of the mirror. The curve is, in fact, part of an epicycloid.

5.25 By considering the differential

$$dG = d(U + PV - ST),$$

where G is the Gibbs free energy, P the pressure, V the volume, S the entropy and T the temperature of a system, and given further that the internal energy U satisfies

$$dU = T\,dS - P\,dV,$$

derive a Maxwell relation connecting $(\partial V/\partial T)_P$ and $(\partial S/\partial P)_T$.

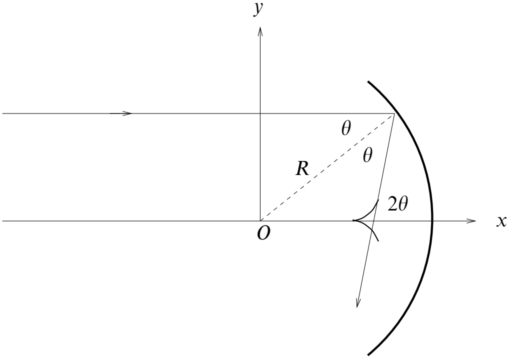


Figure 5.5 The reflecting mirror discussed in exercise 5.24.

5.26 Functions $P(V, T)$, $U(V, T)$ and $S(V, T)$ are related by

$$T\,dS = dU + P\,dV,$$

where the symbols have the same meaning as in the previous question. The pressure P is known from experiment to have the form

$$P = \frac{T^4}{3} + \frac{T}{V},$$

in appropriate units. If

$$U = \alpha V T^4 + \beta T,$$

where α, β, are constants (or, at least, do not depend on T or V), deduce that α must have a specific value, but that β may have any value. Find the corresponding form of S.

5.27 As in the previous two exercises on the thermodynamics of a simple gas, the quantity $dS = T^{-1}(dU + P\,dV)$ is an exact differential. Use this to prove that

$$\left(\frac{\partial U}{\partial V}\right)_T = T\left(\frac{\partial P}{\partial T}\right)_V - P.$$

In the van der Waals model of a gas, P obeys the equation

$$P = \frac{RT}{V-b} - \frac{a}{V^2},$$

where R, a and b are constants. Further, in the limit $V \to \infty$, the form of U becomes $U = cT$, where c is another constant. Find the complete expression for $U(V, T)$.

5.28 The entropy $S(H, T)$, the magnetisation $M(H, T)$ and the internal energy $U(H, T)$ of a magnetic salt placed in a magnetic field of strength H, at temperature T, are connected by the equation

$$T\,dS = dU - H\,dM.$$

By considering $d(U - TS - HM)$ prove that

$$\left(\frac{\partial M}{\partial T}\right)_H = \left(\frac{\partial S}{\partial H}\right)_T.$$

For a particular salt,

$$M(H,T) = M_0[1 - \exp(-\alpha H/T)].$$

Show that if, at a fixed temperature, the applied field is increased from zero to a strength such that the magnetization of the salt is $\frac{3}{4}M_0$, then the salt's entropy *decreases* by an amount

$$\frac{M_0}{4\alpha}(3 - \ln 4).$$

5.29 Using the results of section 5.12, evaluate the integral

$$I(y) = \int_0^\infty \frac{e^{-xy}\sin x}{x}\,dx.$$

Hence show that

$$J = \int_0^\infty \frac{\sin x}{x}\,dx = \frac{\pi}{2}.$$

5.30 The integral

$$\int_{-\infty}^{\infty} e^{-\alpha x^2}\,dx$$

has the value $(\pi/\alpha)^{1/2}$. Use this result to evaluate

$$J(n) = \int_{-\infty}^{\infty} x^{2n} e^{-x^2}\,dx,$$

where n is a positive integer. Express your answer in terms of factorials.

5.31 The function $f(x)$ is differentiable and $f(0) = 0$. A second function $g(y)$ is defined by

$$g(y) = \int_0^y \frac{f(x)\,dx}{\sqrt{y-x}}.$$

Prove that

$$\frac{dg}{dy} = \int_0^y \frac{df}{dx}\frac{dx}{\sqrt{y-x}}.$$

For the case $f(x) = x^n$, prove that

$$\frac{d^n g}{dy^n} = 2(n!)\sqrt{y}.$$

5.32 The functions $f(x,t)$ and $F(x)$ are defined by

$$f(x,t) = e^{-xt},$$
$$F(x) = \int_0^x f(x,t)\,dt.$$

Verify, by explicit calculation, that

$$\frac{dF}{dx} = f(x,x) + \int_0^x \frac{\partial f(x,t)}{\partial x}\,dt.$$

5.33 If

$$I(\alpha) = \int_0^1 \frac{x^\alpha - 1}{\ln x}\,dx, \qquad \alpha > -1,$$

what is the value of $I(0)$? Show that

$$\frac{d}{d\alpha}x^\alpha = x^\alpha \ln x,$$

and deduce that

$$\frac{d}{d\alpha}I(\alpha) = \frac{1}{\alpha + 1}.$$

Hence prove that $I(\alpha) = \ln(1+\alpha)$.

5.34 Find the derivative, with respect to x, of the integral

$$I(x) = \int_x^{3x} \exp xt\, dt.$$

5.35 The function $G(t,\xi)$ is defined for $0 \le t \le \pi$ by

$$G(t,\xi) = \begin{cases} -\cos t \sin \xi & \text{for } \xi \le t, \\ -\sin t \cos \xi & \text{for } \xi > t. \end{cases}$$

Show that the function $x(t)$ defined by

$$x(t) = \int_0^\pi G(t,\xi) f(\xi)\, d\xi$$

satisfies the equation

$$\frac{d^2x}{dt^2} + x = f(t),$$

where $f(t)$ can be *any* arbitrary (continuous) function. Show further that $x(0) = [dx/dt]_{t=\pi} = 0$, again for any $f(t)$, but that the *value* of $x(\pi)$ does depend upon the form of $f(t)$.

[The function $G(t,\xi)$ is an example of a Green's function, an important concept in the solution of differential equations and one studied extensively in later chapters.]

5.14 Hints and answers

5.1 (a) (i) $2xy, x^2$; (ii) $2x, 2y$; (iii) $y^{-1}\cos(x/y), (-x/y^2)\cos(x/y)$; (iv) $-y/(x^2+y^2), x/(x^2+y^2)$; (v) $x/r, y/r, z/r$.
(b) (i) $2y, 0, 2x$; (ii) $2, 2, 0$; (v) $(y^2+z^2)r^{-3}, (x^2+z^2)r^{-3}, -xyr^{-3}$.
(c) Both second derivatives are equal to $(y^2-x^2)(x^2+y^2)^{-2}$.

5.3 $2x \neq -2y - x$. For g, both sides of equation (5.9) equal y^{-2}.

5.5 $\partial^2 z/\partial x^2 = 2xz(z^2+x)^{-3}$, $\partial^2 z/\partial x \partial y = (z^2-x)(z^2+x)^{-3}$, $\partial^2 z/\partial y^2 = -2z(z^2+x)^{-3}$.

5.7 $(0,0)$, $(a/4,-a)$ and $(16a,-8a)$. Only the saddle point at $(0,0)$.

5.9 The transformed equation is $2(x^2+y^2)\partial f/\partial v = 0$; hence f does not depend on v.

5.11 Maxima, equal to $1/8$, at $\pm(1/2,-1/2)$, minima, equal to $-1/8$, at $\pm(1/2,1/2)$, saddle points, equalling 0, at $(0,0)$, $(0,\pm1)$, $(\pm1,0)$.

5.13 Maxima equal to a^2e^{-1} at $(\pm a, 0)$, minima equal to $-2a^2e^{-1}$ at $(0,\pm a)$, saddle point equalling 0 at $(0,0)$.

5.15 Minimum at $(0,0)$; saddle points at $(\pm1,\pm1)$. To help with sketching the contours, determine the behaviour of $g(x) = f(x,x)$.

5.17 The Lagrange multiplier method gives $z = y = x/2$, for a maximal area of 4.

5.19 The cost always includes $2\alpha h$, which can therefore be ignored in the optimisation. With Lagrange multiplier λ, $\sin\theta = \lambda w/(4\beta)$ and $\beta\sec\theta - \frac{1}{2}\lambda w\tan\theta = \lambda h$, leading to the stated results.

5.21 The envelope of the lines $x/a + y/(c-a) - 1 = 0$, as a is varied, is $\sqrt{x} + \sqrt{y} = \sqrt{c}$. Area $= c^2/6$.

5.23 (a) Using $\alpha = \cot\theta$, where θ is the initial angle a jet makes with the vertical, the equation is $f(z,\rho,\alpha) = z - \rho\alpha + [g\rho^2(1+\alpha^2)/(2v_0^2)]$, and setting $\partial f/\partial\alpha = 0$ gives $\alpha = v_0^2/(g\rho)$. The water bell has a parabolic profile $z = v_0^2/(2g) - g\rho^2/(2v_0^2)$.

(b) Setting $z = 0$ gives the minimum diameter as $2v_0^2/g$.

5.25 Show that $(\partial G/\partial P)_T = V$ and $(\partial G/\partial T)_P = -S$. From each result, obtain an expression for $\partial^2 G/\partial T\partial P$ and equate these, giving $(\partial V/\partial T)_P = -(\partial S/\partial P)_T$.

5.27 Find expressions for $(\partial S/\partial V)_T$ and $(\partial S/\partial T)_V$, and equate $\partial^2 S/\partial V\partial T$ with $\partial^2 S/\partial T\partial V$. $U(V,T) = cT - aV^{-1}$.

5.29 $dI/dy = -\mathrm{Im}[\int_0^\infty \exp(-xy + ix)\,dx] = -1/(1+y^2)$. Integrate dI/dy from 0 to ∞. $I(\infty) = 0$ and $I(0) = J$.

5.31 Integrate the RHS of the equation by parts, before differentiating with respect to y. Repeated application of the method establishes the result for all orders of derivative.

5.33 $I(0) = 0$; use Leibnitz' rule.

5.35 Write $x(t) = -\cos t\int_0^t \sin\xi\, f(\xi)\,d\xi - \sin t\int_t^\pi \cos\xi\, f(\xi)\,d\xi$ and differentiate each term as a product to obtain dx/dt. Obtain d^2x/dt^2 in a similar way. Note that integrals that have equal lower and upper limits have value zero. The value of $x(\pi)$ is $\int_0^\pi \sin\xi\, f(\xi)\,d\xi$.

6

Multiple integrals

For functions of several variables, just as we may consider derivatives with respect to two or more of them, so may the integral of the function with respect to more than one variable be formed. The formal definitions of such multiple integrals are extensions of that for a single variable, discussed in chapter 2. We first discuss double and triple integrals and illustrate some of their applications. We then consider changing the variables in multiple integrals and discuss some general properties of Jacobians.

6.1 Double integrals

For an integral involving two variables – a double integral – we have a function, $f(x, y)$ say, to be integrated with respect to x and y between certain limits. These limits can usually be represented by a closed curve C bounding a region R in the xy-plane. Following the discussion of single integrals given in chapter 2, let us divide the region R into N subregions ΔR_p of area ΔA_p, $p = 1, 2, \ldots, N$, and let (x_p, y_p) be any point in subregion ΔR_p. Now consider the sum

$$S = \sum_{p=1}^{N} f(x_p, y_p)\Delta A_p,$$

and let $N \to \infty$ as each of the areas $\Delta A_p \to 0$. If the sum S tends to a unique limit, I, then this is called the *double integral of $f(x, y)$ over the region R* and is written

$$I = \int_R f(x, y)\, dA, \tag{6.1}$$

where dA stands for the element of area in the xy-plane. By choosing the subregions to be small rectangles each of area $\Delta A = \Delta x \Delta y$, and letting both Δx

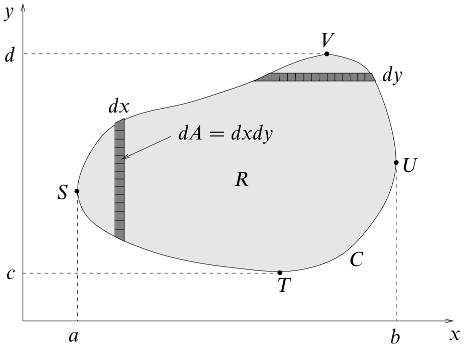


Figure 6.1 A simple curve C in the xy-plane, enclosing a region R.

and $\Delta y \to 0$, we can also write the integral as

$$I = \iint_R f(x, y)\, dx\, dy, \tag{6.2}$$

where we have written out the element of area explicitly as the product of the two coordinate differentials (see figure 6.1).

Some authors use a single integration symbol whatever the dimension of the integral; others use as many symbols as the dimension. In different circumstances both have their advantages. We will adopt the convention used in (6.1) and (6.2), that as many integration symbols will be used as differentials *explicitly* written.

The form (6.2) gives us a clue as to how we may proceed in the evaluation of a double integral. Referring to figure 6.1, the limits on the integration may be written as an equation $c(x, y) = 0$ giving the boundary curve C. However, an explicit statement of the limits can be written in two distinct ways.

One way of evaluating the integral is first to sum up the contributions from the small rectangular elemental areas in a horizontal strip of width dy (as shown in the figure) and then to combine the contributions of these horizontal strips to cover the region R. In this case, we write

$$I = \int_{y=c}^{y=d} \left\{ \int_{x=x_1(y)}^{x=x_2(y)} f(x, y)\, dx \right\} dy, \tag{6.3}$$

where $x = x_1(y)$ and $x = x_2(y)$ are the equations of the curves TSV and TUV respectively. This expression indicates that first $f(x, y)$ is to be integrated with respect to x (treating y as a constant) between the values $x = x_1(y)$ and $x = x_2(y)$ and then the result, considered as a function of y, is to be integrated between the limits $y = c$ and $y = d$. Thus the double integral is evaluated by expressing it in terms of two single integrals called *iterated* (or *repeated*) integrals.

An alternative way of evaluating the integral, however, is first to sum up the contributions from the elemental rectangles arranged into *vertical* strips and then to combine these vertical strips to cover the region R. We then write

$$I = \int_{x=a}^{x=b} \left\{ \int_{y=y_1(x)}^{y=y_2(x)} f(x,y)\,dy \right\} dx, \tag{6.4}$$

where $y = y_1(x)$ and $y = y_2(x)$ are the equations of the curves STU and SVU respectively. In going to (6.4) from (6.3), we have essentially interchanged the order of integration.

In the discussion above we assumed that the curve C was such that any line parallel to either the x- or y-axis intersected C at most twice. In general, provided $f(x,y)$ is continuous everywhere in R and the boundary curve C has this simple shape, the same result is obtained irrespective of the order of integration. In cases where the region R has a more complicated shape, it can usually be subdivided into smaller simpler regions R_1, R_2 etc. that satisfy this criterion. The double integral over R is then merely the sum of the double integrals over the subregions.

►*Evaluate the double integral*

$$I = \iint_R x^2 y\,dx\,dy,$$

where R is the triangular area bounded by the lines $x = 0$, $y = 0$ *and* $x + y = 1$. *Reverse the order of integration and demonstrate that the same result is obtained.*

The area of integration is shown in figure 6.2. Suppose we choose to carry out the integration with respect to y first. With x fixed, the range of y is 0 to $1 - x$. We can therefore write

$$\begin{aligned} I &= \int_{x=0}^{x=1} \left\{ \int_{y=0}^{y=1-x} x^2 y\,dy \right\} dx \\ &= \int_{x=0}^{x=1} \left[\frac{x^2 y^2}{2} \right]_{y=0}^{y=1-x} dx = \int_0^1 \frac{x^2(1-x)^2}{2}\,dx = \frac{1}{60}. \end{aligned}$$

Alternatively, we may choose to perform the integration with respect to x first. With y fixed, the range of x is 0 to $1 - y$, so we have

$$\begin{aligned} I &= \int_{y=0}^{y=1} \left\{ \int_{x=0}^{x=1-y} x^2 y\,dx \right\} dy \\ &= \int_{y=0}^{y=1} \left[\frac{x^3 y}{3} \right]_{x=0}^{x=1-y} dx = \int_0^1 \frac{(1-y)^3 y}{3}\,dy = \frac{1}{60}. \end{aligned}$$

As expected, we obtain the same result irrespective of the order of integration. ◄

We may avoid the use of braces in expressions such as (6.3) and (6.4) by writing (6.4), for example, as

$$I = \int_a^b dx \int_{y_1(x)}^{y_2(x)} dy\, f(x,y),$$

where it is understood that each integral symbol acts on everything to its right,

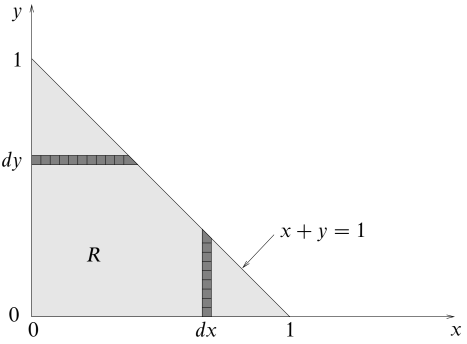


Figure 6.2 The triangular region whose sides are the axes $x = 0$, $y = 0$ and the line $x + y = 1$.

and that the order of integration is from right to left. So, in this example, the integrand $f(x, y)$ is first to be integrated with respect to y and then with respect to x. With the double integral expressed in this way, we will no longer write the independent variables explicitly in the limits of integration, since the differential of the variable with respect to which we are integrating is always adjacent to the relevant integral sign.

Using the order of integration in (6.3), we could also write the double integral as

$$I = \int_c^d dy \int_{x_1(y)}^{x_2(y)} dx\, f(x, y).$$

Occasionally, however, interchange of the order of integration in a double integral is not permissible, as it yields a different result. For example, difficulties might arise if the region R were unbounded with some of the limits infinite, though in many cases involving infinite limits the same result is obtained whichever order of integration is used. Difficulties can also occur if the integrand $f(x, y)$ has any discontinuities in the region R or on its boundary C.

6.2 Triple integrals

The above discussion for double integrals can easily be extended to triple integrals. Consider the function $f(x, y, z)$ defined in a closed three-dimensional region R. Proceeding as we did for double integrals, let us divide the region R into N subregions ΔR_p of volume ΔV_p, $p = 1, 2, \ldots, N$, and let (x_p, y_p, z_p) be any point in the subregion ΔR_p. Now we form the sum

$$S = \sum_{p=1}^{N} f(x_p, y_p, z_p)\Delta V_p,$$

and let $N \to \infty$ as each of the volumes $\Delta V_p \to 0$. If the sum S tends to a unique limit, I, then this is called the *triple integral of* $f(x,y,z)$ *over the region* R and is written

$$I = \int_R f(x,y,z)\, dV, \tag{6.5}$$

where dV stands for the element of volume. By choosing the subregions to be small cuboids, each of volume $\Delta V = \Delta x \Delta y \Delta z$, and proceeding to the limit, we can also write the integral as

$$I = \iiint_R f(x,y,z)\, dx\, dy\, dz, \tag{6.6}$$

where we have written out the element of volume explicitly as the product of the three coordinate differentials. Extending the notation used for double integrals, we may write triple integrals as three iterated integrals, for example,

$$I = \int_{x_1}^{x_2} dx \int_{y_1(x)}^{y_2(x)} dy \int_{z_1(x,y)}^{z_2(x,y)} dz\, f(x,y,z),$$

where the limits on each of the integrals describe the values that x, y and z take on the boundary of the region R. As for double integrals, in most cases the order of integration does not affect the value of the integral.

We can extend these ideas to define multiple integrals of higher dimensionality in a similar way.

6.3 Applications of multiple integrals

Multiple integrals have many uses in the physical sciences, since there are numerous physical quantities which can be written in terms of them. We now discuss a few of the more common examples.

6.3.1 Areas and volumes

Multiple integrals are often used in finding areas and volumes. For example, the integral

$$A = \int_R dA = \iint_R dx\, dy$$

is simply equal to the area of the region R. Similarly, if we consider the surface $z = f(x,y)$ in three-dimensional Cartesian coordinates then the volume under this surface that stands vertically above the region R is given by the integral

$$V = \int_R z\, dA = \iint_R f(x,y)\, dx\, dy,$$

where volumes above the xy-plane are counted as positive, and those below as negative.

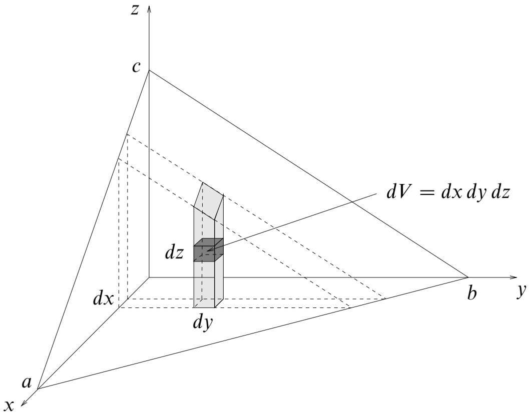


Figure 6.3 The tetrahedron bounded by the coordinate surfaces and the plane $x/a + y/b + z/c = 1$ is divided up into vertical slabs, the slabs into columns and the columns into small boxes.

►*Find the volume of the tetrahedron bounded by the three coordinate surfaces* $x = 0$*,* $y = 0$ *and* $z = 0$ *and the plane* $x/a + y/b + z/c = 1$.

Referring to figure 6.3, the elemental volume of the shaded region is given by $dV = z\,dx\,dy$, and we must integrate over the triangular region R in the xy-plane whose sides are $x = 0$, $y = 0$ and $y = b - bx/a$. The total volume of the tetrahedron is therefore given by

$$\begin{aligned} V = \iint_R z\,dx\,dy &= \int_0^a dx \int_0^{b-bx/a} dy\, c\left(1 - \frac{y}{b} - \frac{x}{a}\right) \\ &= c\int_0^a dx \left[y - \frac{y^2}{2b} - \frac{xy}{a}\right]_{y=0}^{y=b-bx/a} \\ &= c\int_0^a dx \left(\frac{bx^2}{2a^2} - \frac{bx}{a} + \frac{b}{2}\right) = \frac{abc}{6}. \end{aligned}$$ ◄

Alternatively, we can write the volume of a three-dimensional region R as

$$V = \int_R dV = \iiint_R dx\,dy\,dz, \tag{6.7}$$

where the only difficulty occurs in setting the correct limits on each of the integrals. For the above example, writing the volume in this way corresponds to dividing the tetrahedron into elemental boxes of volume $dx\,dy\,dz$ (as shown in figure 6.3); integration over z then adds up the boxes to form the shaded column in the figure. The limits of integration are $z = 0$ to $z = c\,(1 - y/b - x/a)$, and the total volume of the tetrahedron is given by

$$V = \int_0^a dx \int_0^{b-bx/a} dy \int_0^{c(1-y/b-x/a)} dz, \tag{6.8}$$

which clearly gives the same result as above. This method is illustrated further in the following example.

►*Find the volume of the region bounded by the paraboloid $z = x^2 + y^2$ and the plane $z = 2y$.*

The required region is shown in figure 6.4. In order to write the volume of the region in the form (6.7), we must deduce the limits on each of the integrals. Since the integrations can be performed in any order, let us first divide the region into vertical slabs of thickness dy perpendicular to the y-axis, and then as shown in the figure we cut each slab into horizontal strips of height dz, and each strip into elemental boxes of volume $dV = dx\,dy\,dz$. Integrating first with respect to x (adding up the elemental boxes to get a horizontal strip), the limits on x are $x = -\sqrt{z-y^2}$ to $x = \sqrt{z-y^2}$. Now integrating with respect to z (adding up the strips to form a vertical slab) the limits on z are $z = y^2$ to $z = 2y$. Finally, integrating with respect to y (adding up the slabs to obtain the required region), the limits on y are $y = 0$ and $y = 2$, the solutions of the simultaneous equations $z = 0^2 + y^2$ and $z = 2y$. So the volume of the region is

$$V = \int_0^2 dy \int_{y^2}^{2y} dz \int_{-\sqrt{z-y^2}}^{\sqrt{z-y^2}} dx = \int_0^2 dy \int_{y^2}^{2y} dz\, 2\sqrt{z-y^2}$$

$$= \int_0^2 dy\, \left[\tfrac{4}{3}(z-y^2)^{3/2}\right]_{z=y^2}^{z=2y} = \int_0^2 dy\, \tfrac{4}{3}(2y-y^2)^{3/2}.$$

The integral over y may be evaluated straightforwardly by making the substitution $y = 1 + \sin u$, and gives $V = \pi/2$. ◄

In general, when calculating the volume (area) of a region, the volume (area) elements need not be small boxes as in the previous example, but may be of any convenient shape. The latter is usually chosen to make evaluation of the integral as simple as possible.

6.3.2 Masses, centres of mass and centroids

It is sometimes necessary to calculate the mass of a given object having a non-uniform density. Symbolically, this mass is given simply by

$$M = \int dM,$$

where dM is the element of mass and the integral is taken over the extent of the object. For a solid three-dimensional body the element of mass is just $dM = \rho\, dV$, where dV is an element of volume and ρ is the variable density. For a laminar body (i.e. a uniform sheet of material) the element of mass is $dM = \sigma\, dA$, where σ is the mass per unit area of the body and dA is an area element. Finally, for a body in the form of a thin wire we have $dM = \lambda\, ds$, where λ is the mass per

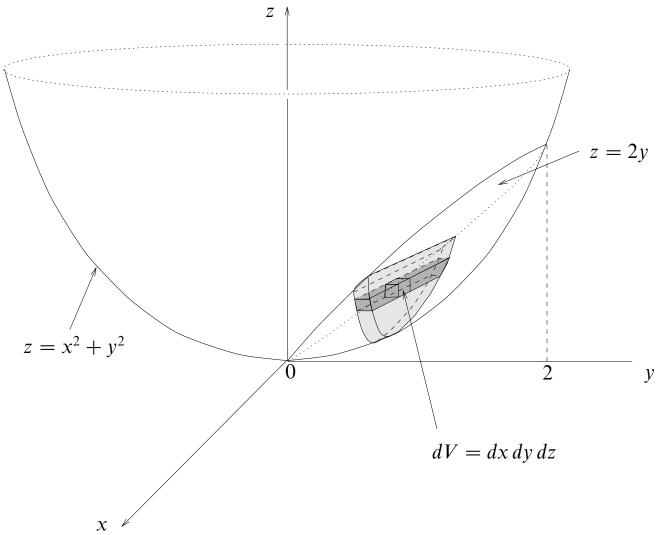


Figure 6.4 The region bounded by the paraboloid $z = x^2 + y^2$ and the plane $z = 2y$ is divided into vertical slabs, the slabs into horizontal strips and the strips into boxes.

unit length and ds is an element of arc length along the wire. When evaluating the required integral, we are free to divide up the body into mass elements in the most convenient way, provided that over each mass element the density is approximately constant.

► *Find the mass of the tetrahedron bounded by the three coordinate surfaces and the plane* $x/a + y/b + z/c = 1$, *if its density is given by* $\rho(x, y, z) = \rho_0(1 + x/a)$.

From (6.8), we can immediately write down the mass of the tetrahedron as

$$M = \int_R \rho_0 \left(1 + \frac{x}{a}\right)\, dV = \int_0^a dx\, \rho_0 \left(1 + \frac{x}{a}\right) \int_0^{b-bx/a} dy \int_0^{c\left(1-y/b-x/a\right)} dz,$$

where we have taken the density outside the integrations with respect to z and y since it depends only on x. Therefore the integrations with respect to z and y proceed exactly as they did when finding the volume of the tetrahedron, and we have

$$M = c\rho_0 \int_0^a dx\, \left(1 + \frac{x}{a}\right) \left(\frac{bx^2}{2a^2} - \frac{bx}{a} + \frac{b}{2}\right). \tag{6.9}$$

We could have arrived at (6.9) more directly by dividing the tetrahedron into triangular slabs of thickness dx perpendicular to the x-axis (see figure 6.3), each of which is of constant density, since ρ depends on x alone. A slab at a position x has volume $dV = \frac{1}{2}c(1 - x/a)(b - bx/a)\, dx$ and mass $dM = \rho\, dV = \rho_0(1 + x/a)\, dV$. Integrating over x we again obtain (6.9). This integral is easily evaluated and gives $M = \frac{5}{24}abc\rho_0$. ◄

The coordinates of the centre of mass of a solid or laminar body may also be written as multiple integrals. The centre of mass of a body has coordinates $\bar{x}$, $\bar{y}$, $\bar{z}$ given by the three equations

$$\bar{x}\int dM = \int x\,dM$$

$$\bar{y}\int dM = \int y\,dM$$

$$\bar{z}\int dM = \int z\,dM,$$

where again dM is an element of mass as described above, x, y, z are the coordinates of the centre of mass of the element dM and the integrals are taken over the entire body. Obviously, for any body that lies entirely in, or is symmetrical about, the xy-plane (say), we immediately have $\bar{z} = 0$. For completeness, we note that the three equations above can be written as the single vector equation (see chapter 7)

$$\bar{\mathbf{r}} = \frac{1}{M}\int \mathbf{r}\,dM,$$

where $\bar{\mathbf{r}}$ is the position vector of the body's centre of mass with respect to the origin, $\mathbf{r}$ is the position vector of the centre of mass of the element dM and $M = \int dM$ is the total mass of the body. As previously, we may divide the body into the most convenient mass elements for evaluating the necessary integrals, provided each mass element is of constant density.

We further note that the coordinates of the *centroid* of a body are defined as those of its centre of mass if the body had uniform density.

►*Find the centre of mass of the solid hemisphere bounded by the surfaces $x^2 + y^2 + z^2 = a^2$ and the xy-plane, assuming that it has a uniform density ρ.*

Referring to figure 6.5, we know from symmetry that the centre of mass must lie on the z-axis. Let us divide the hemisphere into volume elements that are circular slabs of thickness dz parallel to the xy-plane. For a slab at a height z, the mass of the element is $dM = \rho\,dV = \rho\pi(a^2 - z^2)\,dz$. Integrating over z, we find that the z-coordinate of the centre of mass of the hemisphere is given by

$$\bar{z}\int_0^a \rho\pi(a^2 - z^2)\,dz = \int_0^a z\rho\pi(a^2 - z^2)\,dz.$$

The integrals are easily evaluated and give $\bar{z} = 3a/8$. Since the hemisphere is of uniform density, this is also the position of its centroid. ◄

6.3.3 Pappus' theorems

The theorems of Pappus (which are about seventeen centuries old) relate centroids to volumes of revolution and areas of surfaces, discussed in chapter 2, and may be useful for finding one quantity given another that can be calculated more easily.

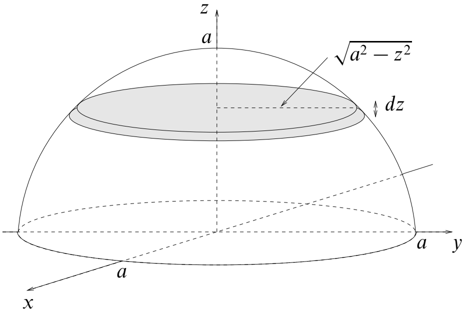


Figure 6.5 The solid hemisphere bounded by the surfaces $x^2 + y^2 + z^2 = a^2$ and the xy-plane.

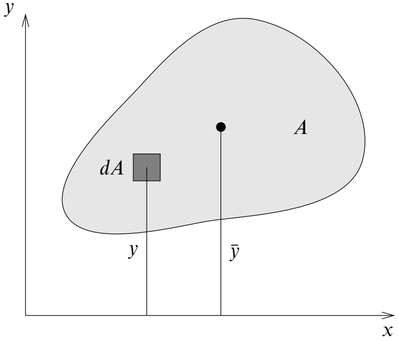


Figure 6.6 An area A in the xy-plane, which may be rotated about the x-axis to form a volume of revolution.

If a plane area is rotated about an axis that does not intersect it then the solid so generated is called a *volume of revolution. Pappus' first theorem* states that the volume of such a solid is given by the plane area A multiplied by the distance moved by its centroid (see figure 6.6). This may be proved by considering the definition of the centroid of the plane area as the position of the centre of mass if the density is uniform, so that

$$\bar{y} = \frac{1}{A}\int y\,dA.$$

Now the volume generated by rotating the plane area about the x-axis is given by

$$V = \int 2\pi y\,dA = 2\pi\bar{y}A,$$

which is the area multiplied by the distance moved by the centroid.

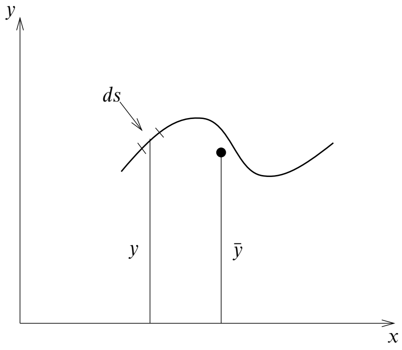


Figure 6.7 A curve in the xy-plane, which may be rotated about the x-axis to form a surface of revolution.

Pappus' second theorem states that if a plane curve is rotated about a coplanar axis that does not intersect it then the area of the *surface of revolution* so generated is given by the length of the curve L multiplied by the distance moved by its centroid (see figure 6.7). This may be proved in a similar manner to the first theorem by considering the definition of the centroid of a plane curve,

$$\bar{y} = \frac{1}{L}\int y\, ds,$$

and noting that the surface area generated is given by

$$S = \int 2\pi y\, ds = 2\pi \bar{y} L,$$

which is equal to the length of the curve multiplied by the distance moved by its centroid.

► *A semicircular uniform lamina is freely suspended from one of its corners. Show that its straight edge makes an angle of* 23.0° *with the vertical.*

Referring to figure 6.8, the suspended lamina will have its centre of gravity C vertically below the suspension point and its straight edge will make an angle $\theta = \tan^{-1}(d/a)$ with the vertical, where $2a$ is the diameter of the semicircle and d is the distance of its centre of mass from the diameter.

Since rotating the lamina about the diameter generates a sphere of volume $\frac{4}{3}\pi a^3$, Pappus' first theorem requires that

$$\tfrac{4}{3}\pi a^3 = 2\pi d \times \tfrac{1}{2}\pi a^2.$$

Hence $d = \frac{4a}{3\pi}$ and $\theta = \tan^{-1}\frac{4}{3\pi} = 23.0°$. ◄

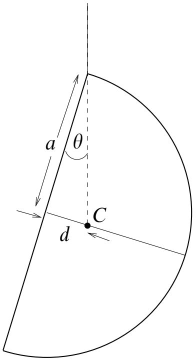


Figure 6.8 Suspending a semicircular lamina from one of its corners.

6.3.4 Moments of inertia

For problems in rotational mechanics it is often necessary to calculate the moment of inertia of a body about a given axis. This is defined by the multiple integral

$$I = \int l^2 \, dM,$$

where l is the distance of a mass element dM from the axis. We may again choose mass elements convenient for evaluating the integral. In this case, however, in addition to elements of constant density we require all parts of each element to be at approximately the same distance from the axis about which the moment of inertia is required.

► *Find the moment of inertia of a uniform rectangular lamina of mass M with sides a and b about one of the sides of length b.*

Referring to figure 6.9, we wish to calculate the moment of inertia about the y-axis. We therefore divide the rectangular lamina into elemental strips parallel to the y-axis of width dx. The mass of such a strip is $dM = \sigma b \, dx$, where σ is the mass per unit area of the lamina. The moment of inertia of a strip at a distance x from the y-axis is simply $dI = x^2 \, dM = \sigma b x^2 \, dx$. The total moment of inertia of the lamina about the y-axis is therefore

$$I = \int_0^a \sigma b x^2 \, dx = \frac{\sigma b a^3}{3}.$$

Since the total mass of the lamina is $M = \sigma ab$, we can write $I = \frac{1}{3} M a^2$. ◄

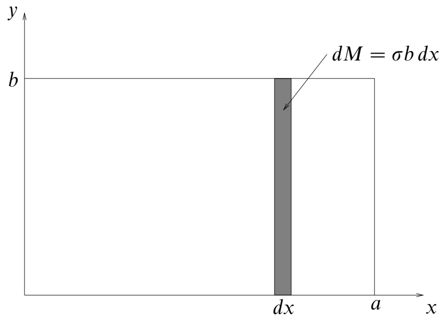


Figure 6.9 A uniform rectangular lamina of mass M with sides a and b can be divided into vertical strips.

6.3.5 Mean values of functions

In chapter 2 we discussed average values for functions of a single variable. This is easily extended to functions of several variables. Let us consider, for example, a function $f(x, y)$ defined in some region R of the xy-plane. Then the average value $\bar{f}$ of the function is given by

$$\bar{f} \int_R dA = \int_R f(x, y)\, dA. \tag{6.10}$$

This definition is easily extended to three (and higher) dimensions; if a function $f(x, y, z)$ is defined in some three-dimensional region of space R then the average value $\bar{f}$ of the function is given by

$$\bar{f} \int_R dV = \int_R f(x, y, z)\, dV. \tag{6.11}$$

►*A tetrahedron is bounded by the three coordinate surfaces and the plane $x/a+y/b+z/c = 1$ and has density $\rho(x, y, z) = \rho_0(1 + x/a)$. Find the average value of the density.*

From (6.11), the average value of the density is given by

$$\bar{\rho} \int_R dV = \int_R \rho(x, y, z)\, dV.$$

Now the integral on the LHS is just the volume of the tetrahedron, which we found in subsection 6.3.1 to be $V = \frac{1}{6}abc$, and the integral on the RHS is its mass $M = \frac{5}{24}abc\rho_0$, calculated in subsection 6.3.2. Therefore $\bar{\rho} = M/V = \frac{5}{4}\rho_0$. ◄

6.4 Change of variables in multiple integrals

It often happens that, either because of the form of the integrand involved or because of the boundary shape of the region of integration, it is desirable to

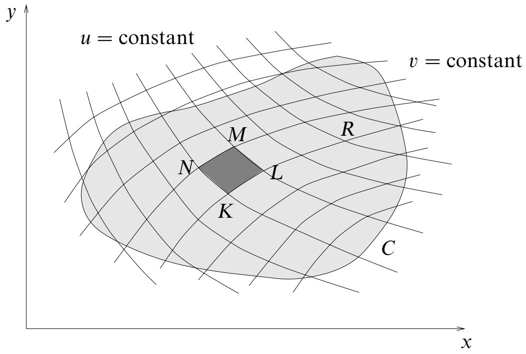


Figure 6.10 A region of integration R overlaid with a grid formed by the family of curves $u =$ constant and $v =$ constant. The parallelogram $KLMN$ defines the area element dA_{uv}.

express a multiple integral in terms of a new set of variables. We now consider how to do this.

6.4.1 Change of variables in double integrals

Let us begin by examining the change of variables in a double integral. Suppose that we require to change an integral

$$I = \iint_R f(x, y)\, dx\, dy,$$

in terms of coordinates x and y, into one expressed in new coordinates u and v, given in terms of x and y by differentiable equations $u = u(x, y)$ and $v = v(x, y)$ with inverses $x = x(u, v)$ and $y = y(u, v)$. The region R in the xy-plane and the curve C that bounds it will become a new region R' and a new boundary C' in the uv-plane, and so we must change the limits of integration accordingly. Also, the function $f(x, y)$ becomes a new function $g(u, v)$ of the new coordinates.

Now the part of the integral that requires most consideration is the area element. In the xy-plane the element is the rectangular area $dA_{xy} = dx\, dy$ generated by constructing a grid of straight lines parallel to the x- and y- axes respectively. Our task is to determine the corresponding area element in the uv-coordinates. In general the corresponding element dA_{uv} will not be the same shape as dA_{xy}, but this does not matter since all elements are infinitesimally small and the value of the integrand is considered constant over them. Since the sides of the area element are infinitesimal, dA_{uv} will in general have the shape of a parallelogram. We can find the connection between dA_{xy} and dA_{uv} by considering the grid formed by the family of curves $u =$ constant and $v =$ constant, as shown in figure 6.10. Since v

is constant along the line element KL, the latter has components $(\partial x/\partial u)\,du$ and $(\partial y/\partial u)\,du$ in the directions of the x- and y-axes respectively. Similarly, since u is constant along the line element KN, the latter has corresponding components $(\partial x/\partial v)\,dv$ and $(\partial y/\partial v)\,dv$. Using the result for the area of a parallelogram given in chapter 7, we find that the area of the parallelogram $KLMN$ is given by

$$\begin{aligned} dA_{uv} &= \left|\frac{\partial x}{\partial u}\,du\frac{\partial y}{\partial v}\,dv - \frac{\partial x}{\partial v}\,dv\frac{\partial y}{\partial u}\,du\right| \\ &= \left|\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right|\,du\,dv. \end{aligned}$$

Defining the *Jacobian* of x, y with respect to u, v as

$$J = \frac{\partial(x,y)}{\partial(u,v)} \equiv \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u},$$

we have

$$dA_{uv} = \left|\frac{\partial(x,y)}{\partial(u,v)}\right|\,du\,dv.$$

The reader acquainted with determinants will notice that the Jacobian can also be written as the 2 × 2 determinant

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} \\ \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} \end{vmatrix}.$$

Such determinants can be evaluated using the methods of chapter 8.

So, in summary, the relationship between the size of the area element generated by dx, dy and the size of the corresponding area element generated by du, dv is

$$dx\,dy = \left|\frac{\partial(x,y)}{\partial(u,v)}\right|\,du\,dv.$$

This equality should be taken as meaning that when transforming from coordinates x, y to coordinates u, v, the area element $dx\,dy$ should be replaced by the expression on the RHS of the above equality. Of course, the Jacobian can, and in general will, vary over the region of integration. We may express the double integral in either coordinate system as

$$I = \iint_R f(x,y)\,dx\,dy = \iint_{R'} g(u,v)\left|\frac{\partial(x,y)}{\partial(u,v)}\right|\,du\,dv. \tag{6.12}$$

When evaluating the integral in the new coordinate system, it is usually advisable to sketch the region of integration R' in the uv-plane.

►*Evaluate the double integral*

$$I = \iint_R \left(a + \sqrt{x^2 + y^2}\right)\, dx\, dy,$$

where R is the region bounded by the circle $x^2 + y^2 = a^2$.

In Cartesian coordinates, the integral may be written

$$I = \int_{-a}^{a} dx \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} dy\, \left(a + \sqrt{x^2 + y^2}\right),$$

and can be calculated directly. However, because of the circular boundary of the integration region, a change of variables to plane polar coordinates ρ, ϕ is indicated. The relationship between Cartesian and plane polar coordinates is given by $x = \rho\cos\phi$ and $y = \rho\sin\phi$. Using (6.12) we can therefore write

$$I = \iint_{R'} (a + \rho) \left| \frac{\partial(x, y)}{\partial(\rho, \phi)} \right| d\rho\, d\phi,$$

where R' is the rectangular region in the $\rho\phi$-plane whose sides are $\rho = 0$, $\rho = a$, $\phi = 0$ and $\phi = 2\pi$. The Jacobian is easily calculated, and we obtain

$$J = \frac{\partial(x, y)}{\partial(\rho, \phi)} = \begin{vmatrix} \cos\phi & \sin\phi \\ -\rho\sin\phi & \rho\cos\phi \end{vmatrix} = \rho(\cos^2\phi + \sin^2\phi) = \rho.$$

So the relationship between the area elements in Cartesian and in plane polar coordinates is

$$dx\, dy = \rho\, d\rho\, d\phi.$$

Therefore, when expressed in plane polar coordinates, the integral is given by

$$\begin{aligned} I &= \iint_{R'} (a + \rho)\rho\, d\rho\, d\phi \\ &= \int_0^{2\pi} d\phi \int_0^{a} d\rho\, (a + \rho)\rho = 2\pi \left[\frac{a\rho^2}{2} + \frac{\rho^3}{3} \right]_0^a = \frac{5\pi a^3}{3}. \end{aligned}$$ ◄

6.4.2 Evaluation of the integral $I = \int_{-\infty}^{\infty} e^{-x^2}\, dx$

By making a judicious change of variables, it is sometimes possible to evaluate an integral that would be intractable otherwise. An important example of this method is provided by the evaluation of the integral

$$I = \int_{-\infty}^{\infty} e^{-x^2}\, dx.$$

Its value may be found by first constructing I^2, as follows:

$$\begin{aligned} I^2 = \int_{-\infty}^{\infty} e^{-x^2}\, dx \int_{-\infty}^{\infty} e^{-y^2}\, dy &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\; e^{-(x^2+y^2)} \\ &= \iint_R e^{-(x^2+y^2)}\, dx\, dy, \end{aligned}$$

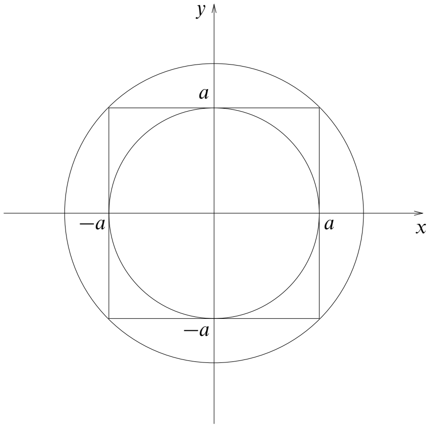


Figure 6.11 The regions used to illustrate the convergence properties of the integral $I(a) = \int_{-a}^{a} e^{-x^2}\,dx$ as $a \to \infty$.

where the region R is the whole xy-plane. Then, transforming to plane polar coordinates, we find

$$I^2 = \iint_{R'} e^{-\rho^2} \rho\, d\rho\, d\phi = \int_0^{2\pi} d\phi \int_0^{\infty} d\rho\, \rho e^{-\rho^2} = 2\pi \left[-\tfrac{1}{2} e^{-\rho^2} \right]_0^{\infty} = \pi.$$

Therefore the original integral is given by $I = \sqrt{\pi}$. Because the integrand is an even function of x, it follows that the value of the integral from 0 to ∞ is simply $\sqrt{\pi}/2$.

We note, however, that unlike in all the previous examples, the regions of integration R and R' are both infinite in extent (i.e. unbounded). It is therefore prudent to derive this result more rigorously; this we do by considering the integral

$$I(a) = \int_{-a}^{a} e^{-x^2}\, dx.$$

We then have

$$I^2(a) = \iint_R e^{-(x^2+y^2)}\, dx\, dy,$$

where R is the square of side $2a$ centred on the origin. Referring to figure 6.11, since the integrand is always positive the value of the integral taken over the square lies between the value of the integral taken over the region bounded by the inner circle of radius a and the value of the integral taken over the outer circle of radius $\sqrt{2}a$. Transforming to plane polar coordinates as above, we may

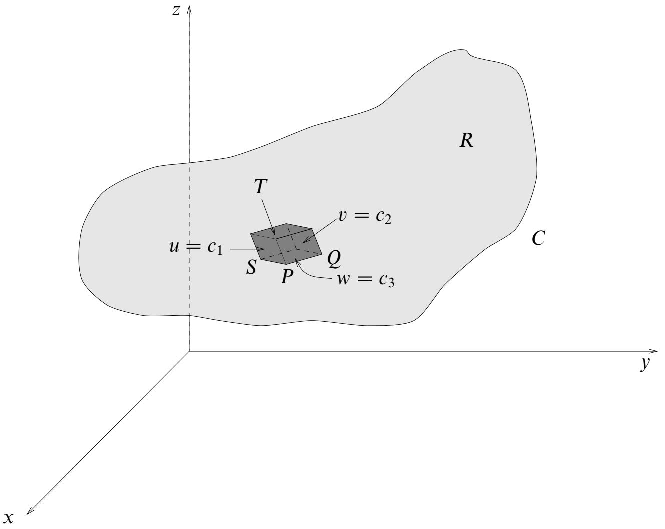


Figure 6.12 A three-dimensional region of integration R, showing an element of volume in u, v, w coordinates formed by the coordinate surfaces $u =$ constant, $v =$ constant, $w =$ constant.

evaluate the integrals over the inner and outer circles respectively, and we find

$$\pi\left(1 - e^{-a^2}\right) < I^2(a) < \pi\left(1 - e^{-2a^2}\right).$$

Taking the limit $a \to \infty$, we find $I^2(a) \to \pi$. Therefore $I = \sqrt{\pi}$, as we found previously. Substituting $x = \sqrt{\alpha}y$ shows that the corresponding integral of $\exp(-\alpha x^2)$ has the value $\sqrt{\pi/\alpha}$. We use this result in the discussion of the normal distribution in chapter 30.

6.4.3 Change of variables in triple integrals

A change of variable in a triple integral follows the same general lines as that for a double integral. Suppose we wish to change variables from x, y, z to u, v, w. In the x, y, z coordinates the element of volume is a cuboid of sides dx, dy, dz and volume $dV_{xyz} = dx\,dy\,dz$. If, however, we divide up the total volume into infinitesimal elements by constructing a grid formed from the coordinate surfaces $u =$ constant, $v =$ constant and $w =$ constant, then the element of volume dV_{uvw} in the new coordinates will have the shape of a parallelepiped whose faces are the coordinate surfaces and whose edges are the curves formed by the intersections of these surfaces (see figure 6.12). Along the line element PQ the coordinates v and

w are constant, and so PQ has components $(\partial x/\partial u)\,du$, $(\partial y/\partial u)\,du$ and $(\partial z/\partial u)\,du$ in the directions of the x-, y- and z- axes respectively. The components of the line elements PS and ST are found by replacing u by v and w respectively.

The expression for the volume of a parallelepiped in terms of the components of its edges with respect to the x-, y- and z-axes is given in chapter 7. Using this, we find that the element of volume in u, v, w coordinates is given by

$$dV_{uvw} = \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\,dv\,dw,$$

where the Jacobian of x, y, z with respect to u, v, w is a short-hand for a 3×3 determinant:

$$\frac{\partial(x,y,z)}{\partial(u,v,w)} \equiv \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} & \dfrac{\partial z}{\partial u} \\ \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} & \dfrac{\partial z}{\partial v} \\ \dfrac{\partial x}{\partial w} & \dfrac{\partial y}{\partial w} & \dfrac{\partial z}{\partial w} \end{vmatrix}.$$

So, in summary, the relationship between the elemental volumes in multiple integrals formulated in the two coordinate systems is given in Jacobian form by

$$dx\,dy\,dz = \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\,dv\,dw,$$

and we can write a triple integral in either set of coordinates as

$$I = \iiint_R f(x,y,z)\,dx\,dy\,dz = \iiint_{R'} g(u,v,w) \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\,dv\,dw.$$

► *Find an expression for a volume element in spherical polar coordinates, and hence calculate the moment of inertia about a diameter of a uniform sphere of radius a and mass M.*

Spherical polar coordinates r, θ, ϕ are defined by

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta$$

(and are discussed fully in chapter 10). The required Jacobian is therefore

$$J = \frac{\partial(x,y,z)}{\partial(r,\theta,\phi)} = \begin{vmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ r\cos\theta\cos\phi & r\cos\theta\sin\phi & -r\sin\theta \\ -r\sin\theta\sin\phi & r\sin\theta\cos\phi & 0 \end{vmatrix}.$$

The determinant is most easily evaluated by expanding it with respect to the last column (see chapter 8), which gives

$$\begin{aligned} J &= \cos\theta(r^2\sin\theta\cos\theta) + r\sin\theta(r\sin^2\theta) \\ &= r^2\sin\theta(\cos^2\theta + \sin^2\theta) = r^2\sin\theta. \end{aligned}$$

Therefore the volume element in spherical polar coordinates is given by

$$dV = \frac{\partial(x,y,z)}{\partial(r,\theta,\phi)}\,dr\,d\theta\,d\phi = r^2\sin\theta\,dr\,d\theta\,d\phi,$$

which agrees with the result given in chapter 10.

If we place the sphere with its centre at the origin of an x, y, z coordinate system then its moment of inertia about the z-axis (which is, of course, a diameter of the sphere) is

$$I = \int \left(x^2 + y^2\right)\, dM = \rho \int \left(x^2 + y^2\right)\, dV,$$

where the integral is taken over the sphere, and ρ is the density. Using spherical polar coordinates, we can write this as

$$\begin{aligned} I &= \rho \iiint_V \left(r^2 \sin^2\theta\right) r^2 \sin\theta \, dr\, d\theta\, d\phi \\ &= \rho \int_0^{2\pi} d\phi \int_0^{\pi} d\theta\, \sin^3\theta \int_0^a dr\, r^4 \\ &= \rho \times 2\pi \times \tfrac{4}{3} \times \tfrac{1}{5}a^5 = \tfrac{8}{15}\pi a^5 \rho. \end{aligned}$$

Since the mass of the sphere is $M = \frac{4}{3}\pi a^3 \rho$, the moment of inertia can also be written as $I = \frac{2}{5}Ma^2$. ◀

6.4.4 General properties of Jacobians

Although we will not prove it, the general result for a change of coordinates in an n-dimensional integral from a set x_i to a set y_j (where i and j both run from 1 to n) is

$$dx_1\, dx_2 \cdots dx_n = \left| \frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)} \right| dy_1\, dy_2 \cdots dy_n,$$

where the n-dimensional Jacobian can be written as an $n \times n$ determinant (see chapter 8) in an analogous way to the two- and three-dimensional cases.

For readers who already have sufficient familiarity with matrices (see chapter 8) and their properties, a fairly compact proof of some useful general properties of Jacobians can be given as follows. Other readers should turn straight to the results (6.16) and (6.17) and return to the proof at some later time.

Consider three sets of variables x_i, y_i and z_i, with i running from 1 to n for each set. From the chain rule in partial differentiation (see (5.17)), we know that

$$\frac{\partial x_i}{\partial z_j} = \sum_{k=1}^{n} \frac{\partial x_i}{\partial y_k} \frac{\partial y_k}{\partial z_j}. \qquad (6.13)$$

Now let A, B and C be the matrices whose ijth elements are $\partial x_i/\partial y_j$, $\partial y_i/\partial z_j$ and $\partial x_i/\partial z_j$ respectively. We can then write (6.13) as the matrix product

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \qquad \text{or} \qquad \mathsf{C} = \mathsf{AB}. \qquad (6.14)$$

We may now use the general result for the determinant of the product of two matrices, namely $|\mathsf{AB}| = |\mathsf{A}||\mathsf{B}|$, and recall that the Jacobian

$$J_{xy} = \frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)} = |\mathsf{A}|, \qquad (6.15)$$

and similarly for J_{yz} and J_{xz}. On taking the determinant of (6.14), we therefore obtain

$$J_{xz} = J_{xy}J_{yz}$$

or, in the usual notation,

$$\frac{\partial(x_1,\ldots,x_n)}{\partial(z_1,\ldots,z_n)} = \frac{\partial(x_1,\ldots,x_n)}{\partial(y_1,\ldots,y_n)}\frac{\partial(y_1,\ldots,y_n)}{\partial(z_1,\ldots,z_n)}. \tag{6.16}$$

As a special case, if the set z_i is taken to be identical to the set x_i, and the obvious result $J_{xx} = 1$ is used, we obtain

$$J_{xy}J_{yx} = 1$$

or, in the usual notation,

$$\frac{\partial(x_1,\ldots,x_n)}{\partial(y_1,\ldots,y_n)} = \left[\frac{\partial(y_1,\ldots,y_n)}{\partial(x_1,\ldots,x_n)}\right]^{-1}. \tag{6.17}$$

The similarity between the properties of Jacobians and those of derivatives is apparent, and to some extent is suggested by the notation. We further note from (6.15) that since $|\mathsf{A}| = |\mathsf{A}^{\mathrm{T}}|$, where A^{T} is the transpose of A, we can interchange the rows and columns in the determinantal form of the Jacobian without changing its value.

6.5 Exercises

6.1 Identify the curved wedge bounded by the surfaces $y^2 = 4ax$, $x + z = a$ and $z = 0$, and hence calculate its volume V.

6.2 Evaluate the volume integral of $x^2 + y^2 + z^2$ over the rectangular parallelepiped bounded by the six surfaces $x = \pm a$, $y = \pm b$ and $z = \pm c$.

6.3 Find the volume integral of x^2y over the tetrahedral volume bounded by the planes $x = 0$, $y = 0$, $z = 0$, and $x + y + z = 1$.

6.4 Evaluate the surface integral of $f(x, y)$ over the rectangle $0 \le x \le a$, $0 \le y \le b$ for the functions

$$\text{(a) } f(x,y) = \frac{x}{x^2 + y^2}, \qquad \text{(b) } f(x,y) = (b - y + x)^{-3/2}.$$

6.5 Calculate the volume of an ellipsoid as follows:

(a) Prove that the area of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

is πab.

(b) Use this result to obtain an expression for the volume of a slice of thickness dz of the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

Hence show that the volume of the ellipsoid is $4\pi abc/3$.

6.6 The function

$$\Psi(r) = A\left(2 - \frac{Zr}{a}\right) e^{-Zr/2a}$$

gives the form of the quantum-mechanical wavefunction representing the electron in a hydrogen-like atom of atomic number Z, when the electron is in its first allowed spherically symmetric excited state. Here r is the usual spherical polar coordinate, but, because of the spherical symmetry, the coordinates θ and ϕ do not appear explicitly in Ψ. Determine the value that A (assumed real) must have if the wavefunction is to be correctly normalised, i.e. if the volume integral of $|\Psi|^2$ over all space is to be equal to unity.

6.7 In quantum mechanics the electron in a hydrogen atom in some particular state is described by a wavefunction Ψ, which is such that $|\Psi|^2\,dV$ is the probability of finding the electron in the infinitesimal volume dV. In spherical polar coordinates $\Psi = \Psi(r,\theta,\phi)$ and $dV = r^2 \sin\theta\,dr\,d\theta\,d\phi$. Two such states are described by

$$\Psi_1 = \left(\frac{1}{4\pi}\right)^{1/2}\left(\frac{1}{a_0}\right)^{3/2} 2e^{-r/a_0},$$

$$\Psi_2 = -\left(\frac{3}{8\pi}\right)^{1/2} \sin\theta\, e^{i\phi}\left(\frac{1}{2a_0}\right)^{3/2}\frac{re^{-r/2a_0}}{a_0\sqrt{3}}.$$

(a) Show that each Ψ_i is normalised, i.e. the integral over all space $\int |\Psi|^2\,dV$ is equal to unity – physically, this means that the electron must be somewhere.

(b) The (so-called) dipole matrix element between the states 1 and 2 is given by the integral

$$p_x = \int \Psi_1^* qr\sin\theta\cos\phi\;\Psi_2\,dV,$$

where q is the charge on the electron. Prove that p_x has the value $-2^7 qa_0/3^5$.

6.8 A planar figure is formed from uniform wire and consists of two equal semicircular arcs, each with its own closing diameter, joined so as to form a letter ‘B’. The figure is freely suspended from its top left-hand corner. Show that the straight edge of the figure makes an angle θ with the vertical given by $\tan\theta = (2+\pi)^{-1}$.

6.9 A certain torus has a circular vertical cross-section of radius a centred on a horizontal circle of radius c ($> a$).

(a) Find the volume V and surface area A of the torus, and show that they can be written as

$$V = \frac{\pi^2}{4}(r_{\rm o}^2 - r_{\rm i}^2)(r_{\rm o} - r_{\rm i}), \qquad A = \pi^2(r_{\rm o}^2 - r_{\rm i}^2),$$

where $r_{\rm o}$ and $r_{\rm i}$ are, respectively, the outer and inner radii of the torus.

(b) Show that a vertical circular cylinder of radius c, coaxial with the torus, divides A in the ratio

$$\pi c + 2a \;:\; \pi c - 2a.$$

6.10 A thin uniform circular disc has mass M and radius a.

(a) Prove that its moment of inertia about an axis perpendicular to its plane and passing through its centre is $\frac{1}{2}Ma^2$.

(b) Prove that the moment of inertia of the same disc about a diameter is $\frac{1}{4}Ma^2$.

This is an example of the general result for planar bodies that the moment of inertia of the body about an axis perpendicular to the plane is equal to the sum of the moments of inertia about two perpendicular axes lying in the plane; in an obvious notation

$$I_z = \int r^2\,dm = \int (x^2 + y^2)\,dm = \int x^2\,dm + \int y^2\,dm = I_y + I_x.$$

6.11 In some applications in mechanics the moment of inertia of a body about a single point (as opposed to about an axis) is needed. The moment of inertia, I, about the origin of a uniform solid body of density ρ is given by the volume integral

$$I = \int_V (x^2 + y^2 + z^2)\rho\,dV.$$

Show that the moment of inertia of a right circular cylinder of radius a, length $2b$ and mass M about its centre is

$$M\left(\frac{a^2}{2} + \frac{b^2}{3}\right).$$

6.12 The shape of an axially symmetric hard-boiled egg, of uniform density ρ_0, is given in spherical polar coordinates by $r = a(2 - \cos\theta)$, where θ is measured from the axis of symmetry.

(a) Prove that the mass M of the egg is $M = \frac{40}{3}\pi\rho_0 a^3$.

(b) Prove that the egg's moment of inertia about its axis of symmetry is $\frac{342}{175}Ma^2$.

6.13 In spherical polar coordinates r, θ, ϕ the element of volume for a body that is symmetrical about the polar axis is $dV = 2\pi r^2 \sin\theta\,dr\,d\theta$, whilst its element of surface area is $2\pi r \sin\theta[(dr)^2 + r^2(d\theta)^2]^{1/2}$. A particular surface is defined by $r = 2a\cos\theta$, where a is a constant and $0 \le \theta \le \pi/2$. Find its total surface area and the volume it encloses, and hence identify the surface.

6.14 By expressing both the integrand and the surface element in spherical polar coordinates, show that the surface integral

$$\int \frac{x^2}{x^2 + y^2}\,dS$$

over the surface $x^2 + y^2 = z^2$, $0 \le z \le 1$, has the value $\pi/\sqrt{2}$.

6.15 By transforming to cylindrical polar coordinates, evaluate the integral

$$I = \int\int\int \ln(x^2 + y^2)\,dx\,dy\,dz$$

over the interior of the conical region $x^2 + y^2 \le z^2$, $0 \le z \le 1$.

6.16 Sketch the two families of curves

$$y^2 = 4u(u - x), \qquad y^2 = 4v(v + x),$$

where u and v are parameters.

By transforming to the uv-plane, evaluate the integral of $y/(x^2 + y^2)^{1/2}$ over the part of the quadrant $x > 0$, $y > 0$ that is bounded by the lines $x = 0$, $y = 0$ and the curve $y^2 = 4a(a - x)$.

6.17 By making two successive simple changes of variables, evaluate

$$I = \int\int\int x^2\,dx\,dy\,dz$$

over the ellipsoidal region

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \le 1.$$

6.18 Sketch the domain of integration for the integral

$$I = \int_0^1 \int_{x=y}^{1/y} \frac{y^3}{x} \exp[y^2(x^2 + x^{-2})]\, dx\, dy$$

and characterise its boundaries in terms of new variables $u = xy$ and $v = y/x$. Show that the Jacobian for the change from (x, y) to (u, v) is equal to $(2v)^{-1}$, and hence evaluate I.

6.19 Sketch the part of the region $0 \le x$, $0 \le y \le \pi/2$ that is bounded by the curves $x = 0$, $y = 0$, $\sinh x \cos y = 1$ and $\cosh x \sin y = 1$. By making a suitable change of variables, evaluate the integral

$$I = \int\int (\sinh^2 x + \cos^2 y) \sinh 2x \sin 2y\, dx\, dy$$

over the bounded subregion.

6.20 Define a coordinate system u, v whose origin coincides with that of the usual x, y system and whose u-axis coincides with the x-axis, whilst the v-axis makes an angle α with it. By considering the integral $I = \int \exp(-r^2)\, dA$, where r is the radial distance from the origin, over the area defined by $0 \le u < \infty$, $0 \le v < \infty$, prove that

$$\int_0^\infty \int_0^\infty \exp(-u^2 - v^2 - 2uv\cos\alpha)\, du\, dv = \frac{\alpha}{2\sin\alpha}.$$

6.21 As stated in section 5.11, the first law of thermodynamics can be expressed as

$$dU = T\, dS - P\, dV.$$

By calculating and equating $\partial^2 U/\partial Y \partial X$ and $\partial^2 U/\partial X \partial Y$, where X and Y are an unspecified pair of variables (drawn from P, V, T and S), prove that

$$\frac{\partial(S, T)}{\partial(X, Y)} = \frac{\partial(V, P)}{\partial(X, Y)}.$$

Using the properties of Jacobians, deduce that

$$\frac{\partial(S, T)}{\partial(V, P)} = 1.$$

6.22 The distances of the variable point P, which has coordinates x, y, z, from the fixed points $(0, 0, 1)$ and $(0, 0, -1)$ are denoted by u and v respectively. New variables ξ, η, ϕ are defined by

$$\xi = \tfrac{1}{2}(u + v), \qquad \eta = \tfrac{1}{2}(u - v),$$

and ϕ is the angle between the plane $y = 0$ and the plane containing the three points. Prove that the Jacobian $\partial(\xi, \eta, \phi)/\partial(x, y, z)$ has the value $(\xi^2 - \eta^2)^{-1}$ and that

$$\int\int\int_{\text{all space}} \frac{(u - v)^2}{uv} \exp\left(-\frac{u + v}{2}\right) dx\, dy\, dz = \frac{16\pi}{3e}.$$

6.23 This is a more difficult question about 'volumes' in an increasing number of dimensions.

(a) Let R be a real positive number and define K_m by

$$K_m = \int_{-R}^{R} \left(R^2 - x^2\right)^m \, dx.$$

Show, using integration by parts, that K_m satisfies the recurrence relation

$$(2m+1)K_m = 2mR^2K_{m-1}.$$

(b) For integer n, define $I_n = K_n$ and $J_n = K_{n+1/2}$. Evaluate I_0 and J_0 directly and hence prove that

$$I_n = \frac{2^{2n+1}(n!)^2R^{2n+1}}{(2n+1)!} \qquad \text{and} \qquad J_n = \frac{\pi(2n+1)!R^{2n+2}}{2^{2n+1}n!(n+1)!}.$$

(c) A sequence of functions $V_n(R)$ is defined by

$$\begin{aligned} V_0(R) &= 1, \\ V_n(R) &= \int_{-R}^{R} V_{n-1}\left(\sqrt{R^2 - x^2}\right) dx, \qquad n \geq 1. \end{aligned}$$

Prove by induction that

$$V_{2n}(R) = \frac{\pi^n R^{2n}}{n!}, \qquad V_{2n+1}(R) = \frac{\pi^n 2^{2n+1} n! R^{2n+1}}{(2n+1)!}.$$

(d) For interest,

(i) show that $V_{2n+2}(1) < V_{2n}(1)$ and $V_{2n+1}(1) < V_{2n-1}(1)$ for all $n \geq 3$;

(ii) hence, by explicitly writing out $V_k(R)$ for $1 \leq k \leq 8$ (say), show that the 'volume' of the totally symmetric solid of unit radius is a maximum in five dimensions.

6.6 Hints and answers

6.1 For integration order z, y, x, the limits are $(0, a-x)$, $(-\sqrt{4ax}, \sqrt{4ax})$ and $(0, a)$. For integration order y, x, z, the limits are $(-\sqrt{4ax}, \sqrt{4ax})$, $(0, a-z)$ and $(0, a)$. $V = 16a^3/15$.

6.3 1/360.

6.5 (a) Evaluate $\int 2b[1-(x/a)^2]^{1/2}\,dx$ by setting $x = a\cos\phi$;
(b) $dV = \pi \times a[1-(z/c)^2]^{1/2} \times b[1-(z/c)^2]^{1/2}\,dz$.

6.7 Write $\sin^3\theta$ as $(1-\cos^2\theta)\sin\theta$ when integrating $|\Psi_2|^2$.

6.9 (a) $V = 2\pi c \times \pi a^2$ and $A = 2\pi a \times 2\pi c$. Setting $r_{\mathrm{o}} = c + a$ and $r_{\mathrm{i}} = c - a$ gives the stated results. (b) Show that the centre of gravity of either half is $2a/\pi$ from the cylinder.

6.11 Transform to cylindrical polar coordinates.

6.13 $4\pi a^2$; $4\pi a^3/3$; a sphere.

6.15 The volume element is $\rho\, d\phi\, d\rho\, dz$. The integrand for the final z-integration is given by $2\pi[(z^2\ln z) - (z^2/2)]$; $I = -5\pi/9$.

6.17 Set $\xi = x/a$, $\eta = y/b$, $\zeta = z/c$ to map the ellipsoid onto the unit sphere, and then change from (ξ, η, ζ) coordinates to spherical polar coordinates; $I = 4\pi a^3bc/15$.

6.19 Set $u = \sinh x \cos y$ and $v = \cosh x \sin y$; $J_{xy,uv} = (\sinh^2 x + \cos^2 y)^{-1}$ and the integrand reduces to $4uv$ over the region $0 \leq u \leq 1$, $0 \leq v \leq 1$; $I = 1$.

6.21 Terms such as $T\partial^2S/\partial Y\,\partial X$ cancel in pairs. Use equations (6.17) and (6.16).

6.23 (c) Show that the two expressions mutually support the integration formula given for computing a volume in the next higher dimension.
(d)(ii) 2, π, $4\pi/3$, $\pi^2/2$, $8\pi^2/15$, $\pi^3/6$, $16\pi^3/105$, $\pi^4/24$.

7

Vector algebra

This chapter introduces space vectors and their manipulation. Firstly we deal with the description and algebra of vectors, then we consider how vectors may be used to describe lines and planes and finally we look at the practical use of vectors in finding distances. Much use of vectors will be made in subsequent chapters; this chapter gives only some basic rules.

7.1 Scalars and vectors

The simplest kind of physical quantity is one that can be completely specified by its magnitude, a single number, together with the units in which it is measured. Such a quantity is called a *scalar* and examples include temperature, time and density.

A *vector* is a quantity that requires both a magnitude (≥ 0) and a direction in space to specify it completely; we may think of it as an arrow in space. A familiar example is force, which has a magnitude (strength) measured in newtons and a direction of application. The large number of vectors that are used to describe the physical world include velocity, displacement, momentum and electric field. Vectors are also used to describe quantities such as angular momentum and surface elements (a surface element has an area and a direction defined by the normal to its tangent plane); in such cases their definitions may seem somewhat arbitrary (though in fact they are standard) and not as physically intuitive as for vectors such as force. A vector is denoted by bold type, the convention of this book, or by underlining, the latter being much used in handwritten work.

This chapter considers basic vector algebra and illustrates just how powerful vector analysis can be. All the techniques are presented for three-dimensional space but most can be readily extended to more dimensions.

Throughout the book we will represent a vector in diagrams as a line together with an arrowhead. We will make no distinction between an arrowhead at the

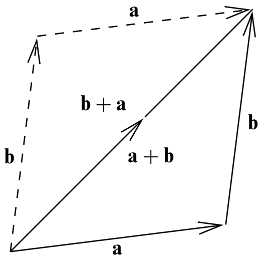


Figure 7.1 Addition of two vectors showing the commutation relation. We make no distinction between an arrowhead at the end of the line and one along the line's length, but rather use that which gives the clearer diagram.

end of the line and one along the line's length but, rather, use that which gives the clearer diagram. Furthermore, even though we are considering three-dimensional vectors, we have to draw them in the plane of the paper. It should not be assumed that vectors drawn thus are coplanar, unless this is explicitly stated.

7.2 Addition and subtraction of vectors

The *resultant* or *vector sum* of two displacement vectors is the displacement vector that results from performing first one and then the other displacement, as shown in figure 7.1; this process is known as vector addition. However, the principle of addition has physical meaning for vector quantities other than displacements; for example, if two forces act on the same body then the resultant force acting on the body is the vector sum of the two. The addition of vectors only makes physical sense if they are of a like kind, for example if they are both forces acting in three dimensions. It may be seen from figure 7.1 that vector addition is commutative, i.e.

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}. \tag{7.1}$$

The generalisation of this procedure to the addition of three (or more) vectors is clear and leads to the associativity property of addition (see figure 7.2), e.g.

$$\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}. \tag{7.2}$$

Thus, it is immaterial in what order any number of vectors are added.

The subtraction of two vectors is very similar to their addition (see figure 7.3), that is,

$$\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b})$$

where $-\mathbf{b}$ is a vector of equal magnitude but exactly opposite direction to vector $\mathbf{b}$.

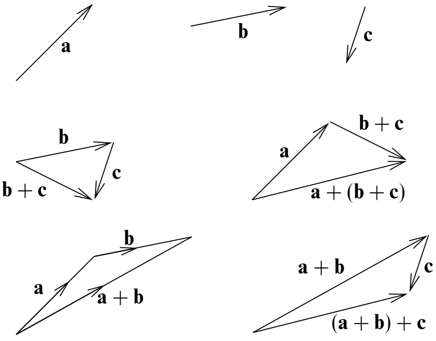


Figure 7.2 Addition of three vectors showing the associativity relation.

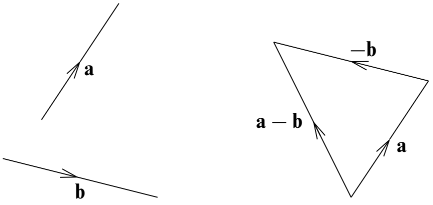


Figure 7.3 Subtraction of two vectors.

The subtraction of two equal vectors yields the zero vector, **0**, which has zero magnitude and no associated direction.

7.3 Multiplication by a scalar

Multiplication of a vector by a scalar (not to be confused with the 'scalar product', to be discussed in subsection 7.6.1) gives a vector in the same direction as the original but of a proportional magnitude. This can be seen in figure 7.4. The scalar may be positive, negative or zero. It can also be complex in some applications. Clearly, when the scalar is negative we obtain a vector pointing in the opposite direction to the original vector. Multiplication by a scalar is associative, commutative and distributive over addition. These properties may be summarised for arbitrary vectors **a** and **b** and arbitrary scalars λ and μ by

$$(\lambda\mu)\mathbf{a} = \lambda(\mu\mathbf{a}) = \mu(\lambda\mathbf{a}), \tag{7.3}$$

$$\lambda(\mathbf{a} + \mathbf{b}) = \lambda\mathbf{a} + \lambda\mathbf{b}, \tag{7.4}$$

$$(\lambda + \mu)\mathbf{a} = \lambda\mathbf{a} + \mu\mathbf{a}. \tag{7.5}$$

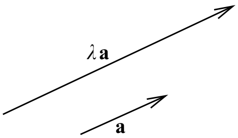


Figure 7.4 Scalar multiplication of a vector (for $\lambda > 1$).

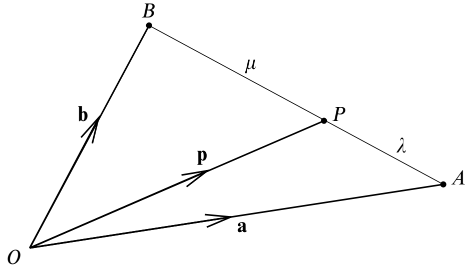


Figure 7.5 An illustration of the ratio theorem. The point P divides the line segment AB in the ratio $\lambda : \mu$.

Having defined the operations of addition, subtraction and multiplication by a scalar, we can now use vectors to solve simple problems in geometry.

►*A point P divides a line segment AB in the ratio $\lambda : \mu$ (see figure 7.5). If the position vectors of the points A and B are* **a** *and* **b**, *respectively, find the position vector of the point P.*

As is conventional for vector geometry problems, we denote the vector from the point A to the point B by **AB**. If the position vectors of the points A and B, relative to some origin O, are **a** and **b**, it should be clear that $\mathbf{AB} = \mathbf{b} - \mathbf{a}$.

Now, from figure 7.5 we see that one possible way of reaching the point P from O is first to go from O to A and to go along the line AB for a distance equal to the the fraction $\lambda/(\lambda + \mu)$ of its total length. We may express this in terms of vectors as

$$\begin{aligned}\mathbf{OP} = \mathbf{p} &= \mathbf{a} + \frac{\lambda}{\lambda + \mu}\mathbf{AB} \\ &= \mathbf{a} + \frac{\lambda}{\lambda + \mu}(\mathbf{b} - \mathbf{a}) \\ &= \left(1 - \frac{\lambda}{\lambda + \mu}\right)\mathbf{a} + \frac{\lambda}{\lambda + \mu}\mathbf{b} \\ &= \frac{\mu}{\lambda + \mu}\mathbf{a} + \frac{\lambda}{\lambda + \mu}\mathbf{b}, \qquad (7.6)\end{aligned}$$

which expresses the position vector of the point P in terms of those of A and B. We would, of course, obtain the same result by considering the path from O to B and then to P. ◄

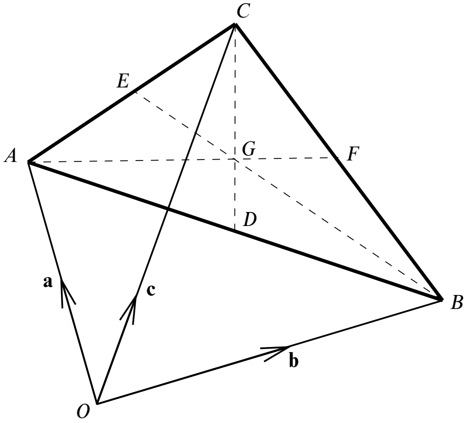


Figure 7.6 The centroid of a triangle. The triangle is defined by the points A, B and C that have position vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. The broken lines CD, BE, AF connect the vertices of the triangle to the mid-points of the opposite sides; these lines intersect at the centroid G of the triangle.

Result (7.6) is a version of the *ratio theorem* and we may use it in solving more complicated problems.

►*The vertices of triangle ABC have position vectors* $\mathbf{a}$, $\mathbf{b}$ *and* $\mathbf{c}$ *relative to some origin O (see figure 7.6). Find the position vector of the centroid G of the triangle.*

From figure 7.6, the points D and E bisect the lines AB and AC respectively. Thus from the ratio theorem (7.6), with $\lambda = \mu = 1/2$, the position vectors of D and E relative to the origin are

$$\mathbf{d} = \tfrac{1}{2}\mathbf{a} + \tfrac{1}{2}\mathbf{b},$$
$$\mathbf{e} = \tfrac{1}{2}\mathbf{a} + \tfrac{1}{2}\mathbf{c}.$$

Using the ratio theorem again, we may write the position vector of a general point on the line CD that divides the line in the ratio $\lambda : (1-\lambda)$ as

$$\begin{aligned}\mathbf{r} &= (1-\lambda)\mathbf{c} + \lambda\mathbf{d},\\ &= (1-\lambda)\mathbf{c} + \tfrac{1}{2}\lambda(\mathbf{a}+\mathbf{b}),\end{aligned} \tag{7.7}$$

where we have expressed $\mathbf{d}$ in terms of $\mathbf{a}$ and $\mathbf{b}$. Similarly, the position vector of a general point on the line BE can be expressed as

$$\begin{aligned}\mathbf{r} &= (1-\mu)\mathbf{b} + \mu\mathbf{e},\\ &= (1-\mu)\mathbf{b} + \tfrac{1}{2}\mu(\mathbf{a}+\mathbf{c}).\end{aligned} \tag{7.8}$$

Thus, at the intersection of the lines CD and BE we require, from (7.7), (7.8),

$$(1-\lambda)\mathbf{c} + \tfrac{1}{2}\lambda(\mathbf{a}+\mathbf{b}) = (1-\mu)\mathbf{b} + \tfrac{1}{2}\mu(\mathbf{a}+\mathbf{c}).$$

By equating the coefficents of the vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ we find

$$\lambda = \mu, \qquad \tfrac{1}{2}\lambda = 1-\mu, \qquad 1-\lambda = \tfrac{1}{2}\mu.$$

These equations are consistent and have the solution $\lambda = \mu = 2/3$. Substituting these values into either (7.7) or (7.8) we find that the position vector of the centroid G is given by

$$\mathbf{g} = \tfrac{1}{3}(\mathbf{a} + \mathbf{b} + \mathbf{c}). ◄$$

7.4 Basis vectors and components

Given any three different vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$, which do not all lie in a plane, it is possible, in a three-dimensional space, to write any other vector in terms of scalar multiples of them:

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3. \tag{7.9}$$

The three vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are said to form a *basis* (for the three-dimensional space); the scalars a_1, a_2 and a_3, which may be positive, negative or zero, are called the *components* of the vector $\mathbf{a}$ with respect to this basis. We say that the vector has been *resolved* into components.

Most often we shall use basis vectors that are mutually perpendicular, for ease of manipulation, though this is not necessary. In general, a basis set must

(i) have as many basis vectors as the number of dimensions (in more formal language, the basis vectors must span the space) and

(ii) be such that no basis vector may be described as a sum of the others, or, more formally, the basis vectors must be *linearly independent*. Putting this mathematically, in N dimensions, we require

$$c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_N\mathbf{e}_N \neq \mathbf{0},$$

for any set of coefficients $c_1, c_2, \ldots, c_N$ except $c_1 = c_2 = \cdots = c_N = 0$.

In this chapter we will only consider vectors in three dimensions; higher dimensionality can be achieved by simple extension.

If we wish to label points in space using a Cartesian coordinate system (x, y, z), we may introduce the unit vectors $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$, which point along the positive x-, y- and z- axes respectively. A vector $\mathbf{a}$ may then be written as a sum of three vectors, each parallel to a different coordinate axis:

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}. \tag{7.10}$$

A vector in three-dimensional space thus requires three components to describe fully both its direction and its magnitude. A displacement in space may be thought of as the sum of displacements along the x-, y- and z- directions (see figure 7.7). For brevity, the components of a vector $\mathbf{a}$ with respect to a particular coordinate system are sometimes written in the form (a_x, a_y, a_z). Note that the

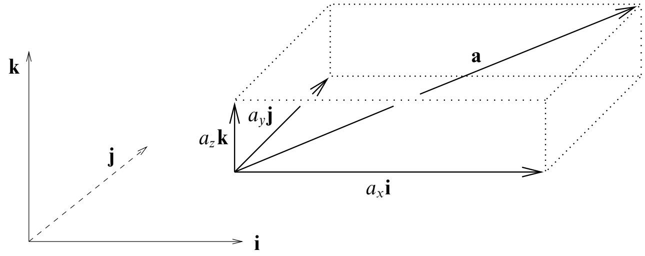


Figure 7.7 A Cartesian basis set. The vector **a** is the sum of $a_x\mathbf{i}$, $a_y\mathbf{j}$ and $a_z\mathbf{k}$.

basis vectors **i**, **j** and **k** may themselves be represented by $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ respectively.

We can consider the addition and subtraction of vectors in terms of their components. The sum of two vectors **a** and **b** is found by simply adding their components, i.e.

$$\begin{aligned}\mathbf{a}+\mathbf{b} &= a_x\mathbf{i}+a_y\mathbf{j}+a_z\mathbf{k}+b_x\mathbf{i}+b_y\mathbf{j}+b_z\mathbf{k}\\ &= (a_x+b_x)\mathbf{i}+(a_y+b_y)\mathbf{j}+(a_z+b_z)\mathbf{k},\end{aligned} \tag{7.11}$$

and their difference by subtracting them,

$$\begin{aligned}\mathbf{a}-\mathbf{b} &= a_x\mathbf{i}+a_y\mathbf{j}+a_z\mathbf{k}-(b_x\mathbf{i}+b_y\mathbf{j}+b_z\mathbf{k})\\ &= (a_x-b_x)\mathbf{i}+(a_y-b_y)\mathbf{j}+(a_z-b_z)\mathbf{k}.\end{aligned} \tag{7.12}$$

►*Two particles have velocities* $\mathbf{v}_1 = \mathbf{i}+3\mathbf{j}+6\mathbf{k}$ *and* $\mathbf{v}_2 = \mathbf{i}-2\mathbf{k}$*, respectively. Find the velocity* **u** *of the second particle relative to the first.*

The required relative velocity is given by

$$\begin{aligned}\mathbf{u}=\mathbf{v}_2-\mathbf{v}_1 &= (1-1)\mathbf{i}+(0-3)\mathbf{j}+(-2-6)\mathbf{k}\\ &= -3\mathbf{j}-8\mathbf{k}.\ ◄\end{aligned}$$

7.5 Magnitude of a vector

The magnitude of the vector **a** is denoted by $|\mathbf{a}|$ or a. In terms of its components in three-dimensional Cartesian coordinates, the magnitude of **a** is given by

$$a \equiv |\mathbf{a}| = \sqrt{a_x^2+a_y^2+a_z^2}. \tag{7.13}$$

Hence, the magnitude of a vector is a measure of its length. Such an analogy is useful for displacement vectors but magnitude is better described, for example, by 'strength' for vectors such as force or by 'speed' for velocity vectors. For instance,

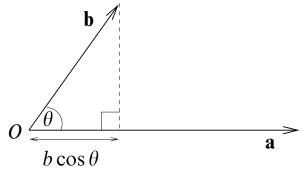


Figure 7.8 The projection of **b** onto the direction of **a** is $b\cos\theta$. The scalar product of **a** and **b** is $ab\cos\theta$.

in the previous example, the speed of the second particle relative to the first is given by

$$u = |\mathbf{u}| = \sqrt{(-3)^2 + (-8)^2} = \sqrt{73}.$$

A vector whose magnitude equals unity is called a *unit vector*. The unit vector in the direction **a** is usually notated $\hat{\mathbf{a}}$ and may be evaluated as

$$\hat{\mathbf{a}} = \frac{\mathbf{a}}{|\mathbf{a}|}. \tag{7.14}$$

The unit vector is a useful concept because a vector written as $\lambda\hat{\mathbf{a}}$ then has magnitude λ and direction $\hat{\mathbf{a}}$. Thus magnitude and direction are explicitly separated.

7.6 Multiplication of vectors

We have already considered multiplying a vector by a scalar. Now we consider the concept of multiplying one vector by another vector. It is not immediately obvious what the product of two vectors represents and in fact two products are commonly defined, the *scalar product* and the *vector product*. As their names imply, the scalar product of two vectors is just a number, whereas the vector product is itself a vector. Although neither the scalar nor the vector product is what we might normally think of as a product, their use is widespread and numerous examples will be described elsewhere in this book.

7.6.1 Scalar product

The scalar product (or dot product) of two vectors **a** and **b** is denoted by $\mathbf{a}\cdot\mathbf{b}$ and is given by

$$\mathbf{a}\cdot\mathbf{b} \equiv |\mathbf{a}||\mathbf{b}|\cos\theta, \qquad 0 \le \theta \le \pi, \tag{7.15}$$

where θ is the angle between the two vectors, placed 'tail to tail' or 'head to head'. Thus, the value of the scalar product $\mathbf{a}\cdot\mathbf{b}$ equals the magnitude of **a** multiplied by the projection of **b** onto **a** (see figure 7.8).

From (7.15) we see that the scalar product has the particularly useful property that

$$\mathbf{a}\cdot\mathbf{b}=0 \tag{7.16}$$

is a necessary and sufficient condition for **a** to be perpendicular to **b** (unless either of them is zero). It should be noted in particular that the Cartesian basis vectors **i**, **j** and **k**, being mutually orthogonal unit vectors, satisfy the equations

$$\mathbf{i}\cdot\mathbf{i}=\mathbf{j}\cdot\mathbf{j}=\mathbf{k}\cdot\mathbf{k}=1, \tag{7.17}$$

$$\mathbf{i}\cdot\mathbf{j}=\mathbf{j}\cdot\mathbf{k}=\mathbf{k}\cdot\mathbf{i}=0. \tag{7.18}$$

Examples of scalar products arise naturally throughout physics and in particular in connection with energy. Perhaps the simplest is the work done $\mathbf{F}\cdot\mathbf{r}$ in moving the point of application of a constant force **F** through a displacement **r**; notice that, as expected, if the displacement is perpendicular to the direction of the force then $\mathbf{F}\cdot\mathbf{r}=0$ and no work is done. A second simple example is afforded by the potential energy $-\mathbf{m}\cdot\mathbf{B}$ of a magnetic dipole, represented in strength and orientation by a vector **m**, placed in an external magnetic field **B**.

As the name implies, the scalar product has a magnitude but no direction. The scalar product is commutative and distributive over addition:

$$\mathbf{a}\cdot\mathbf{b}=\mathbf{b}\cdot\mathbf{a} \tag{7.19}$$

$$\mathbf{a}\cdot(\mathbf{b}+\mathbf{c})=\mathbf{a}\cdot\mathbf{b}+\mathbf{a}\cdot\mathbf{c}. \tag{7.20}$$

►*Four non-coplanar points A, B, C, D are positioned such that the line AD is perpendicular to BC and BD is perpendicular to AC. Show that CD is perpendicular to AB.*

Denote the four position vectors by **a**, **b**, **c**, **d**. As none of the three pairs of lines actually intersect, it is difficult to indicate their orthogonality in the diagram we would normally draw. However, the orthogonality can be expressed in vector form and we start by noting that, since $AD \perp BC$, it follows from (7.16) that

$$(\mathbf{d}-\mathbf{a})\cdot(\mathbf{c}-\mathbf{b})=0.$$

Similarly, since $BD \perp AC$,

$$(\mathbf{d}-\mathbf{b})\cdot(\mathbf{c}-\mathbf{a})=0.$$

Combining these two equations we find

$$(\mathbf{d}-\mathbf{a})\cdot(\mathbf{c}-\mathbf{b})=(\mathbf{d}-\mathbf{b})\cdot(\mathbf{c}-\mathbf{a}),$$

which, on mutliplying out the parentheses, gives

$$\mathbf{d}\cdot\mathbf{c}-\mathbf{a}\cdot\mathbf{c}-\mathbf{d}\cdot\mathbf{b}+\mathbf{a}\cdot\mathbf{b}=\mathbf{d}\cdot\mathbf{c}-\mathbf{b}\cdot\mathbf{c}-\mathbf{d}\cdot\mathbf{a}+\mathbf{b}\cdot\mathbf{a}.$$

Cancelling terms that appear on both sides and rearranging yields

$$\mathbf{d}\cdot\mathbf{b}-\mathbf{d}\cdot\mathbf{a}-\mathbf{c}\cdot\mathbf{b}+\mathbf{c}\cdot\mathbf{a}=0,$$

which simplifies to give

$$(\mathbf{d}-\mathbf{c})\cdot(\mathbf{b}-\mathbf{a})=0.$$

From (7.16), we see that this implies that CD is perpendicular to AB. ◄

If we introduce a set of basis vectors that are mutually orthogonal, such as **i**, **j**, **k**, we can write the components of a vector **a**, with respect to that basis, in terms of the scalar product of **a** with each of the basis vectors, i.e. $a_x = \mathbf{a}\cdot\mathbf{i}$, $a_y = \mathbf{a}\cdot\mathbf{j}$ and $a_z = \mathbf{a}\cdot\mathbf{k}$. In terms of the components a_x, a_y and a_z the scalar product is given by

$$\mathbf{a}\cdot\mathbf{b} = (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k})\cdot(b_x\mathbf{i} + b_y\mathbf{j} + b_z\mathbf{k}) = a_xb_x + a_yb_y + a_zb_z, \qquad (7.21)$$

where the cross terms such as $a_x\mathbf{i}\cdot b_y\mathbf{j}$ are zero because the basis vectors are mutually perpendicular; see equation (7.18). It should be clear from (7.15) that the value of $\mathbf{a}\cdot\mathbf{b}$ has a geometrical definition and that this value is independent of the actual basis vectors used.

►*Find the angle between the vectors* $\mathbf{a} = \mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$ *and* $\mathbf{b} = 2\mathbf{i} + 3\mathbf{j} + 4\mathbf{k}$.

From (7.15) the cosine of the angle θ between **a** and **b** is given by

$$\cos\theta = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}.$$

From (7.21) the scalar product $\mathbf{a}\cdot\mathbf{b}$ has the value

$$\mathbf{a}\cdot\mathbf{b} = 1\times 2 + 2\times 3 + 3\times 4 = 20,$$

and from (7.13) the lengths of the vectors are

$$|\mathbf{a}| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14} \qquad \text{and} \qquad |\mathbf{b}| = \sqrt{2^2 + 3^2 + 4^2} = \sqrt{29}.$$

Thus,

$$\cos\theta = \frac{20}{\sqrt{14}\sqrt{29}} \approx 0.9926 \qquad \Rightarrow \qquad \theta = 0.12 \text{ rad.} ◄$$

We can see from the expressions (7.15) and (7.21) for the scalar product that if θ is the angle between **a** and **b** then

$$\cos\theta = \frac{a_x}{a}\frac{b_x}{b} + \frac{a_y}{a}\frac{b_y}{b} + \frac{a_z}{a}\frac{b_z}{b}$$

where a_x/a, a_y/a and a_z/a are called the *direction cosines* of **a**, since they give the cosine of the angle made by **a** with each of the basis vectors. Similarly b_x/b, b_y/b and b_z/b are the direction cosines of **b**.

If we take the scalar product of any vector **a** with itself then clearly $\theta = 0$ and from (7.15) we have

$$\mathbf{a}\cdot\mathbf{a} = |\mathbf{a}|^2.$$

Thus the magnitude of **a** can be written in a coordinate-independent form as $|\mathbf{a}| = \sqrt{\mathbf{a}\cdot\mathbf{a}}$.

Finally, we note that the scalar product may be extended to vectors with complex components if it is redefined as

$$\mathbf{a}\cdot\mathbf{b} = a_x^*b_x + a_y^*b_y + a_z^*b_z,$$

where the asterisk represents the operation of complex conjugation. To accom-

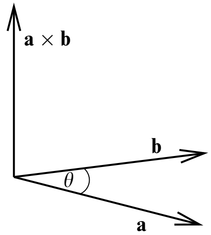


Figure 7.9 The vector product. The vectors **a**, **b** and **a** × **b** form a right-handed set.

modate this extension the commutation property (7.19) must be modified to read

$$\mathbf{a}\cdot\mathbf{b} = (\mathbf{b}\cdot\mathbf{a})^*. \qquad (7.22)$$

In particular it should be noted that $(\lambda\mathbf{a})\cdot\mathbf{b} = \lambda^*\mathbf{a}\cdot\mathbf{b}$, whereas $\mathbf{a}\cdot(\lambda\mathbf{b}) = \lambda\mathbf{a}\cdot\mathbf{b}$. However, the magnitude of a complex vector is still given by $|\mathbf{a}| = \sqrt{\mathbf{a}\cdot\mathbf{a}}$, since $\mathbf{a}\cdot\mathbf{a}$ is always real.

7.6.2 Vector product

The vector product (or cross product) of two vectors **a** and **b** is denoted by $\mathbf{a}\times\mathbf{b}$ and is defined to be a vector of magnitude $|\mathbf{a}||\mathbf{b}|\sin\theta$ in a direction perpendicular to both **a** and **b**;

$$|\mathbf{a}\times\mathbf{b}| = |\mathbf{a}||\mathbf{b}|\,\sin\theta.$$

The direction is found by 'rotating' **a** into **b** through the smallest possible angle. The sense of rotation is that of a right-handed screw that moves forward in the direction $\mathbf{a}\times\mathbf{b}$ (see figure 7.9). Again, θ is the angle between the two vectors placed 'tail to tail' or 'head to head'. With this definition **a**, **b** and $\mathbf{a}\times\mathbf{b}$ form a right-handed set. A more directly usable description of the relative directions in a vector product is provided by a right hand whose first two fingers and thumb are held to be as nearly mutually perpendicular as possible. If the first finger is pointed in the direction of the first vector and the second finger in the direction of the second vector, then the thumb gives the direction of the vector product.

The vector product is distributive over addition, but *anticommutative* and *non-associative*:

$$(\mathbf{a}+\mathbf{b})\times\mathbf{c} = (\mathbf{a}\times\mathbf{c}) + (\mathbf{b}\times\mathbf{c}), \qquad (7.23)$$

$$\mathbf{b}\times\mathbf{a} = -(\mathbf{a}\times\mathbf{b}), \qquad (7.24)$$

$$(\mathbf{a}\times\mathbf{b})\times\mathbf{c} \neq \mathbf{a}\times(\mathbf{b}\times\mathbf{c}). \qquad (7.25)$$

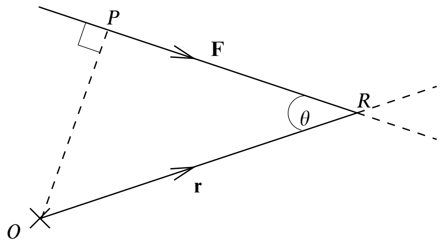


Figure 7.10 The moment of the force **F** about O is $\mathbf{r}\times\mathbf{F}$. The cross represents the direction of $\mathbf{r}\times\mathbf{F}$, which is perpendicularly into the plane of the paper.

From its definition, we see that the vector product has the very useful property that if $\mathbf{a}\times\mathbf{b}=\mathbf{0}$ then **a** is parallel or antiparallel to **b** (unless either of them is zero). We also note that

$$\mathbf{a}\times\mathbf{a}=\mathbf{0}. \tag{7.26}$$

►*Show that if* $\mathbf{a}=\mathbf{b}+\lambda\mathbf{c}$, *for some scalar* λ, *then* $\mathbf{a}\times\mathbf{c}=\mathbf{b}\times\mathbf{c}$.

From (7.23) we have

$$\mathbf{a}\times\mathbf{c}=(\mathbf{b}+\lambda\mathbf{c})\times\mathbf{c}=\mathbf{b}\times\mathbf{c}+\lambda\mathbf{c}\times\mathbf{c}.$$

However, from (7.26), $\mathbf{c}\times\mathbf{c}=\mathbf{0}$ and so

$$\mathbf{a}\times\mathbf{c}=\mathbf{b}\times\mathbf{c}. \tag{7.27}$$

We note in passing that the fact that (7.27) is satisfied does *not* imply that $\mathbf{a}=\mathbf{b}$. ◄

An example of the use of the vector product is that of finding the area, A, of a parallelogram with sides **a** and **b**, using the formula

$$A=|\mathbf{a}\times\mathbf{b}|. \tag{7.28}$$

Another example is afforded by considering a force **F** acting through a point R, whose vector position relative to the origin O is **r** (see figure 7.10). Its *moment* or *torque* about O is the strength of the force times the perpendicular distance OP, which numerically is just $Fr\sin\theta$, i.e. the magnitude of $\mathbf{r}\times\mathbf{F}$. Furthermore, the sense of the moment is clockwise about an axis through O that points perpendicularly into the plane of the paper (the axis is represented by a cross in the figure). Thus the moment is completely represented by the vector $\mathbf{r}\times\mathbf{F}$, in both magnitude and spatial sense. It should be noted that the same vector product is obtained wherever the point R is chosen, so long as it lies on the line of action of **F**.

Similarly, if a solid body is rotating about some axis that passes through the origin, with an angular velocity ω then we can describe this rotation by a vector $\boldsymbol{\omega}$ that has magnitude ω and points along the axis of rotation. The direction of $\boldsymbol{\omega}$

is the forward direction of a right-handed screw rotating in the same sense as the body. The velocity of any point in the body with position vector $\mathbf{r}$ is then given by $\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r}$.

Since the basis vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are mutually perpendicular unit vectors, forming a right-handed set, their vector products are easily seen to be

$$\mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = \mathbf{0}, \tag{7.29}$$

$$\mathbf{i} \times \mathbf{j} = -\mathbf{j} \times \mathbf{i} = \mathbf{k}, \tag{7.30}$$

$$\mathbf{j} \times \mathbf{k} = -\mathbf{k} \times \mathbf{j} = \mathbf{i}, \tag{7.31}$$

$$\mathbf{k} \times \mathbf{i} = -\mathbf{i} \times \mathbf{k} = \mathbf{j}. \tag{7.32}$$

Using these relations, it is straightforward to show that the vector product of two general vectors $\mathbf{a}$ and $\mathbf{b}$ is given in terms of their components with respect to the basis set $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, by

$$\mathbf{a} \times \mathbf{b} = (a_y b_z - a_z b_y)\mathbf{i} + (a_z b_x - a_x b_z)\mathbf{j} + (a_x b_y - a_y b_x)\mathbf{k}. \tag{7.33}$$

For the reader who is familiar with determinants (see chapter 8), we record that this can also be written as

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix}.$$

That the cross product $\mathbf{a} \times \mathbf{b}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$ can be verified in component form by forming its dot products with each of the two vectors and showing that it is zero in both cases.

►*Find the area A of the parallelogram with sides* $\mathbf{a} = \mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$ *and* $\mathbf{b} = 4\mathbf{i} + 5\mathbf{j} + 6\mathbf{k}$.

The vector product $\mathbf{a} \times \mathbf{b}$ is given in component form by

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &= (2 \times 6 - 3 \times 5)\mathbf{i} + (3 \times 4 - 1 \times 6)\mathbf{j} + (1 \times 5 - 2 \times 4)\mathbf{k} \\ &= -3\mathbf{i} + 6\mathbf{j} - 3\mathbf{k}. \end{aligned}$$

Thus the area of the parallelogram is

$$A = |\mathbf{a} \times \mathbf{b}| = \sqrt{(-3)^2 + 6^2 + (-3)^2} = \sqrt{54}.$$ ◄

7.6.3 Scalar triple product

Now that we have defined the scalar and vector products, we can extend our discussion to define products of three vectors. Again, there are two possibilities, the *scalar triple product* and the *vector triple product*.

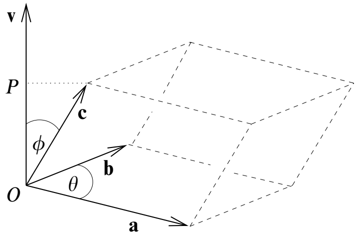


Figure 7.11 The scalar triple product gives the volume of a parallelepiped.

The scalar triple product is denoted by

$$[\mathbf{a}, \mathbf{b}, \mathbf{c}] \equiv \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$$

and, as its name suggests, it is just a number. It is most simply interpreted as the volume of a parallelepiped whose edges are given by **a**, **b** and **c** (see figure 7.11). The vector $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ is perpendicular to the base of the solid and has magnitude $v = ab \sin\theta$, i.e. the area of the base. Further, $\mathbf{v} \cdot \mathbf{c} = vc \cos\phi$. Thus, since $c \cos\phi = OP$ is the vertical height of the parallelepiped, it is clear that $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} =$ area of the base × perpendicular height = volume. It follows that, if the vectors **a**, **b** and **c** are coplanar, $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = 0$.

Expressed in terms of the components of each vector with respect to the Cartesian basis set **i**, **j**, **k** the scalar triple product is

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = a_x(b_y c_z - b_z c_y) + a_y(b_z c_x - b_x c_z) + a_z(b_x c_y - b_y c_x), \tag{7.34}$$

which can also be written as a determinant:

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}.$$

By writing the vectors in component form, it can be shown that

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c},$$

so that the dot and cross symbols can be interchanged without changing the result. More generally, the scalar triple product is unchanged under cyclic permutation of the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Other permutations simply give the negative of the original scalar triple product. These results can be summarised by

$$[\mathbf{a}, \mathbf{b}, \mathbf{c}] = [\mathbf{b}, \mathbf{c}, \mathbf{a}] = [\mathbf{c}, \mathbf{a}, \mathbf{b}] = -[\mathbf{a}, \mathbf{c}, \mathbf{b}] = -[\mathbf{b}, \mathbf{a}, \mathbf{c}] = -[\mathbf{c}, \mathbf{b}, \mathbf{a}]. \tag{7.35}$$

►*Find the volume V of the parallelepiped with sides* $\mathbf{a} = \mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$, $\mathbf{b} = 4\mathbf{i} + 5\mathbf{j} + 6\mathbf{k}$ *and* $\mathbf{c} = 7\mathbf{i} + 8\mathbf{j} + 10\mathbf{k}$.

We have already found that $\mathbf{a} \times \mathbf{b} = -3\mathbf{i} + 6\mathbf{j} - 3\mathbf{k}$, in subsection 7.6.2. Hence the volume of the parallelepiped is given by

$$\begin{aligned} V = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})| &= |(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}| \\ &= |(-3\mathbf{i} + 6\mathbf{j} - 3\mathbf{k}) \cdot (7\mathbf{i} + 8\mathbf{j} + 10\mathbf{k})| \\ &= |(-3)(7) + (6)(8) + (-3)(10)| = 3. \blacktriangleleft \end{aligned}$$

Another useful formula involving both the scalar and vector products is Lagrange's identity (see exercise 7.9), i.e.

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) \equiv (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c}). \tag{7.36}$$

7.6.4 Vector triple product

By the vector triple product of three vectors **a**, **b**, **c** we mean the vector $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$. Clearly, $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ is perpendicular to **a** and lies in the plane of **b** and **c** and so can be expressed in terms of them (see (7.37) below). We note, from (7.25), that the vector triple product is not associative, i.e. $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \neq (\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$.

Two useful formulae involving the vector triple product are

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}, \tag{7.37}$$

$$(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{b} \cdot \mathbf{c})\mathbf{a}, \tag{7.38}$$

which may be derived by writing each vector in component form (see exercise 7.8). It can also be shown that for any three vectors **a**, **b**, **c**,

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) + \mathbf{b} \times (\mathbf{c} \times \mathbf{a}) + \mathbf{c} \times (\mathbf{a} \times \mathbf{b}) = \mathbf{0}.$$

7.7 Equations of lines, planes and spheres

Now that we have described the basic algebra of vectors, we can apply the results to a variety of problems, the first of which is to find the equation of a line in vector form.

7.7.1 Equation of a line

Consider the line passing through the fixed point A with position vector **a** and having a direction **b** (see figure 7.12). It is clear that the position vector **r** of a general point R on the line can be written as

$$\mathbf{r} = \mathbf{a} + \lambda\mathbf{b}, \tag{7.39}$$

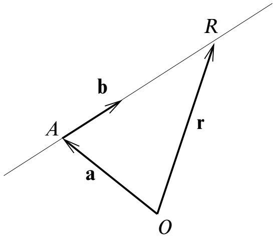


Figure 7.12 The equation of a line. The vector **b** is in the direction AR and $\lambda\mathbf{b}$ is the vector from A to R.

since R can be reached by starting from O, going along the translation vector **a** to the point A on the line and then adding some multiple $\lambda\mathbf{b}$ of the vector **b**. Different values of λ give different points R on the line.

Taking the components of (7.39), we see that the equation of the line can also be written in the form

$$\frac{x - a_x}{b_x} = \frac{y - a_y}{b_y} = \frac{z - a_z}{b_z} = \text{constant}. \tag{7.40}$$

Taking the vector product of (7.39) with **b** and remembering that $\mathbf{b} \times \mathbf{b} = \mathbf{0}$ gives an alternative equation for the line

$$(\mathbf{r} - \mathbf{a}) \times \mathbf{b} = \mathbf{0}.$$

We may also find the equation of the line that passes through two fixed points A and C with position vectors **a** and **c**. Since AC is given by $\mathbf{c} - \mathbf{a}$, the position vector of a general point on the line is

$$\mathbf{r} = \mathbf{a} + \lambda(\mathbf{c} - \mathbf{a}).$$

7.7.2 Equation of a plane

The equation of a plane through a point A with position vector **a** and perpendicular to a unit position vector $\hat{\mathbf{n}}$ (see figure 7.13) is

$$(\mathbf{r} - \mathbf{a}) \cdot \hat{\mathbf{n}} = 0. \tag{7.41}$$

This follows since the vector joining A to a general point R with position vector **r** is $\mathbf{r} - \mathbf{a}$; **r** will lie in the plane if this vector is perpendicular to the normal to the plane. Rewriting (7.41) as $\mathbf{r} \cdot \hat{\mathbf{n}} = \mathbf{a} \cdot \hat{\mathbf{n}}$, we see that the equation of the plane may also be expressed in the form $\mathbf{r} \cdot \hat{\mathbf{n}} = d$, or in component form as

$$lx + my + nz = d, \tag{7.42}$$

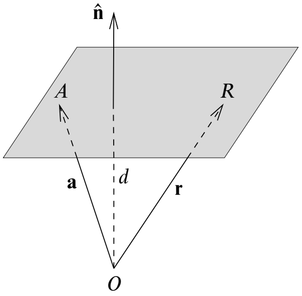


Figure 7.13 The equation of the plane is $(\mathbf{r}-\mathbf{a})\cdot\hat{\mathbf{n}}=0$.

where the unit normal to the plane is $\hat{\mathbf{n}} = l\mathbf{i} + m\mathbf{j} + n\mathbf{k}$ and $d = \mathbf{a}\cdot\hat{\mathbf{n}}$ is the perpendicular distance of the plane from the origin.

The equation of a plane containing points $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ is

$$\mathbf{r} = \mathbf{a} + \lambda(\mathbf{b}-\mathbf{a}) + \mu(\mathbf{c}-\mathbf{a}).$$

This is apparent because starting from the point $\mathbf{a}$ in the plane, all other points may be reached by moving a distance along each of two (non-parallel) directions in the plane. Two such directions are given by $\mathbf{b}-\mathbf{a}$ and $\mathbf{c}-\mathbf{a}$. It can be shown that the equation of this plane may also be written in the more symmetrical form

$$\mathbf{r} = \alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c},$$

where $\alpha+\beta+\gamma = 1$.

►*Find the direction of the line of intersection of the two planes* $x+3y-z=5$ *and* $2x-2y+4z=3$.

The two planes have normal vectors $\mathbf{n}_1 = \mathbf{i}+3\mathbf{j}-\mathbf{k}$ and $\mathbf{n}_2 = 2\mathbf{i}-2\mathbf{j}+4\mathbf{k}$. It is clear that these are not parallel vectors and so the planes must intersect along some line. The direction $\mathbf{p}$ of this line must be parallel to both planes and hence perpendicular to both normals. Therefore

$$\begin{aligned}\mathbf{p} &= \mathbf{n}_1\times\mathbf{n}_2\\ &= [(3)(4)-(-2)(-1)]\,\mathbf{i} + [(-1)(2)-(1)(4)]\,\mathbf{j} + [(1)(-2)-(3)(2)]\,\mathbf{k}\\ &= 10\mathbf{i}-6\mathbf{j}-8\mathbf{k}.\ \blacktriangleleft\end{aligned}$$

7.7.3 Equation of a sphere

Clearly, the defining property of a sphere is that all points on it are equidistant from a fixed point in space and that the common distance is equal to the radius of the sphere. This is easily expressed in vector notation as

$$|\mathbf{r}-\mathbf{c}|^2 = (\mathbf{r}-\mathbf{c})\cdot(\mathbf{r}-\mathbf{c}) = a^2, \tag{7.43}$$

where $\mathbf{c}$ is the position vector of the centre of the sphere and a is its radius.

►*Find the radius ρ of the circle that is the intersection of the plane $\hat{\mathbf{n}}\cdot\mathbf{r} = p$ and the sphere of radius a centred on the point with position vector $\mathbf{c}$.*

The equation of the sphere is

$$|\mathbf{r}-\mathbf{c}|^2 = a^2, \tag{7.44}$$

and that of the circle of intersection is

$$|\mathbf{r}-\mathbf{b}|^2 = \rho^2, \tag{7.45}$$

where $\mathbf{r}$ is restricted to lie in the plane and $\mathbf{b}$ is the position of the circle's centre.

As $\mathbf{b}$ lies on the plane whose normal is $\hat{\mathbf{n}}$, the vector $\mathbf{b}-\mathbf{c}$ must be parallel to $\hat{\mathbf{n}}$, i.e. $\mathbf{b}-\mathbf{c} = \lambda\hat{\mathbf{n}}$ for some λ. Further, by Pythagoras, we must have $\rho^2 + |\mathbf{b}-\mathbf{c}|^2 = a^2$. Thus $\lambda^2 = a^2 - \rho^2$.

Writing $\mathbf{b} = \mathbf{c} + \sqrt{a^2-\rho^2}\,\hat{\mathbf{n}}$ and substituting in (7.45) gives

$$r^2 - 2\mathbf{r}\cdot\left(\mathbf{c} + \sqrt{a^2-\rho^2}\,\hat{\mathbf{n}}\right) + c^2 + 2(\mathbf{c}\cdot\hat{\mathbf{n}})\sqrt{a^2-\rho^2} + a^2 - \rho^2 = \rho^2,$$

whilst, on expansion, (7.44) becomes

$$r^2 - 2\mathbf{r}\cdot\mathbf{c} + c^2 = a^2.$$

Subtracting these last two equations, using $\hat{\mathbf{n}}\cdot\mathbf{r} = p$ and simplifying yields

$$p - \mathbf{c}\cdot\hat{\mathbf{n}} = \sqrt{a^2-\rho^2}.$$

On rearrangement, this gives ρ as $\sqrt{a^2-(p-\mathbf{c}\cdot\hat{\mathbf{n}})^2}$, which places obvious geometrical constraints on the values $a, \mathbf{c}, \hat{\mathbf{n}}$ and p can take if a real intersection between the sphere and the plane is to occur. ◄

7.8 Using vectors to find distances

This section deals with the practical application of vectors to finding distances. Some of these problems are extremely cumbersome in component form, but they all reduce to neat solutions when general vectors, with no explicit basis set, are used. These examples show the power of vectors in simplifying geometrical problems.

7.8.1 Distance from a point to a line

Figure 7.14 shows a line having direction $\mathbf{b}$ that passes through a point A whose position vector is $\mathbf{a}$. To find the *minimum distance* d of the line from a point P whose position vector is $\mathbf{p}$, we must solve the right-angled triangle shown. We see that $d = |\mathbf{p}-\mathbf{a}|\sin\theta$; so, from the definition of the vector product, it follows that

$$d = |(\mathbf{p}-\mathbf{a})\times\hat{\mathbf{b}}|.$$

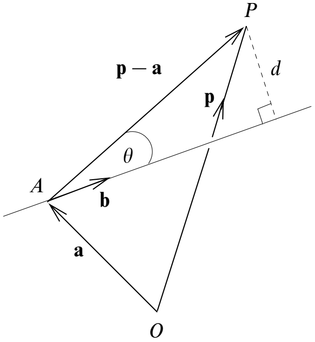


Figure 7.14 The minimum distance from a point to a line.

►*Find the minimum distance from the point P with coordinates* $(1, 2, 1)$ *to the line* $\mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, *where* $\mathbf{a} = \mathbf{i} + \mathbf{j} + \mathbf{k}$ *and* $\mathbf{b} = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k}$.

Comparison with (7.39) shows that the line passes through the point $(1, 1, 1)$ and has direction $2\mathbf{i} - \mathbf{j} + 3\mathbf{k}$. The unit vector in this direction is

$$\hat{\mathbf{b}} = \frac{1}{\sqrt{14}}(2\mathbf{i} - \mathbf{j} + 3\mathbf{k}).$$

The position vector of P is $\mathbf{p} = \mathbf{i} + 2\mathbf{j} + \mathbf{k}$ and we find

$$\begin{aligned}(\mathbf{p} - \mathbf{a}) \times \hat{\mathbf{b}} &= \frac{1}{\sqrt{14}}\,[\,\mathbf{j} \times (2\mathbf{i} - 3\mathbf{j} + 3\mathbf{k})] \\ &= \frac{1}{\sqrt{14}}(3\mathbf{i} - 2\mathbf{k}).\end{aligned}$$

Thus the minimum distance from the line to the point P is $d = \sqrt{13/14}$. ◄

7.8.2 Distance from a point to a plane

The minimum distance d from a point P whose position vector is $\mathbf{p}$ to the plane defined by $(\mathbf{r} - \mathbf{a}) \cdot \hat{\mathbf{n}} = 0$ may be deduced by finding any vector from P to the plane and then determining its component in the normal direction. This is shown in figure 7.15. Consider the vector $\mathbf{a} - \mathbf{p}$, which is a particular vector from P to the plane. Its component normal to the plane, and hence its distance from the plane, is given by

$$d = (\mathbf{a} - \mathbf{p}) \cdot \hat{\mathbf{n}}, \tag{7.46}$$

where the sign of d depends on which side of the plane P is situated.

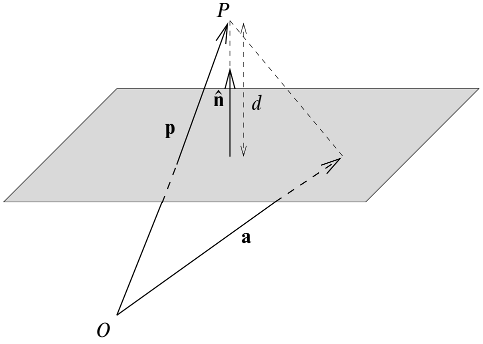


Figure 7.15 The minimum distance d from a point to a plane.

►*Find the distance from the point P with coordinates* $(1, 2, 3)$ *to the plane that contains the points A, B and C having coordinates* $(0, 1, 0)$, $(2, 3, 1)$ *and* $(5, 7, 2)$.

Let us denote the position vectors of the points A, B, C by $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$. Two vectors in the plane are

$$\mathbf{b} - \mathbf{a} = 2\mathbf{i} + 2\mathbf{j} + \mathbf{k} \qquad \text{and} \qquad \mathbf{c} - \mathbf{a} = 5\mathbf{i} + 6\mathbf{j} + 2\mathbf{k},$$

and hence a vector normal to the plane is

$$\mathbf{n} = (2\mathbf{i} + 2\mathbf{j} + \mathbf{k}) \times (5\mathbf{i} + 6\mathbf{j} + 2\mathbf{k}) = -2\mathbf{i} + \mathbf{j} + 2\mathbf{k},$$

and its unit normal is

$$\hat{\mathbf{n}} = \frac{\mathbf{n}}{|\mathbf{n}|} = \tfrac{1}{3}(-2\mathbf{i} + \mathbf{j} + 2\mathbf{k}).$$

Denoting the position vector of P by $\mathbf{p}$, the minimum distance from the plane to P is given by

$$\begin{aligned} d &= (\mathbf{a} - \mathbf{p}) \cdot \hat{\mathbf{n}} \\ &= (-\mathbf{i} - \mathbf{j} - 3\mathbf{k}) \cdot \tfrac{1}{3}(-2\mathbf{i} + \mathbf{j} + 2\mathbf{k}) \\ &= \tfrac{2}{3} - \tfrac{1}{3} - 2 \;=\; -\tfrac{5}{3}. \end{aligned}$$

If we take P to be the origin O, then we find $d = \frac{1}{3}$, i.e. a positive quantity. It follows from this that the original point P with coordinates $(1, 2, 3)$, for which d was negative, is on the opposite side of the plane from the origin. ◄

7.8.3 Distance from a line to a line

Consider two lines in the directions $\mathbf{a}$ and $\mathbf{b}$, as shown in figure 7.16. Since $\mathbf{a} \times \mathbf{b}$ is by definition perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, the unit vector normal to both these lines is

$$\hat{\mathbf{n}} = \frac{\mathbf{a} \times \mathbf{b}}{|\mathbf{a} \times \mathbf{b}|}.$$

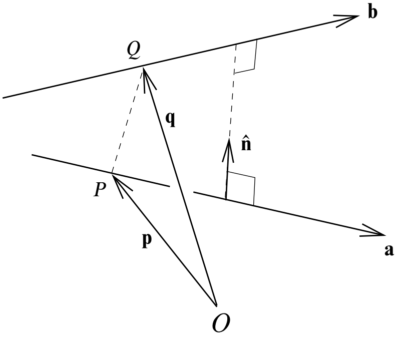


Figure 7.16 The minimum distance from one line to another.

If **p** and **q** are the position vectors of any two points P and Q on different lines then the vector connecting them is $\mathbf{p}-\mathbf{q}$. Thus, the minimum distance d between the lines is this vector's component along the unit normal, i.e.

$$d = |(\mathbf{p}-\mathbf{q})\cdot\hat{\mathbf{n}}|.$$

►*A line is inclined at equal angles to the x-, y- and z-axes and passes through the origin. Another line passes through the points* $(1,2,4)$ *and* $(0,0,1)$. *Find the minimum distance between the two lines.*

The first line is given by

$$\mathbf{r}_1 = \lambda(\mathbf{i}+\mathbf{j}+\mathbf{k}),$$

and the second by

$$\mathbf{r}_2 = \mathbf{k}+\mu(\mathbf{i}+2\mathbf{j}+3\mathbf{k}).$$

Hence a vector normal to both lines is

$$\mathbf{n} = (\mathbf{i}+\mathbf{j}+\mathbf{k})\times(\mathbf{i}+2\mathbf{j}+3\mathbf{k}) = \mathbf{i}-2\mathbf{j}+\mathbf{k},$$

and the unit normal is

$$\hat{\mathbf{n}} = \frac{1}{\sqrt{6}}(\mathbf{i}-2\mathbf{j}+\mathbf{k}).$$

A vector between the two lines is, for example, the one connecting the points $(0,0,0)$ and $(0,0,1)$, which is simply **k**. Thus it follows that the minimum distance between the two lines is

$$d = \frac{1}{\sqrt{6}}|\mathbf{k}\cdot(\mathbf{i}-2\mathbf{j}+\mathbf{k})| = \frac{1}{\sqrt{6}}. ◄$$

7.8.4 Distance from a line to a plane

Let us consider the line $\mathbf{r} = \mathbf{a}+\lambda\mathbf{b}$. This line will intersect any plane to which it is not parallel. Thus, if a plane has a normal $\hat{\mathbf{n}}$ then the minimum distance from the line to the plane is zero unless

$$\mathbf{b}\cdot\hat{\mathbf{n}} = 0,$$

in which case the distance, d, will be

$$d = |(\mathbf{a}-\mathbf{r})\cdot\hat{\mathbf{n}}|,$$

where $\mathbf{r}$ is any point in the plane.

►*A line is given by* $\mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$, *where* $\mathbf{a} = \mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$ *and* $\mathbf{b} = 4\mathbf{i} + 5\mathbf{j} + 6\mathbf{k}$. *Find the coordinates of the point P at which the line intersects the plane*

$$x + 2y + 3z = 6.$$

A vector normal to the plane is

$$\mathbf{n} = \mathbf{i} + 2\mathbf{j} + 3\mathbf{k},$$

from which we find that $\mathbf{b}\cdot\mathbf{n} \neq 0$. Thus the line does indeed intersect the plane. To find the point of intersection we merely substitute the x-, y- and z- values of a general point on the line into the equation of the plane, obtaining

$$1 + 4\lambda + 2(2 + 5\lambda) + 3(3 + 6\lambda) = 6 \quad\Rightarrow\quad 14 + 32\lambda = 6.$$

This gives $\lambda = -\frac{1}{4}$, which we may substitute into the equation for the line to obtain $x = 1 - \frac{1}{4}(4) = 0$, $y = 2 - \frac{1}{4}(5) = \frac{3}{4}$ and $z = 3 - \frac{1}{4}(6) = \frac{3}{2}$. Thus the point of intersection is $(0, \frac{3}{4}, \frac{3}{2})$. ◄

7.9 Reciprocal vectors

The final section of this chapter introduces the concept of reciprocal vectors, which have particular uses in crystallography.

The two sets of vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{a}'$, $\mathbf{b}'$, $\mathbf{c}'$ are called *reciprocal sets* if

$$\mathbf{a}\cdot\mathbf{a}' = \mathbf{b}\cdot\mathbf{b}' = \mathbf{c}\cdot\mathbf{c}' = 1 \tag{7.47}$$

and

$$\mathbf{a}'\cdot\mathbf{b} = \mathbf{a}'\cdot\mathbf{c} = \mathbf{b}'\cdot\mathbf{a} = \mathbf{b}'\cdot\mathbf{c} = \mathbf{c}'\cdot\mathbf{a} = \mathbf{c}'\cdot\mathbf{b} = 0. \tag{7.48}$$

It can be verified (see exercise 7.19) that the reciprocal vectors of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are given by

$$\mathbf{a}' = \frac{\mathbf{b}\times\mathbf{c}}{\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})}, \tag{7.49}$$

$$\mathbf{b}' = \frac{\mathbf{c}\times\mathbf{a}}{\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})}, \tag{7.50}$$

$$\mathbf{c}' = \frac{\mathbf{a}\times\mathbf{b}}{\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})}, \tag{7.51}$$

where $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) \neq 0$. In other words, reciprocal vectors only exist if $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are not coplanar. Moreover, if **a**, **b** and **c** are mutually orthogonal unit vectors then $\mathbf{a}' = \mathbf{a}$, $\mathbf{b}' = \mathbf{b}$ and $\mathbf{c}' = \mathbf{c}$, so that the two systems of vectors are identical.

►*Construct the reciprocal vectors of* $\mathbf{a} = 2\mathbf{i}$, $\mathbf{b} = \mathbf{j} + \mathbf{k}$, $\mathbf{c} = \mathbf{i} + \mathbf{k}$.

First we evaluate the triple scalar product:

$$\begin{aligned}\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) &= 2\mathbf{i} \cdot [(\mathbf{j} + \mathbf{k}) \times (\mathbf{i} + \mathbf{k})] \\ &= 2\mathbf{i} \cdot (\mathbf{i} + \mathbf{j} - \mathbf{k}) \;=\; 2.\end{aligned}$$

Now we find the reciprocal vectors:

$$\begin{aligned}\mathbf{a}' &= \tfrac{1}{2}(\mathbf{j} + \mathbf{k}) \times (\mathbf{i} + \mathbf{k}) \;=\; \tfrac{1}{2}(\mathbf{i} + \mathbf{j} - \mathbf{k}), \\ \mathbf{b}' &= \tfrac{1}{2}(\mathbf{i} + \mathbf{k}) \times 2\mathbf{i} \;=\; \mathbf{j}, \\ \mathbf{c}' &= \tfrac{1}{2}(2\mathbf{i}) \times (\mathbf{j} + \mathbf{k}) \;=\; -\mathbf{j} + \mathbf{k}.\end{aligned}$$

It is easily verified that these reciprocal vectors satisfy their defining properties (7.47), (7.48). ◄

We may also use the concept of reciprocal vectors to define the components of a vector **a** with respect to basis vectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ that are not mutually orthogonal. If the basis vectors are of unit length and mutually orthogonal, such as the Cartesian basis vectors **i**, **j**, **k**, then (see the text preceeding (7.21)) the vector **a** can be written in the form

$$\mathbf{a} = (\mathbf{a} \cdot \mathbf{i})\mathbf{i} + (\mathbf{a} \cdot \mathbf{j})\mathbf{j} + (\mathbf{a} \cdot \mathbf{k})\mathbf{k}.$$

If the basis is not orthonormal, however, then this is no longer true. Nevertheless, we may write the components of **a** with respect to a non-orthonormal basis $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ in terms of its reciprocal basis vectors $\mathbf{e}'_1$, $\mathbf{e}'_2$, $\mathbf{e}'_3$, which are defined as in (7.49)–(7.51). If we let

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3,$$

then the scalar product $\mathbf{a} \cdot \mathbf{e}'_1$ is given by

$$\mathbf{a} \cdot \mathbf{e}'_1 = a_1\mathbf{e}_1 \cdot \mathbf{e}'_1 + a_2\mathbf{e}_2 \cdot \mathbf{e}'_1 + a_3\mathbf{e}_3 \cdot \mathbf{e}'_1 = a_1,$$

where we have used the relations (7.48). Similarly, $a_2 = \mathbf{a}\cdot\mathbf{e}'_2$ and $a_3 = \mathbf{a}\cdot\mathbf{e}'_3$; so now

$$\mathbf{a} = (\mathbf{a} \cdot \mathbf{e}'_1)\mathbf{e}_1 + (\mathbf{a} \cdot \mathbf{e}'_2)\mathbf{e}_2 + (\mathbf{a} \cdot \mathbf{e}'_3)\mathbf{e}_3. \qquad (7.52)$$

7.10 Exercises

7.1 Which of the following statements about general vectors **a**, **b** and **c** are true?

(a) $\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = (\mathbf{b} \times \mathbf{a}) \cdot \mathbf{c}$.
(b) $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$.
(c) $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$.
(d) $\mathbf{d} = \lambda\mathbf{a} + \mu\mathbf{b}$ implies $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{d} = 0$.
(e) $\mathbf{a} \times \mathbf{c} = \mathbf{b} \times \mathbf{c}$ implies $\mathbf{c} \cdot \mathbf{a} - \mathbf{c} \cdot \mathbf{b} = c|\mathbf{a} - \mathbf{b}|$.
(f) $(\mathbf{a} \times \mathbf{b}) \times (\mathbf{c} \times \mathbf{b}) = \mathbf{b}[\mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})]$.

7.2 A unit cell of diamond is a cube of side A, with carbon atoms at each corner, at the centre of each face and, in addition, at positions displaced by $\frac{1}{4}A(\mathbf{i}+\mathbf{j}+\mathbf{k})$ from each of those already mentioned; **i**, **j**, **k** are unit vectors along the cube axes. One corner of the cube is taken as the origin of coordinates. What are the vectors joining the atom at $\frac{1}{4}A(\mathbf{i}+\mathbf{j}+\mathbf{k})$ to its four nearest neighbours? Determine the angle between the carbon bonds in diamond.

7.3 Identify the following surfaces:

(a) $|\mathbf{r}| = k$; (b) $\mathbf{r}\cdot\mathbf{u} = l$; (c) $\mathbf{r}\cdot\mathbf{u} = m|\mathbf{r}|$ for $-1 \le m \le +1$;
(d) $|\mathbf{r} - (\mathbf{r}\cdot\mathbf{u})\mathbf{u}| = n$.

Here k, l, m and n are fixed scalars and **u** is a fixed unit vector.

7.4 Find the angle between the position vectors to the points $(3, -4, 0)$ and $(-2, 1, 0)$ and find the direction cosines of a vector perpendicular to both.

7.5 A, B, C and D are the four corners, in order, of one face of a cube of side 2 units. The opposite face has corners E, F, G and H, with AE, BF, CG and DH as parallel edges of the cube. The centre O of the cube is taken as the origin and the x-, y- and z-axes are parallel to AD, AE and AB, respectively. Find the following:

(a) the angle between the face diagonal AF and the body diagonal AG;
(b) the equation of the plane through B that is parallel to the plane CGE;
(c) the perpendicular distance from the centre J of the face $BCGF$ to the plane OCG;
(d) the volume of the tetrahedron $JOCG$.

7.6 Use vector methods to prove that the lines joining the mid-points of the opposite edges of a tetrahedron $OABC$ meet at a point and that this point bisects each of the lines.

7.7 The edges OP, OQ and OR of a tetrahedron $OPQR$ are vectors **p**, **q** and **r**, respectively, where $\mathbf{p} = 2\mathbf{i} + 4\mathbf{j}$, $\mathbf{q} = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k}$ and $\mathbf{r} = 4\mathbf{i} - 2\mathbf{j} + 5\mathbf{k}$. Show that OP is perpendicular to the plane containing OQR. Express the volume of the tetrahedron in terms of **p**, **q** and **r** and hence calculate the volume.

7.8 Prove, by writing it out in component form, that

$$(\mathbf{a}\times\mathbf{b})\times\mathbf{c} = (\mathbf{a}\cdot\mathbf{c})\mathbf{b} - (\mathbf{b}\cdot\mathbf{c})\mathbf{a},$$

and deduce the result, stated in equation (7.25), that the operation of forming the vector product is non-associative.

7.9 Prove Lagrange's identity, i.e.

$$(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{c}\times\mathbf{d}) = (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{d}) - (\mathbf{a}\cdot\mathbf{d})(\mathbf{b}\cdot\mathbf{c}).$$

7.10 For four arbitrary vectors **a**, **b**, **c** and **d**, evaluate

$$(\mathbf{a}\times\mathbf{b})\times(\mathbf{c}\times\mathbf{d})$$

in two different ways and so prove that

$$\mathbf{a}[\mathbf{b},\mathbf{c},\mathbf{d}] - \mathbf{b}[\mathbf{c},\mathbf{d},\mathbf{a}] + \mathbf{c}[\mathbf{d},\mathbf{a},\mathbf{b}] - \mathbf{d}[\mathbf{a},\mathbf{b},\mathbf{c}] = \mathbf{0}.$$

Show that this reduces to the normal Cartesian representation of the vector **d**, i.e. $d_x\mathbf{i} + d_y\mathbf{j} + d_z\mathbf{k}$, if **a**, **b** and **c** are taken as **i**, **j** and **k**, the Cartesian base vectors.

7.11 Show that the points $(1, 0, 1)$, $(1, 1, 0)$ and $(1, -3, 4)$ lie on a straight line. Give the equation of the line in the form

$$\mathbf{r} = \mathbf{a} + \lambda\mathbf{b}.$$

7.12 The plane P_1 contains the points A, B and C, which have position vectors $\mathbf{a} = -3\mathbf{i} + 2\mathbf{j}$, $\mathbf{b} = 7\mathbf{i} + 2\mathbf{j}$ and $\mathbf{c} = 2\mathbf{i} + 3\mathbf{j} + 2\mathbf{k}$, respectively. Plane P_2 passes through A and is orthogonal to the line BC, whilst plane P_3 passes through B and is orthogonal to the line AC. Find the coordinates of $\mathbf{r}$, the point of intersection of the three planes.

7.13 Two planes have non-parallel unit normals $\hat{\mathbf{n}}$ and $\hat{\mathbf{m}}$ and their closest distances from the origin are λ and μ, respectively. Find the vector equation of their line of intersection in the form $\mathbf{r} = \nu\mathbf{p} + \mathbf{a}$.

7.14 Two fixed points, A and B, in three-dimensional space have position vectors $\mathbf{a}$ and $\mathbf{b}$. Identify the plane P given by

$$(\mathbf{a} - \mathbf{b}) \cdot \mathbf{r} = \tfrac{1}{2}(a^2 - b^2),$$

where a and b are the magnitudes of $\mathbf{a}$ and $\mathbf{b}$.
Show also that the equation

$$(\mathbf{a} - \mathbf{r}) \cdot (\mathbf{b} - \mathbf{r}) = 0$$

describes a sphere S of radius $|\mathbf{a} - \mathbf{b}|/2$. Deduce that the intersection of P and S is also the intersection of two spheres, centred on A and B, and each of radius $|\mathbf{a} - \mathbf{b}|/\sqrt{2}$.

7.15 Let O, A, B and C be four points with position vectors $\mathbf{0}$, $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$, and denote by $\mathbf{g} = \lambda\mathbf{a} + \mu\mathbf{b} + \nu\mathbf{c}$ the position of the centre of the sphere on which they all lie.

(a) Prove that λ, μ and ν simultaneously satisfy

$$(\mathbf{a} \cdot \mathbf{a})\lambda + (\mathbf{a} \cdot \mathbf{b})\mu + (\mathbf{a} \cdot \mathbf{c})\nu = \tfrac{1}{2}a^2$$

and two other similar equations.

(b) By making a change of origin, find the centre and radius of the sphere on which the points $\mathbf{p} = 3\mathbf{i} + \mathbf{j} - 2\mathbf{k}$, $\mathbf{q} = 4\mathbf{i} + 3\mathbf{j} - 3\mathbf{k}$, $\mathbf{r} = 7\mathbf{i} - 3\mathbf{k}$ and $\mathbf{s} = 6\mathbf{i} + \mathbf{j} - \mathbf{k}$ all lie.

7.16 The vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are coplanar and related by

$$\lambda\mathbf{a} + \mu\mathbf{b} + \nu\mathbf{c} = 0,$$

where λ, μ, ν are not all zero. Show that the condition for the points with position vectors $\alpha\mathbf{a}$, $\beta\mathbf{b}$ and $\gamma\mathbf{c}$ to be collinear is

$$\frac{\lambda}{\alpha} + \frac{\mu}{\beta} + \frac{\nu}{\gamma} = 0.$$

7.17 Using vector methods:

(a) Show that the line of intersection of the planes $x + 2y + 3z = 0$ and $3x + 2y + z = 0$ is equally inclined to the x- and z-axes and makes an angle $\cos^{-1}(-2/\sqrt{6})$ with the y-axis.

(b) Find the perpendicular distance between one corner of a unit cube and the major diagonal not passing through it.

7.18 Four points X_i, $i = 1, 2, 3, 4$, taken for simplicity as all lying within the octant $x, y, z \geq 0$, have position vectors $\mathbf{x}_i$. Convince yourself that the direction of vector $\mathbf{x}_n$ lies within the sector of space defined by the directions of the other three vectors if

$$\min_{\text{over } j} \left[\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{|\mathbf{x}_i||\mathbf{x}_j|} \right],$$

considered for $i = 1, 2, 3, 4$ in turn, takes its maximum value for $i = n$, i.e. n equals that value of i for which the largest of the set of angles which $\mathbf{x}_i$ makes with the other vectors, is found to be the lowest. Determine whether any of the four

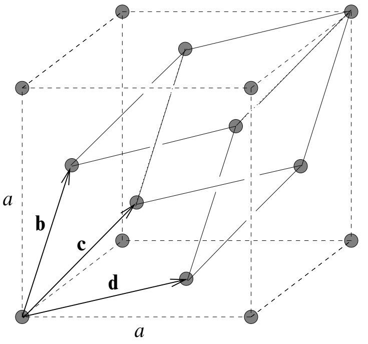


Figure 7.17 A face-centred cubic crystal.

points with coordinates

$$X_1 = (3, 2, 2), \quad X_2 = (2, 3, 1), \quad X_3 = (2, 1, 3), \quad X_4 = (3, 0, 3)$$

lies within the tetrahedron defined by the origin and the other three points.

7.19 The vectors **a**, **b** and **c** are not coplanar. The vectors $\mathbf{a}'$, $\mathbf{b}'$ and $\mathbf{c}'$ are the associated reciprocal vectors. Verify that the expressions (7.49)–(7.51) define a set of reciprocal vectors $\mathbf{a}'$, $\mathbf{b}'$ and $\mathbf{c}'$ with the following properties:

(a) $\mathbf{a}' \cdot \mathbf{a} = \mathbf{b}' \cdot \mathbf{b} = \mathbf{c}' \cdot \mathbf{c} = 1$;
(b) $\mathbf{a}' \cdot \mathbf{b} = \mathbf{a}' \cdot \mathbf{c} = \mathbf{b}' \cdot \mathbf{a} \quad \text{etc} = 0$;
(c) $[\mathbf{a}', \mathbf{b}', \mathbf{c}'] = 1/[\mathbf{a}, \mathbf{b}, \mathbf{c}]$;
(d) $\mathbf{a} = (\mathbf{b}' \times \mathbf{c}')/[\mathbf{a}', \mathbf{b}', \mathbf{c}']$.

7.20 Three non-coplanar vectors **a**, **b** and **c**, have as their respective reciprocal vectors the set $\mathbf{a}'$, $\mathbf{b}'$ and $\mathbf{c}'$. Show that the normal to the plane containing the points $k^{-1}\mathbf{a}$, $l^{-1}\mathbf{b}$ and $m^{-1}\mathbf{c}$ is in the direction of the vector $k\mathbf{a}' + l\mathbf{b}' + m\mathbf{c}'$.

7.21 In a crystal with a face-centred cubic structure, the basic cell can be taken as a cube of edge a with its centre at the origin of coordinates and its edges parallel to the Cartesian coordinate axes; atoms are sited at the eight corners and at the centre of each face. However, other basic cells are possible. One is the rhomboid shown in figure 7.17, which has the three vectors **b**, **c** and **d** as edges.

(a) Show that the volume of the rhomboid is one-quarter that of the cube.
(b) Show that the angles between pairs of edges of the rhomboid are 60° and that the corresponding angles between pairs of edges of the rhomboid defined by the reciprocal vectors to **b**, **c**, **d** are each 109.5°. (This rhomboid can be used as the basic cell of a body-centred cubic structure, more easily visualised as a cube with an atom at each corner and one at its centre.)
(c) In order to use the Bragg formula, $2d \sin\theta = n\lambda$, for the scattering of X-rays by a crystal, it is necessary to know the perpendicular distance d between successive planes of atoms; for a given crystal structure, d has a particular value for each set of planes considered. For the face-centred cubic structure find the distance between successive planes with normals in the **k**, $\mathbf{i} + \mathbf{j}$ and $\mathbf{i} + \mathbf{j} + \mathbf{k}$ directions.

7.22 In subsection 7.6.2 we showed how the moment or torque of a force about an axis could be represented by a vector in the direction of the axis. The magnitude of the vector gives the size of the moment and the sign of the vector gives the sense. Similar representations can be used for angular velocities and angular momenta.

(a) The magnitude of the angular momentum about the origin of a particle of mass m moving with velocity $\mathbf{v}$ on a path that is a perpendicular distance d from the origin is given by $m|\mathbf{v}|d$. Show that if $\mathbf{r}$ is the position of the particle then the vector $\mathbf{J} = \mathbf{r} \times m\mathbf{v}$ represents the angular momentum.

(b) Now consider a rigid collection of particles (or a solid body) rotating about an axis through the origin, the angular velocity of the collection being represented by $\boldsymbol{\omega}$.

(i) Show that the velocity of the ith particle is

$$\mathbf{v}_i = \boldsymbol{\omega} \times \mathbf{r}_i$$

and that the total angular momentum $\mathbf{J}$ is

$$\mathbf{J} = \sum_i m_i[r_i^2\boldsymbol{\omega} - (\mathbf{r}_i \cdot \boldsymbol{\omega})\mathbf{r}_i].$$

(ii) Show further that the component of $\mathbf{J}$ along the axis of rotation can be written as $I\boldsymbol{\omega}$, where I, the moment of inertia of the collection about the axis or rotation, is given by

$$I = \sum_i m_i\rho_i^2.$$

Interpret ρ_i geometrically.

(iii) Prove that the total kinetic energy of the particles is $\frac{1}{2}I\omega^2$.

7.23 By proceeding as indicated below, prove the *parallel axis theorem*, which states that, for a body of mass M, the moment of inertia I about any axis is related to the corresponding moment of inertia I_0 about a parallel axis that passes through the centre of mass of the body by

$$I = I_0 + Ma_\perp^2,$$

where $a_\perp$ is the perpendicular distance between the two axes. Note that I_0 can be written as

$$\int (\hat{\mathbf{n}} \times \mathbf{r}) \cdot (\hat{\mathbf{n}} \times \mathbf{r})\, dm,$$

where $\mathbf{r}$ is the vector position, relative to the centre of mass, of the infinitesimal mass dm and $\hat{\mathbf{n}}$ is a unit vector in the direction of the axis of rotation. Write a similar expression for I in which $\mathbf{r}$ is replaced by $\mathbf{r}' = \mathbf{r} - \mathbf{a}$, where $\mathbf{a}$ is the vector position of any point on the axis to which I refers. Use Lagrange's identity and the fact that $\int \mathbf{r}\, dm = \mathbf{0}$ (by the definition of the centre of mass) to establish the result.

7.24 Without carrying out any further integration, use the results of the previous exercise, the worked example in subsection 6.3.4 and exercise 6.10 to prove that the moment of inertia of a uniform rectangular lamina, of mass M and sides a and b, about an axis perpendicular to its plane and passing through the point $(\alpha a/2, \beta b/2)$, with $-1 \leq \alpha, \beta \leq 1$, is

$$\frac{M}{12}[a^2(1 + 3\alpha^2) + b^2(1 + 3\beta^2)].$$

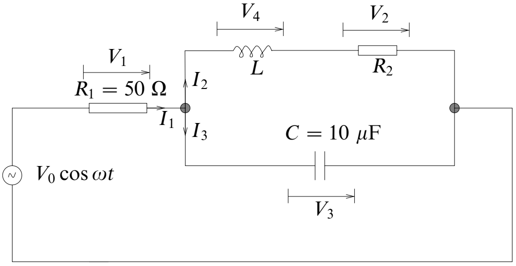


Figure 7.18 An oscillatory electric circuit. The power supply has angular frequency $\omega = 2\pi f = 400\pi\ \text{s}^{-1}$.

7.25 Define a set of (non-orthogonal) base vectors $\mathbf{a} = \mathbf{j} + \mathbf{k}$, $\mathbf{b} = \mathbf{i} + \mathbf{k}$ and $\mathbf{c} = \mathbf{i} + \mathbf{j}$.

(a) Establish their reciprocal vectors and hence express the vectors $\mathbf{p} = 3\mathbf{i} - 2\mathbf{j} + \mathbf{k}$, $\mathbf{q} = \mathbf{i} + 4\mathbf{j}$ and $\mathbf{r} = -2\mathbf{i} + \mathbf{j} + \mathbf{k}$ in terms of the base vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$.

(b) Verify that the scalar product $\mathbf{p} \cdot \mathbf{q}$ has the same value, -5, when evaluated using either set of components.

7.26 Systems that can be modelled as damped harmonic oscillators are widespread; pendulum clocks, car shock absorbers, tuning circuits in television sets and radios, and collective electron motions in plasmas and metals are just a few examples.

In all these cases, one or more variables describing the system obey(s) an equation of the form

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x = P \cos \omega t,$$

where $\dot{x} = dx/dt$, etc. and the inclusion of the factor 2 is conventional. In the steady state (i.e. after the effects of any initial displacement or velocity have been damped out) the solution of the equation takes the form

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

By expressing each term in the form $B \cos(\omega\, t + \epsilon)$, and representing it by a vector of magnitude B making an angle ϵ with the x-axis, draw a closed vector diagram, at $t = 0$, say, that is equivalent to the equation.

(a) Convince yourself that whatever the value of ω (> 0) ϕ must be negative ($-\pi < \phi \leq 0$) and that

$$\phi = \tan^{-1}\left(\frac{-2\gamma\omega}{\omega_0^2 - \omega^2}\right).$$

(b) Obtain an expression for A in terms of P, ω_0 and ω.

7.27 According to alternating current theory, the currents and potential differences in the components of the circuit shown in figure 7.18 are determined by Kirchhoff's laws and the relationships

$$I_1 = \frac{V_1}{R_1}, \qquad I_2 = \frac{V_2}{R_2}, \qquad I_3 = i\omega C V_3, \qquad V_4 = i\omega L I_2.$$

The factor $i = \sqrt{-1}$ in the expression for I_3 indicates that the phase of I_3 is 90° ahead of V_3. Similarly the phase of V_4 is 90° ahead of I_2.

Measurement shows that V_3 has an amplitude of $0.661 V_0$ and a phase of $+13.4°$ relative to that of the power supply. Taking $V_0 = 1\,\text{V}$, and using a series of vector plots for potential differences and currents (they could all be on the same plot if suitable scales were chosen), determine all unknown currents and potential differences and find values for the inductance of L and the resistance of R_2.

[Scales of 1 cm = 0.1 V for potential differences and 1 cm = 1 mA for currents are convenient.]

7.11 Hints and answers

7.1 (c), (d) and (e).

7.3 (a) A sphere of radius k centred on the origin; (b) a plane with its normal in the direction of $\mathbf{u}$ and at a distance l from the origin; (c) a cone with its axis parallel to $\mathbf{u}$ and of semiangle $\cos^{-1} m$; (d) a circular cylinder of radius n with its axis parallel to $\mathbf{u}$.

7.5 (a) $\cos^{-1}\sqrt{2/3}$; (b) $z - x = 2$; (c) $1/\sqrt{2}$; (d) $\frac{1}{3}\frac{1}{2}(\mathbf{c} \times \mathbf{g}) \cdot \mathbf{j} = \frac{1}{3}$.

7.7 Show that $\mathbf{q} \times \mathbf{r}$ is parallel to $\mathbf{p}$; volume $= \frac{1}{3}\left[\frac{1}{2}(\mathbf{q} \times \mathbf{r}) \cdot \mathbf{p}\right] = \frac{5}{3}$.

7.9 Note that $(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = \mathbf{d} \cdot [(\mathbf{a} \times \mathbf{b}) \times \mathbf{c}]$ and use the result for a triple vector product to expand the expression in square brackets.

7.11 Show that the position vectors of the points are linearly dependent; $\mathbf{r} = \mathbf{a} + \lambda\mathbf{b}$ where $\mathbf{a} = \mathbf{i} + \mathbf{k}$ and $\mathbf{b} = -\mathbf{j} + \mathbf{k}$.

7.13 Show that $\mathbf{p}$ must have the direction $\hat{\mathbf{n}} \times \hat{\mathbf{m}}$ and write $\mathbf{a}$ as $x\hat{\mathbf{n}} + y\hat{\mathbf{m}}$. By obtaining a pair of simultaneous equations for x and y, prove that $x = (\lambda - \mu\hat{\mathbf{n}} \cdot \hat{\mathbf{m}})/[1 - (\hat{\mathbf{n}} \cdot \hat{\mathbf{m}})^2]$ and that $y = (\mu - \lambda\hat{\mathbf{n}} \cdot \hat{\mathbf{m}})/[1 - (\hat{\mathbf{n}} \cdot \hat{\mathbf{m}})^2]$.

7.15 (a) Note that $|\mathbf{a} - \mathbf{g}|^2 = R^2 = |\mathbf{0} - \mathbf{g}|^2$, leading to $\mathbf{a} \cdot \mathbf{a} = 2\mathbf{a} \cdot \mathbf{g}$.
(b) Make $\mathbf{p}$ the new origin and solve the three simultaneous linear equations to obtain $\lambda = 5/18$, $\mu = 10/18$, $\nu = -3/18$, giving $\mathbf{g} = 2\mathbf{i} - \mathbf{k}$ and a sphere of radius $\sqrt{5}$ centred on $(5, 1, -3)$.

7.17 (a) Find two points on both planes, say $(0, 0, 0)$ and $(1, -2, 1)$, and hence determine the direction cosines of the line of intersection; (b) $(\frac{2}{3})^{1/2}$.

7.19 For (c) and (d), treat $(\mathbf{c} \times \mathbf{a}) \times (\mathbf{a} \times \mathbf{b})$ as a triple vector product with $\mathbf{c} \times \mathbf{a}$ as one of the three vectors.

7.21 (b) $\mathbf{b}' = a^{-1}(-\mathbf{i} + \mathbf{j} + \mathbf{k})$, $\mathbf{c}' = a^{-1}(\mathbf{i} - \mathbf{j} + \mathbf{k})$, $\mathbf{d}' = a^{-1}(\mathbf{i} + \mathbf{j} - \mathbf{k})$; (c) $a/2$ for direction $\mathbf{k}$; successive planes through $(0, 0, 0)$ and $(a/2, 0, a/2)$ give a spacing of $a/\sqrt{8}$ for direction $\mathbf{i} + \mathbf{j}$; successive planes through $(-a/2, 0, 0)$ and $(a/2, 0, 0)$ give a spacing of $a/\sqrt{3}$ for direction $\mathbf{i} + \mathbf{j} + \mathbf{k}$.

7.23 Note that $a^2 - (\hat{\mathbf{n}} \cdot \mathbf{a})^2 = a_\perp^2$.

7.25 $\mathbf{p} = -2\mathbf{a} + 3\mathbf{b}$, $\mathbf{q} = \frac{3}{2}\mathbf{a} - \frac{3}{2}\mathbf{b} + \frac{5}{2}\mathbf{c}$ and $\mathbf{r} = 2\mathbf{a} - \mathbf{b} - \mathbf{c}$. Remember that $\mathbf{a} \cdot \mathbf{a} = \mathbf{b} \cdot \mathbf{b} = \mathbf{c} \cdot \mathbf{c} = 2$ and $\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} = \mathbf{b} \cdot \mathbf{c} = 1$.

7.27 With currents in mA and potential differences in volts:
$I_1 = (7.76, -23.2^\circ)$, $I_2 = (14.36, -50.8^\circ)$, $I_3 = (8.30, 103.4^\circ)$;
$V_1 = (0.388, -23.2^\circ)$, $V_2 = (0.287, -50.8^\circ)$, $V_4 = (0.596, 39.2^\circ)$;
$L = 33$ mH, $R_2 = 20\ \Omega$.

8

Matrices and vector spaces

In the previous chapter we defined a *vector* as a geometrical object which has both a magnitude and a direction and which may be thought of as an arrow fixed in our familiar three-dimensional space, a space which, if we need to, we define by reference to, say, the fixed stars. This geometrical definition of a vector is both useful and important since it is *independent* of any coordinate system with which we choose to label points in space.

In most specific applications, however, it is necessary at some stage to choose a coordinate system and to break down a vector into its *component vectors* in the directions of increasing coordinate values. Thus for a particular Cartesian coordinate system (for example) the component vectors of a vector **a** will be $a_x\mathbf{i}$, $a_y\mathbf{j}$ and $a_z\mathbf{k}$ and the complete vector will be

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}. \tag{8.1}$$

Although we have so far considered only real three-dimensional space, we may extend our notion of a vector to more abstract spaces, which in general can have an arbitrary number of dimensions N. We may still think of such a vector as an ‘arrow’ in this abstract space, so that it is again *independent* of any (N-dimensional) coordinate system with which we choose to label the space. As an example of such a space, which, though abstract, has very practical applications, we may consider the description of a mechanical or electrical system. If the state of a system is uniquely specified by assigning values to a set of N variables, which could be angles or currents, for example, then that state can be represented by a vector in an N-dimensional space, the vector having those values as its components.

In this chapter we first discuss general *vector spaces* and their properties. We then go on to discuss the transformation of one vector into another by a linear operator. This leads naturally to the concept of a *matrix*, a two-dimensional array of numbers. The properties of matrices are then discussed and we conclude with a discussion of how to use these properties to solve systems of linear equations. The application of matrices to the study of oscillations in physical systems is taken up in chapter 9.

8.1 Vector spaces

A set of objects (vectors) **a**, **b**, **c**, … is said to form a *linear vector space V* if:

(i) the set is closed under commutative and associative addition, so that

$$\mathbf{a}+\mathbf{b}=\mathbf{b}+\mathbf{a}, \tag{8.2}$$

$$(\mathbf{a}+\mathbf{b})+\mathbf{c}=\mathbf{a}+(\mathbf{b}+\mathbf{c}); \tag{8.3}$$

(ii) the set is closed under multiplication by a scalar (any complex number) to form a new vector $\lambda\mathbf{a}$, the operation being both distributive and associative so that

$$\lambda(\mathbf{a}+\mathbf{b})=\lambda\mathbf{a}+\lambda\mathbf{b}, \tag{8.4}$$

$$(\lambda+\mu)\mathbf{a}=\lambda\mathbf{a}+\mu\mathbf{a}, \tag{8.5}$$

$$\lambda(\mu\mathbf{a})=(\lambda\mu)\mathbf{a}, \tag{8.6}$$

where λ and μ are arbitrary scalars;

(iii) there exists a *null vector* **0** such that $\mathbf{a}+\mathbf{0}=\mathbf{a}$ for all **a**;

(iv) multiplication by unity leaves any vector unchanged, i.e. $1\times\mathbf{a}=\mathbf{a}$;

(v) all vectors have a corresponding *negative vector* $-\mathbf{a}$ such that $\mathbf{a}+(-\mathbf{a})=\mathbf{0}$. It follows from (8.5) with $\lambda=1$ and $\mu=-1$ that $-\mathbf{a}$ is the same vector as $(-1)\times\mathbf{a}$.

We note that if we restrict all scalars to be real then we obtain a *real vector space* (an example of which is our familiar three-dimensional space); otherwise, in general, we obtain a *complex vector space*. We note that it is common to use the terms 'vector space' and 'space', instead of the more formal 'linear vector space'.

The *span* of a set of vectors $\mathbf{a},\mathbf{b},\ldots,\mathbf{s}$ is defined as the set of all vectors that may be written as a linear sum of the original set, i.e. all vectors

$$\mathbf{x}=\alpha\mathbf{a}+\beta\mathbf{b}+\cdots+\sigma\mathbf{s} \tag{8.7}$$

that result from the infinite number of possible values of the (in general complex) scalars $\alpha,\beta,\ldots,\sigma$. If **x** in (8.7) is equal to **0** for some choice of $\alpha,\beta,\ldots,\sigma$ (not *all* zero), i.e. if

$$\alpha\mathbf{a}+\beta\mathbf{b}+\cdots+\sigma\mathbf{s}=\mathbf{0}, \tag{8.8}$$

then the set of vectors $\mathbf{a},\mathbf{b},\ldots,\mathbf{s}$, is said to be *linearly dependent*. In such a set at least one vector is redundant, since it can be expressed as a linear sum of the others. If, however, (8.8) is not satisfied by *any* set of coefficients (other than the trivial case in which all the coefficients are zero) then the vectors are *linearly independent*, and no vector in the set can be expressed as a linear sum of the others.

If, in a given vector space, there exist sets of N linearly independent vectors, but no set of $N+1$ linearly independent vectors, then the vector space is said to be N-dimensional. (In this chapter we will limit our discussion to vector spaces of finite dimensionality; spaces of infinite dimensionality are discussed in chapter 17.)

8.1.1 Basis vectors

If V is an N-dimensional vector space then *any* set of N linearly independent vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$ forms a *basis* for V. If $\mathbf{x}$ is an arbitrary vector lying in V then the set of $N+1$ vectors $\mathbf{x}, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$, must be *linearly dependent* and therefore such that

$$\alpha\mathbf{e}_1 + \beta\mathbf{e}_2 + \cdots + \sigma\mathbf{e}_N + \chi\mathbf{x} = \mathbf{0}, \tag{8.9}$$

where the coefficients $\alpha, \beta, \ldots, \chi$ are not all equal to 0, and in particular $\chi \neq 0$. Rearranging (8.9) we may write $\mathbf{x}$ as a linear sum of the vectors $\mathbf{e}_i$ as follows:

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_N\mathbf{e}_N = \sum_{i=1}^{N} x_i\mathbf{e}_i, \tag{8.10}$$

for some set of coefficients x_i that are simply related to the original coefficients, e.g. $x_1 = -\alpha/\chi$, $x_2 = -\beta/\chi$, etc. Since any $\mathbf{x}$ lying in the span of V can be expressed in terms of the *basis* or *base vectors* $\mathbf{e}_i$, the latter are said to form a *complete* set. The coefficients x_i are the *components* of $\mathbf{x}$ with respect to the $\mathbf{e}_i$-basis. These components are *unique*, since if both

$$\mathbf{x} = \sum_{i=1}^{N} x_i\mathbf{e}_i \qquad \text{and} \qquad \mathbf{x} = \sum_{i=1}^{N} y_i\mathbf{e}_i,$$

then

$$\sum_{i=1}^{N}(x_i - y_i)\mathbf{e}_i = \mathbf{0}, \tag{8.11}$$

which, since the $\mathbf{e}_i$ are linearly independent, has only the solution $x_i = y_i$ for all $i = 1, 2, \ldots, N$.

From the above discussion we see that *any* set of N linearly independent vectors can form a basis for an N-dimensional space. If we choose a different set $\mathbf{e}'_i$, $i = 1, \ldots, N$ then we can write $\mathbf{x}$ as

$$\mathbf{x} = x'_1\mathbf{e}'_1 + x'_2\mathbf{e}'_2 + \cdots + x'_N\mathbf{e}'_N = \sum_{i=1}^{N} x'_i\mathbf{e}'_i. \tag{8.12}$$

We reiterate that the vector $\mathbf{x}$ (a geometrical entity) is independent of the basis – it is only the components of $\mathbf{x}$ that depend on the basis. We note, however, that given a set of vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M$, where $M \neq N$, in an N-dimensional vector space, then *either* there exists a vector that cannot be expressed as a linear combination of the $\mathbf{u}_i$ *or*, for some vector that can be so expressed, the components are not unique.

8.1.2 The inner product

We may usefully add to the description of vectors in a vector space by defining the *inner product* of two vectors, denoted in general by $\langle \mathbf{a}|\mathbf{b}\rangle$, which is a scalar function of $\mathbf{a}$ and $\mathbf{b}$. The scalar or dot product, $\mathbf{a}\cdot\mathbf{b} \equiv |\mathbf{a}||\mathbf{b}|\cos\theta$, of vectors in real three-dimensional space (where θ is the angle between the vectors), was introduced in the last chapter and is an example of an inner product. In effect the notion of an inner product $\langle \mathbf{a}|\mathbf{b}\rangle$ is a generalisation of the dot product to more abstract vector spaces. Alternative notations for $\langle \mathbf{a}|\mathbf{b}\rangle$ are $(\mathbf{a}, \mathbf{b})$, or simply $\mathbf{a}\cdot\mathbf{b}$.

The inner product has the following properties:

(i) $\langle \mathbf{a}|\mathbf{b}\rangle = \langle \mathbf{b}|\mathbf{a}\rangle^*$,
(ii) $\langle \mathbf{a}|\lambda\mathbf{b} + \mu\mathbf{c}\rangle = \lambda\langle \mathbf{a}|\mathbf{b}\rangle + \mu\langle \mathbf{a}|\mathbf{c}\rangle$.

We note that in general, for a complex vector space, (i) and (ii) imply that

$$\langle \lambda\mathbf{a} + \mu\mathbf{b}|\mathbf{c}\rangle = \lambda^*\langle \mathbf{a}|\mathbf{c}\rangle + \mu^*\langle \mathbf{b}|\mathbf{c}\rangle, \tag{8.13}$$

$$\langle \lambda\mathbf{a}|\mu\mathbf{b}\rangle = \lambda^*\mu\langle \mathbf{a}|\mathbf{b}\rangle. \tag{8.14}$$

Following the analogy with the dot product in three-dimensional real space, two vectors in a general vector space are defined to be *orthogonal* if $\langle \mathbf{a}|\mathbf{b}\rangle = 0$. Similarly, the *norm* of a vector $\mathbf{a}$ is given by $\|\mathbf{a}\| = \langle \mathbf{a}|\mathbf{a}\rangle^{1/2}$ and is clearly a generalisation of the length or modulus $|\mathbf{a}|$ of a vector $\mathbf{a}$ in three-dimensional space. In a general vector space $\langle \mathbf{a}|\mathbf{a}\rangle$ can be positive or negative; however, we shall be primarily concerned with spaces in which $\langle \mathbf{a}|\mathbf{a}\rangle \geq 0$ and which are thus said to have a *positive semi-definite norm*. In such a space $\langle \mathbf{a}|\mathbf{a}\rangle = 0$ implies $\mathbf{a} = \mathbf{0}$.

Let us now introduce into our N-dimensional vector space a basis $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_N$ that has the desirable property of being *orthonormal* (the basis vectors are mutually orthogonal and each has unit norm), i.e. a basis that has the property

$$\langle \hat{\mathbf{e}}_i|\hat{\mathbf{e}}_j\rangle = \delta_{ij}. \tag{8.15}$$

Here δ_{ij} is the *Kronecker delta* symbol (of which we say more in chapter 26) and has the properties

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

In the above basis we may express any two vectors **a** and **b** as

$$\mathbf{a} = \sum_{i=1}^{N} a_i \hat{\mathbf{e}}_i \qquad \text{and} \qquad \mathbf{b} = \sum_{i=1}^{N} b_i \hat{\mathbf{e}}_i.$$

Furthermore, *in such an orthonormal basis* we have, for any **a**,

$$\langle \hat{\mathbf{e}}_j | \mathbf{a} \rangle = \sum_{i=1}^{N} \langle \hat{\mathbf{e}}_j | a_i \hat{\mathbf{e}}_i \rangle = \sum_{i=1}^{N} a_i \langle \hat{\mathbf{e}}_j | \hat{\mathbf{e}}_i \rangle = a_j. \tag{8.16}$$

Thus the components of **a** are given by $a_i = \langle \hat{\mathbf{e}}_i | \mathbf{a} \rangle$. Note that this is *not* true unless the basis is orthonormal. We can write the inner product of **a** and **b** in terms of their components in an orthonormal basis as

$$\begin{aligned}
\langle \mathbf{a} | \mathbf{b} \rangle &= \langle a_1 \hat{\mathbf{e}}_1 + a_2 \hat{\mathbf{e}}_2 + \cdots + a_N \hat{\mathbf{e}}_N | b_1 \hat{\mathbf{e}}_1 + b_2 \hat{\mathbf{e}}_2 + \cdots + b_N \hat{\mathbf{e}}_N \rangle \\
&= \sum_{i=1}^{N} a_i^* b_i \langle \hat{\mathbf{e}}_i | \hat{\mathbf{e}}_i \rangle + \sum_{i=1}^{N} \sum_{j \neq i}^{N} a_i^* b_j \langle \hat{\mathbf{e}}_i | \hat{\mathbf{e}}_j \rangle \\
&= \sum_{i=1}^{N} a_i^* b_i,
\end{aligned}$$

where the second equality follows from (8.14) and the third from (8.15). This is clearly a generalisation of the expression (7.21) for the dot product of vectors in three-dimensional space.

We may generalise the above to the case where the base vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$ are *not* orthonormal (or orthogonal). In general we can define the N^2 numbers

$$G_{ij} = \langle \mathbf{e}_i | \mathbf{e}_j \rangle. \tag{8.17}$$

Then, if $\mathbf{a} = \sum_{i=1}^{N} a_i \mathbf{e}_i$ and $\mathbf{b} = \sum_{i=1}^{N} b_i \mathbf{e}_i$, the inner product of **a** and **b** is given by

$$\begin{aligned}
\langle \mathbf{a} | \mathbf{b} \rangle &= \left\langle \sum_{i=1}^{N} a_i \mathbf{e}_i \,\middle|\, \sum_{j=1}^{N} b_j \mathbf{e}_j \right\rangle \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i^* b_j \langle \mathbf{e}_i | \mathbf{e}_j \rangle \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i^* G_{ij} b_j.
\end{aligned} \tag{8.18}$$

We further note that from (8.17) and the properties of the inner product we require $G_{ij} = G_{ji}^*$. This in turn ensures that $\|\mathbf{a}\| = \langle \mathbf{a} | \mathbf{a} \rangle$ is real, since then

$$\langle \mathbf{a} | \mathbf{a} \rangle^* = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i G_{ij}^* a_j^* = \sum_{j=1}^{N} \sum_{i=1}^{N} a_j^* G_{ji} a_i = \langle \mathbf{a} | \mathbf{a} \rangle.$$

8.1.3 Some useful inequalities

For a set of objects (vectors) forming a linear vector space in which $\langle \mathbf{a}|\mathbf{a}\rangle \geq 0$ for all $\mathbf{a}$, the following inequalities are often useful.

(i) *Schwarz's inequality* is the most basic result and states that

$$|\langle \mathbf{a}|\mathbf{b}\rangle| \leq \|\mathbf{a}\|\,\|\mathbf{b}\|, \tag{8.19}$$

where the equality holds when $\mathbf{a}$ is a scalar multiple of $\mathbf{b}$, i.e. when $\mathbf{a} = \lambda\mathbf{b}$. It is important here to distinguish between the *absolute value* of a scalar, $|\lambda|$, and the *norm* of a vector, $\|\mathbf{a}\|$. Schwarz's inequality may be proved by considering

$$\begin{aligned}\|\mathbf{a}+\lambda\mathbf{b}\|^2 &= \langle \mathbf{a}+\lambda\mathbf{b}|\mathbf{a}+\lambda\mathbf{b}\rangle \\ &= \langle \mathbf{a}|\mathbf{a}\rangle + \lambda\langle \mathbf{a}|\mathbf{b}\rangle + \lambda^*\langle \mathbf{b}|\mathbf{a}\rangle + \lambda\lambda^*\langle \mathbf{b}|\mathbf{b}\rangle.\end{aligned}$$

If we write $\langle \mathbf{a}|\mathbf{b}\rangle$ as $|\langle \mathbf{a}|\mathbf{b}\rangle|e^{i\alpha}$ then

$$\|\mathbf{a}+\lambda\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + |\lambda|^2\|\mathbf{b}\|^2 + \lambda|\langle \mathbf{a}|\mathbf{b}\rangle|e^{i\alpha} + \lambda^*|\langle \mathbf{a}|\mathbf{b}\rangle|e^{-i\alpha}.$$

However, $\|\mathbf{a}+\lambda\mathbf{b}\|^2 \geq 0$ for all λ, so we may choose $\lambda = re^{-i\alpha}$ and require that, for all r,

$$0 \leq \|\mathbf{a}+\lambda\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + r^2\|\mathbf{b}\|^2 + 2r|\langle \mathbf{a}|\mathbf{b}\rangle|.$$

This means that the quadratic equation in r formed by setting the RHS equal to zero must have no real roots. This, in turn, implies that

$$4|\langle \mathbf{a}|\mathbf{b}\rangle|^2 \leq 4\|\mathbf{a}\|^2\|\mathbf{b}\|^2,$$

which, on taking the square root (all factors are necessarily positive) of both sides, gives Schwarz's inequality.

(ii) The *triangle inequality* states that

$$\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| \tag{8.20}$$

and may be derived from the properties of the inner product and Schwarz's inequality as follows. Let us first consider

$$\|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\ \mathrm{Re}\,\langle \mathbf{a}|\mathbf{b}\rangle \leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2|\langle \mathbf{a}|\mathbf{b}\rangle|.$$

Using Schwarz's inequality we then have

$$\|\mathbf{a}+\mathbf{b}\|^2 \leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{a}\|\,\|\mathbf{b}\| = (\|\mathbf{a}\| + \|\mathbf{b}\|)^2,$$

which, on taking the square root, gives the triangle inequality (8.20).

(iii) *Bessel's inequality* requires the introduction of an orthonormal basis $\hat{\mathbf{e}}_i$, $i = 1, 2, \ldots, N$ into the N-dimensional vector space; it states that

$$\|\mathbf{a}\|^2 \geq \sum_i |\langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle|^2, \tag{8.21}$$

where the equality holds if the sum includes all N basis vectors. If not all the basis vectors are included in the sum then the inequality results (though of course the equality remains if those basis vectors omitted all have $a_i = 0$). Bessel's inequality can also be written

$$\langle \mathbf{a}|\mathbf{a}\rangle \geq \sum_i |a_i|^2,$$

where the a_i are the components of $\mathbf{a}$ in the orthonormal basis. From (8.16) these are given by $a_i = \langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle$. The above may be proved by considering

$$\left\| \mathbf{a} - \sum_i \langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle \hat{\mathbf{e}}_i \right\|^2 = \left\langle \mathbf{a} - \sum_i \langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle \hat{\mathbf{e}}_i \middle| \mathbf{a} - \sum_j \langle \hat{\mathbf{e}}_j|\mathbf{a}\rangle \hat{\mathbf{e}}_j \right\rangle.$$

Expanding out the inner product and using $\langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle^* = \langle \mathbf{a}|\hat{\mathbf{e}}_i\rangle$, we obtain

$$\left\| \mathbf{a} - \sum_i \langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle \hat{\mathbf{e}}_i \right\|^2 = \langle \mathbf{a}|\mathbf{a}\rangle - 2\sum_i \langle \mathbf{a}|\hat{\mathbf{e}}_i\rangle \langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle + \sum_i \sum_j \langle \mathbf{a}|\hat{\mathbf{e}}_i\rangle \langle \hat{\mathbf{e}}_j|\mathbf{a}\rangle \langle \hat{\mathbf{e}}_i|\hat{\mathbf{e}}_j\rangle.$$

Now $\langle \hat{\mathbf{e}}_i|\hat{\mathbf{e}}_j\rangle = \delta_{ij}$, since the basis is orthonormal, and so we find

$$0 \leq \left\| \mathbf{a} - \sum_i \langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle \hat{\mathbf{e}}_i \right\|^2 = \|\mathbf{a}\|^2 - \sum_i |\langle \hat{\mathbf{e}}_i|\mathbf{a}\rangle|^2,$$

which is Bessel's inequality.

We take this opportunity to mention also

(iv) the *parallelogram equality*

$$\|\mathbf{a} + \mathbf{b}\|^2 + \|\mathbf{a} - \mathbf{b}\|^2 = 2\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right), \tag{8.22}$$

which may be proved straightforwardly from the properties of the inner product.

8.2 Linear operators

We now discuss the action of *linear operators* on vectors in a vector space. A linear operator $\mathcal{A}$ associates with every vector $\mathbf{x}$ another vector

$$\mathbf{y} = \mathcal{A}\,\mathbf{x},$$

in such a way that, for two vectors $\mathbf{a}$ and $\mathbf{b}$,

$$\mathcal{A}(\lambda\mathbf{a} + \mu\mathbf{b}) = \lambda\mathcal{A}\,\mathbf{a} + \mu\mathcal{A}\,\mathbf{b},$$

where λ, μ are scalars. We say that $\mathcal{A}$ 'operates' on $\mathbf{x}$ to give the vector $\mathbf{y}$. We note that the action of $\mathcal{A}$ is *independent* of any basis or coordinate system and may be thought of as 'transforming' one geometrical entity (i.e. a vector) into another.

If we now introduce a basis $\mathbf{e}_i$, $i = 1, 2, \ldots, N$, into our vector space then the action of $\mathcal{A}$ on each of the basis vectors is to produce a linear combination of the latter; this may be written as

$$\mathcal{A}\,\mathbf{e}_j = \sum_{i=1}^{N} A_{ij}\mathbf{e}_i, \tag{8.23}$$

where A_{ij} is the ith component of the vector $\mathcal{A}\,\mathbf{e}_j$ in this basis; collectively the numbers A_{ij} are called the components of the linear operator in the $\mathbf{e}_i$-basis. *In this basis* we can express the relation $\mathbf{y} = \mathcal{A}\,\mathbf{x}$ in component form as

$$\mathbf{y} = \sum_{i=1}^{N} y_i\mathbf{e}_i = \mathcal{A}\left(\sum_{j=1}^{N} x_j\mathbf{e}_j\right) = \sum_{j=1}^{N} x_j \sum_{i=1}^{N} A_{ij}\mathbf{e}_i,$$

and hence, in purely component form, in this basis we have

$$y_i = \sum_{j=1}^{N} A_{ij}x_j. \tag{8.24}$$

If we had chosen a different basis $\mathbf{e}'_i$, in which the components of $\mathbf{x}$, $\mathbf{y}$ and $\mathcal{A}$ are x'_i, y'_i and A'_{ij} respectively then the geometrical relationship $\mathbf{y} = \mathcal{A}\,\mathbf{x}$ would be represented in this new basis by

$$y'_i = \sum_{j=1}^{N} A'_{ij}x'_j.$$

We have so far assumed that the vector $\mathbf{y}$ is in the same vector space as $\mathbf{x}$. If, however, $\mathbf{y}$ belongs to a different vector space, which may in general be M-dimensional ($M \neq N$) then the above analysis needs a slight modification. By introducing a basis set $\mathbf{f}_i$, $i = 1, 2, \ldots, M$, into the vector space to which $\mathbf{y}$ belongs we may generalise (8.23) as

$$\mathcal{A}\,\mathbf{e}_j = \sum_{i=1}^{M} A_{ij}\mathbf{f}_i,$$

where the components A_{ij} of the linear operator $\mathcal{A}$ relate to both of the bases $\mathbf{e}_j$ and $\mathbf{f}_i$.

8.2.1 Properties of linear operators

If $\mathbf{x}$ is a vector and $\mathcal{A}$ and $\mathcal{B}$ are two linear operators then it follows that

$$(\mathcal{A} + \mathcal{B})\mathbf{x} = \mathcal{A}\mathbf{x} + \mathcal{B}\mathbf{x},$$
$$(\lambda\mathcal{A})\mathbf{x} = \lambda(\mathcal{A}\mathbf{x}),$$
$$(\mathcal{A}\mathcal{B})\mathbf{x} = \mathcal{A}(\mathcal{B}\mathbf{x}),$$

where in the last equality we see that the action of two linear operators in succession is associative. The product of two linear operators is not in general commutative, however, so that in general $\mathcal{A}\mathcal{B}\mathbf{x} \neq \mathcal{B}\mathcal{A}\mathbf{x}$. In an obvious way we define the null (or zero) and identity operators by

$$\mathcal{O}\mathbf{x} = \mathbf{0} \qquad \text{and} \qquad \mathcal{I}\mathbf{x} = \mathbf{x},$$

for any vector $\mathbf{x}$ in our vector space. Two operators $\mathcal{A}$ and $\mathcal{B}$ are equal if $\mathcal{A}\mathbf{x} = \mathcal{B}\mathbf{x}$ for all vectors $\mathbf{x}$. Finally, if there exists an operator $\mathcal{A}^{-1}$ such that

$$\mathcal{A}\mathcal{A}^{-1} = \mathcal{A}^{-1}\mathcal{A} = \mathcal{I}$$

then $\mathcal{A}^{-1}$ is the *inverse* of $\mathcal{A}$. Some linear operators do not possess an inverse and are called *singular*, whilst those operators that do have an inverse are termed *non-singular*.

8.3 Matrices

We have seen that in a particular basis $\mathbf{e}_i$ both vectors and linear operators can be described in terms of their components with respect to the basis. These components may be displayed as an array of numbers called a *matrix*. In general, if a linear operator $\mathcal{A}$ transforms vectors from an N-dimensional vector space, for which we choose a basis $\mathbf{e}_j$, $j = 1, 2, \ldots, N$, into vectors belonging to an M-dimensional vector space, with basis $\mathbf{f}_i$, $i = 1, 2, \ldots, M$, then we may represent the operator $\mathcal{A}$ by the matrix

$$\mathsf{A} = \begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1N} \\ A_{21} & A_{22} & \ldots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \ldots & A_{MN} \end{pmatrix}. \tag{8.25}$$

The *matrix elements* A_{ij} are the components of the linear operator with respect to the bases $\mathbf{e}_j$ and $\mathbf{f}_i$; the component A_{ij} of the linear operator appears in the ith row and jth column of the matrix. The array has M rows and N columns and is thus called an $M \times N$ matrix. If the dimensions of the two vector spaces are the same, i.e. $M = N$ (for example, if they are the same vector space) then we may represent $\mathcal{A}$ by an $N \times N$ or *square* matrix of *order* N. The component A_{ij}, which in general may be complex, is also denoted by $(\mathsf{A})_{ij}$.

In a similar way we may denote a vector $\mathbf{x}$ in terms of its components x_i in a basis $\mathbf{e}_i$, $i = 1, 2, \ldots, N$, by the array

$$\mathsf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix},$$

which is a special case of (8.25) and is called a *column matrix* (or conventionally, and slightly confusingly, a *column vector* or even just a *vector* – strictly speaking the term 'vector' refers to the geometrical entity $\mathbf{x}$). The column matrix x can also be written as

$$\mathsf{x} = (x_1 \quad x_2 \quad \cdots \quad x_N)^{\mathrm{T}},$$

which is the *transpose* of a *row matrix* (see section 8.6).

We note that in a different basis $\mathbf{e}'_i$ the vector $\mathbf{x}$ would be represented by a *different* column matrix containing the components x'_i in the new basis, i.e.

$$\mathsf{x}' = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_N \end{pmatrix}.$$

Thus, we use x and x' to denote different column matrices which, in different bases $\mathbf{e}_i$ and $\mathbf{e}'_i$, represent the *same* vector $\mathbf{x}$. In many texts, however, this distinction is not made and $\mathbf{x}$ (rather than x) is equated to the corresponding column matrix; if we regard $\mathbf{x}$ as the geometrical entity, however, this can be misleading and so we explicitly make the distinction. A similar argument follows for linear operators; the same linear operator $\mathcal{A}$ is described in different bases by different matrices A and A', containing different matrix elements.

8.4 Basic matrix algebra

The basic algebra of matrices may be deduced from the properties of the linear operators that they represent. In a given basis the action of two linear operators $\mathcal{A}$ and $\mathcal{B}$ on an arbitrary vector $\mathbf{x}$ (see the beginning of subsection 8.2.1), when written in terms of components using (8.24), is given by

$$\sum_j (\mathsf{A}+\mathsf{B})_{ij} x_j = \sum_j A_{ij} x_j + \sum_j B_{ij} x_j,$$

$$\sum_j (\lambda \mathsf{A})_{ij} x_j = \lambda \sum_j A_{ij} x_j,$$

$$\sum_j (\mathsf{AB})_{ij} x_j = \sum_k A_{ik} (\mathsf{Bx})_k = \sum_j \sum_k A_{ik} B_{kj} x_j.$$

Now, since **x** is arbitrary, we can immediately deduce the way in which matrices are added or multiplied, i.e.

$$(\mathsf{A}+\mathsf{B})_{ij} = A_{ij} + B_{ij}, \tag{8.26}$$

$$(\lambda\mathsf{A})_{ij} = \lambda A_{ij}, \tag{8.27}$$

$$(\mathsf{AB})_{ij} = \sum_k A_{ik}B_{kj}. \tag{8.28}$$

We note that a matrix element may, in general, be complex. We now discuss matrix addition and multiplication in more detail.

8.4.1 Matrix addition and multiplication by a scalar

From (8.26) we see that the sum of two matrices, $\mathsf{S} = \mathsf{A} + \mathsf{B}$, is the matrix whose elements are given by

$$S_{ij} = A_{ij} + B_{ij}$$

for every pair of subscripts i, j, with $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$. For example, if A and B are 2×3 matrices then $\mathsf{S} = \mathsf{A} + \mathsf{B}$ is given by

$$\begin{aligned}\begin{pmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \end{pmatrix} &= \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \end{pmatrix} \\ &= \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} & A_{13}+B_{13} \\ A_{21}+B_{21} & A_{22}+B_{22} & A_{23}+B_{23} \end{pmatrix}. \end{aligned} \tag{8.29}$$

Clearly, for the sum of two matrices to have any meaning, the matrices must have the same dimensions, i.e. both be $M \times N$ matrices.

From definition (8.29) it follows that $\mathsf{A} + \mathsf{B} = \mathsf{B} + \mathsf{A}$ and that the sum of a number of matrices can be written unambiguously without bracketting, i.e. matrix addition is *commutative* and *associative*.

The difference of two matrices is defined by direct analogy with addition. The matrix $\mathsf{D} = \mathsf{A} - \mathsf{B}$ has elements

$$D_{ij} = A_{ij} - B_{ij}, \qquad \text{for } i = 1, 2, \ldots, M,\ j = 1, 2, \ldots, N. \tag{8.30}$$

From (8.27) the product of a matrix A with a scalar λ is the matrix with elements λA_{ij}, for example

$$\lambda \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix} = \begin{pmatrix} \lambda A_{11} & \lambda A_{12} & \lambda A_{13} \\ \lambda A_{21} & \lambda A_{22} & \lambda A_{23} \end{pmatrix}. \tag{8.31}$$

Multiplication by a scalar is distributive and associative.

►*The matrices* A, B *and* C *are given by*

$$\mathsf{A} = \begin{pmatrix} 2 & -1 \\ 3 & 1 \end{pmatrix}, \qquad \mathsf{B} = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix}, \qquad \mathsf{C} = \begin{pmatrix} -2 & 1 \\ -1 & 1 \end{pmatrix}.$$

Find the matrix D = A + 2B − C.

$$\begin{aligned} \mathsf{D} &= \begin{pmatrix} 2 & -1 \\ 3 & 1 \end{pmatrix} + 2\begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix} - \begin{pmatrix} -2 & 1 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2+2\times 1-(-2) & -1+2\times 0-1 \\ 3+2\times 0-(-1) & 1+2\times(-2)-1 \end{pmatrix} = \begin{pmatrix} 6 & -2 \\ 4 & -4 \end{pmatrix}. \end{aligned}$$ ◄

From the above considerations we see that the set of all, in general complex, $M \times N$ matrices (with fixed M and N) forms a linear vector space of dimension MN. One basis for the space is the set of $M \times N$ matrices $\mathsf{E}^{(p,q)}$ with the property that $E_{ij}^{(p,q)} = 1$ if $i = p$ and $j = q$ whilst $E_{ij}^{(p,q)} = 0$ for all other values of i and j, i.e. each matrix has only one non-zero entry, which equals unity. Here the pair (p, q) is simply a label that picks out a particular one of the matrices $E^{(p,q)}$, the total number of which is MN.

8.4.2 Multiplication of matrices

Let us consider again the 'transformation' of one vector into another, $\mathbf{y} = \mathcal{A}\,\mathbf{x}$, which, from (8.24), may be described in terms of components with respect to a particular basis as

$$y_i = \sum_{j=1}^{N} A_{ij} x_j \qquad \text{for } i = 1, 2, \ldots, M. \tag{8.32}$$

Writing this in matrix form as $\mathsf{y} = \mathsf{A}\mathsf{x}$ we have

$$\begin{pmatrix} y_1 \\ \boxed{y_2} \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1N} \\ \multicolumn{4}{c}{\boxed{A_{21} \quad A_{22} \quad \ldots \quad A_{2N}}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \ldots & A_{MN} \end{pmatrix} \left(\boxed{\begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{matrix}} \right) \tag{8.33}$$

where we have highlighted with boxes the components used to calculate the element y_2: using (8.32) for $i = 2$,

$$y_2 = A_{21}x_1 + A_{22}x_2 + \cdots + A_{2N}x_N.$$

All the other components y_i are calculated similarly.

If instead we operate with $\mathcal{A}$ on a basis vector $\mathbf{e}_j$ having all components zero except for the jth, which equals unity, then we find

$$\mathsf{A}\mathsf{e}_j = \begin{pmatrix} \begin{matrix} A_{11} & A_{12} & \dots & A_{1N} \end{matrix} \\ \boxed{\begin{matrix} A_{21} & A_{22} & \dots & A_{2N} \end{matrix}} \\ \begin{matrix} \vdots & \vdots & \ddots & \vdots \end{matrix} \\ \begin{matrix} A_{M1} & A_{M2} & \dots & A_{MN} \end{matrix} \end{pmatrix} \begin{pmatrix} \boxed{\begin{matrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{matrix}} \end{pmatrix} = \begin{pmatrix} A_{1j} \\ \boxed{A_{2j}} \\ \vdots \\ A_{Mj} \end{pmatrix},$$

and so confirm our identification of the matrix element A_{ij} as the ith component of $\mathsf{A}\mathsf{e}_j$ in this basis.

From (8.28) we can extend our discussion to the product of two matrices $\mathsf{P} = \mathsf{AB}$, where P is the matrix of the quantities formed by the operation of the rows of A on the columns of B, treating each column of B in turn as the vector $\mathbf{x}$ represented in component form in (8.32). It is clear that, for this to be a meaningful definition, the number of columns in A must equal the number of rows in B. Thus the product AB of an $M \times N$ matrix A with an $N \times R$ matrix B is itself an $M \times R$ matrix P, where

$$P_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj} \quad \text{for } i = 1, 2, \dots, M, \quad j = 1, 2, \dots, R.$$

For example, $\mathsf{P} = \mathsf{AB}$ may be written in matrix form

$$\begin{pmatrix} \boxed{P_{11}} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} \boxed{\begin{matrix} A_{11} & A_{12} & A_{13} \end{matrix}} \\ \begin{matrix} A_{21} & A_{22} & A_{23} \end{matrix} \end{pmatrix} \begin{pmatrix} \boxed{\begin{matrix} B_{11} \\ B_{21} \\ B_{31} \end{matrix}} & \begin{matrix} B_{12} \\ B_{22} \\ B_{32} \end{matrix} \end{pmatrix}$$

where

$$\begin{aligned} P_{11} &= A_{11}B_{11} + A_{12}B_{21} + A_{13}B_{31}, \\ P_{21} &= A_{21}B_{11} + A_{22}B_{21} + A_{23}B_{31}, \\ P_{12} &= A_{11}B_{12} + A_{12}B_{22} + A_{13}B_{32}, \\ P_{22} &= A_{21}B_{12} + A_{22}B_{22} + A_{23}B_{32}. \end{aligned}$$

Multiplication of more than two matrices follows naturally and is associative. So, for example,

$$\mathsf{A}(\mathsf{BC}) \equiv (\mathsf{AB})\mathsf{C}, \tag{8.34}$$

provided, of course, that all the products are defined.

As mentioned above, if A is an $M \times N$ matrix and B is an $N \times M$ matrix then two product matrices are possible, i.e.

$$\mathsf{P} = \mathsf{AB} \qquad \text{and} \qquad \mathsf{Q} = \mathsf{BA}.$$

These are clearly not the same, since P is an $M \times M$ matrix whilst Q is an $N \times N$ matrix. Thus, particular care must be taken to write matrix products in the intended order; $\mathsf{P} = \mathsf{AB}$ but $\mathsf{Q} = \mathsf{BA}$. We note in passing that A^2 means AA, A^3 means $\mathsf{A(AA)} = \mathsf{(AA)A}$ etc. Even if both A and B are square, in general

$$\mathsf{AB} \neq \mathsf{BA}, \tag{8.35}$$

i.e. the multiplication of matrices is not, in general, commutative.

►*Evaluate* $\mathsf{P} = \mathsf{AB}$ *and* $\mathsf{Q} = \mathsf{BA}$ *where*

$$\mathsf{A} = \begin{pmatrix} 3 & 2 & -1 \\ 0 & 3 & 2 \\ 1 & -3 & 4 \end{pmatrix}, \qquad \mathsf{B} = \begin{pmatrix} 2 & -2 & 3 \\ 1 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix}.$$

As we saw for the 2×2 case above, the element P_{ij} of the matrix $\mathsf{P} = \mathsf{AB}$ is found by mentally taking the 'scalar product' of the ith row of A with the jth column of B. For example, $P_{11} = 3 \times 2 + 2 \times 1 + (-1) \times 3 = 5$, $P_{12} = 3 \times (-2) + 2 \times 1 + (-1) \times 2 = -6$, etc. Thus

$$\mathsf{P} = \mathsf{AB} = \begin{pmatrix} 3 & 2 & -1 \\ 0 & 3 & 2 \\ 1 & -3 & 4 \end{pmatrix} \begin{pmatrix} 2 & -2 & 3 \\ 1 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -6 & 8 \\ 9 & 7 & 2 \\ 11 & 3 & 7 \end{pmatrix},$$

and, similarly,

$$\mathsf{Q} = \mathsf{BA} = \begin{pmatrix} 2 & -2 & 3 \\ 1 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 & -1 \\ 0 & 3 & 2 \\ 1 & -3 & 4 \end{pmatrix} = \begin{pmatrix} 9 & -11 & 6 \\ 3 & 5 & 1 \\ 10 & 9 & 5 \end{pmatrix}.$$

These results illustrate that, in general, two matrices do not commute. ◄

The property that matrix multiplication is distributive over addition, i.e. that

$$(\mathsf{A} + \mathsf{B})\mathsf{C} = \mathsf{AC} + \mathsf{BC} \tag{8.36}$$

and

$$\mathsf{C}(\mathsf{A} + \mathsf{B}) = \mathsf{CA} + \mathsf{CB}, \tag{8.37}$$

follows directly from its definition.

8.4.3 The null and identity matrices

Both the null matrix and the identity matrix are frequently encountered, and we take this opportunity to introduce them briefly, leaving their uses until later. The *null* or *zero* matrix 0 has all elements equal to zero, and so its properties are

$$\mathsf{A0} = \mathsf{0} = \mathsf{0A},$$

$$\mathsf{A} + \mathsf{0} = \mathsf{0} + \mathsf{A} = \mathsf{A}.$$

The *identity* matrix I has the property

$$\mathsf{AI} = \mathsf{IA} = \mathsf{A}.$$

It is clear that, in order for the above products to be defined, the identity matrix must be square. The $N \times N$ identity matrix (often denoted by I_N) has the form

$$\mathsf{I}_N = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

8.5 Functions of matrices

If a matrix A is *square* then, as mentioned above, one can define *powers* of A in a straightforward way. For example $\mathsf{A}^2 = \mathsf{AA}$, $\mathsf{A}^3 = \mathsf{AAA}$, or in the general case

$$\mathsf{A}^n = \mathsf{AA} \cdots \mathsf{A} \qquad (n \text{ times}),$$

where n is a positive integer. Having defined powers of a square matrix A, we may construct *functions* of A of the form

$$\mathsf{S} = \sum_n a_n \mathsf{A}^n,$$

where the a_k are simple scalars and the number of terms in the summation may be finite or infinite. In the case where the sum has an infinite number of terms, the sum has meaning only if it converges. A common example of such a function is the *exponential* of a matrix, which is defined by

$$\exp \mathsf{A} = \sum_{n=0}^{\infty} \frac{\mathsf{A}^n}{n!}. \qquad (8.38)$$

This definition can, in turn, be used to define other functions such as $\sin \mathsf{A}$ and $\cos \mathsf{A}$.

8.6 The transpose of a matrix

We have seen that the components of a linear operator in a given coordinate system can be written in the form of a matrix A. We will also find it useful, however, to consider the different (but clearly related) matrix formed by interchanging the rows and columns of A. The matrix is called the *transpose* of A and is denoted by A^{T}.

►*Find the transpose of the matrix*

$$\mathsf{A} = \begin{pmatrix} 3 & 1 & 2 \\ 0 & 4 & 1 \end{pmatrix}.$$

By interchanging the rows and columns of A we immediately obtain

$$\mathsf{A}^{\mathrm{T}} = \begin{pmatrix} 3 & 0 \\ 1 & 4 \\ 2 & 1 \end{pmatrix}. \blacktriangleleft$$

It is obvious that if A is an $M \times N$ matrix then its transpose A^{T} is a $N \times M$ matrix. As mentioned in section 8.3, the transpose of a column matrix is a row matrix and vice versa. An important use of column and row matrices is in the representation of the inner product of two real vectors in terms of their components in a given basis. This notion is discussed fully in the next section, where it is extended to complex vectors.

The transpose of the product of two matrices, $(\mathsf{AB})^{\mathrm{T}}$, is given by the product of their transposes taken in the reverse order, i.e.

$$(\mathsf{AB})^{\mathrm{T}} = \mathsf{B}^{\mathrm{T}}\mathsf{A}^{\mathrm{T}}. \tag{8.39}$$

This is proved as follows:

$$\begin{aligned}(\mathsf{AB})^{\mathrm{T}}_{ij} = (\mathsf{AB})_{ji} &= \sum_k A_{jk}B_{ki} \\ &= \sum_k (\mathsf{A}^{\mathrm{T}})_{kj}(\mathsf{B}^{\mathrm{T}})_{ik} = \sum_k (\mathsf{B}^{\mathrm{T}})_{ik}(\mathsf{A}^{\mathrm{T}})_{kj} = (\mathsf{B}^{\mathrm{T}}\mathsf{A}^{\mathrm{T}})_{ij},\end{aligned}$$

and the proof can be extended to the product of several matrices to give

$$(\mathsf{ABC}\cdots\mathsf{G})^{\mathrm{T}} = \mathsf{G}^{\mathrm{T}}\cdots\mathsf{C}^{\mathrm{T}}\mathsf{B}^{\mathrm{T}}\mathsf{A}^{\mathrm{T}}.$$

8.7 The complex and Hermitian conjugates of a matrix

Two further matrices that can be derived from a given general $M \times N$ matrix are the *complex conjugate*, denoted by A^*, and the *Hermitian conjugate*, denoted by $\mathsf{A}^\dagger$.

The complex conjugate of a matrix A is the matrix obtained by taking the complex conjugate of each of the elements of A, i.e.

$$(\mathsf{A}^*)_{ij} = (A_{ij})^*.$$

Obviously if a matrix is *real* (i.e. it contains only real elements) then $\mathsf{A}^* = \mathsf{A}$.

►*Find the complex conjugate of the matrix*

$$\mathsf{A} = \begin{pmatrix} 1 & 2 & 3i \\ 1+i & 1 & 0 \end{pmatrix}.$$

By taking the complex conjugate of each element we obtain immediately

$$\mathsf{A}^* = \begin{pmatrix} 1 & 2 & -3i \\ 1-i & 1 & 0 \end{pmatrix}. \blacktriangleleft$$

The Hermitian conjugate, or *adjoint*, of a matrix A is the transpose of its complex conjugate, or equivalently, the complex conjugate of its transpose, i.e.

$$\mathsf{A}^\dagger = (\mathsf{A}^*)^{\mathrm{T}} = (\mathsf{A}^{\mathrm{T}})^*.$$

We note that if A is real (and so $\mathsf{A}^* = \mathsf{A}$) then $\mathsf{A}^\dagger = \mathsf{A}^{\mathrm{T}}$, and taking the Hermitian conjugate is equivalent to taking the transpose. Following the previous line of argument for the transpose of the product of several matrices, the Hermitian conjugate of such a product can be shown to be given by

$$(\mathsf{AB}\cdots\mathsf{G})^\dagger = \mathsf{G}^\dagger\cdots\mathsf{B}^\dagger\mathsf{A}^\dagger. \qquad (8.40)$$

►*Find the Hermitian conjugate of the matrix*

$$\mathsf{A} = \begin{pmatrix} 1 & 2 & 3i \\ 1+i & 1 & 0 \end{pmatrix}.$$

Taking the complex conjugate of A and then forming the transpose we find

$$\mathsf{A}^\dagger = \begin{pmatrix} 1 & 1-i \\ 2 & 1 \\ -3i & 0 \end{pmatrix}.$$

We obtain the same result, of course, if we first take the transpose of A and then take the complex conjugate. ◄

An important use of the Hermitian conjugate (or transpose in the real case) is in connection with the inner product of two vectors. Suppose that in a given orthonormal basis the vectors **a** and **b** may be represented by the column matrices

$$\mathsf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} \quad \text{and} \quad \mathsf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix}. \qquad (8.41)$$

Taking the Hermitian conjugate of a, to give a row matrix, and multiplying (on the right) by b we obtain

$$\mathsf{a}^\dagger\mathsf{b} = (a_1^* \; a_2^* \; \cdots \; a_N^*) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix} = \sum_{i=1}^{N} a_i^* b_i, \tag{8.42}$$

which is the expression for the inner product $\langle \mathbf{a}|\mathbf{b}\rangle$ in that basis. We note that for real vectors (8.42) reduces to $\mathsf{a}^{\mathrm{T}}\mathsf{b} = \sum_{i=1}^{N} a_i b_i$.

If the basis $\mathbf{e}_i$ is *not* orthonormal, so that, in general,

$$\langle \mathbf{e}_i|\mathbf{e}_j\rangle = G_{ij} \neq \delta_{ij},$$

then, from (8.18), the scalar product of **a** and **b** in terms of their components with respect to this basis is given by

$$\langle \mathbf{a}|\mathbf{b}\rangle = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i^* G_{ij} b_j = \mathsf{a}^\dagger \mathsf{G}\mathsf{b},$$

where G is the $N \times N$ matrix with elements G_{ij}.

8.8 The trace of a matrix

For a given matrix A, in the previous two sections we have considered various other matrices that can be derived from it. However, sometimes one wishes to derive a single number from a matrix. The simplest example is the *trace* (or *spur*) of a square matrix, which is denoted by $\mathrm{Tr}\,\mathsf{A}$. This quantity is defined as the sum of the diagonal elements of the matrix,

$$\mathrm{Tr}\,\mathsf{A} = A_{11} + A_{22} + \cdots + A_{NN} = \sum_{i=1}^{N} A_{ii}. \tag{8.43}$$

It is clear that taking the trace is a linear operation so that, for example,

$$\mathrm{Tr}(\mathsf{A} \pm \mathsf{B}) = \mathrm{Tr}\,\mathsf{A} \pm \mathrm{Tr}\,\mathsf{B}.$$

A very useful property of traces is that the trace of the product of two matrices is independent of the order of their multiplication; this results holds whether or not the matrices commute and is proved as follows:

$$\mathrm{Tr}\,\mathsf{AB} = \sum_{i=1}^{N} (\mathsf{AB})_{ii} = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}B_{ji} = \sum_{i=1}^{N}\sum_{j=1}^{N} B_{ji}A_{ij} = \sum_{j=1}^{N} (\mathsf{BA})_{jj} = \mathrm{Tr}\,\mathsf{BA}. \tag{8.44}$$

The result can be extended to the product of several matrices. For example, from (8.44), we immediately find

$$\mathrm{Tr}\,\mathsf{ABC} = \mathrm{Tr}\,\mathsf{BCA} = \mathrm{Tr}\,\mathsf{CAB},$$

which shows that the trace of a multiple product is invariant under cyclic permutations of the matrices in the product. Other easily derived properties of the trace are, for example, $\mathrm{Tr}\,\mathsf{A}^{\mathrm{T}} = \mathrm{Tr}\,\mathsf{A}$ and $\mathrm{Tr}\,\mathsf{A}^{\dagger} = (\mathrm{Tr}\,\mathsf{A})^{*}$.

8.9 The determinant of a matrix

For a given matrix A, the determinant det A (like the trace) is a single number (or algebraic expression) that depends upon the elements of A. Also like the trace, the determinant is defined only for *square* matrices. If, for example, A is a 3×3 matrix then its determinant, of *order* 3, is denoted by

$$\det \mathsf{A} = |\mathsf{A}| = \begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix}. \tag{8.45}$$

In order to calculate the value of a determinant, we first need to introduce the notions of the *minor* and the *cofactor* of an element of a matrix. (We shall see that we can use the cofactors to write an order-3 determinant as the weighted sum of three order-2 determinants, thereby simplifying its evaluation.) The minor M_{ij} of the element A_{ij} of an $N \times N$ matrix A is the determinant of the $(N-1) \times (N-1)$ matrix obtained by removing all the elements of the ith row and jth column of A; the associated cofactor, C_{ij}, is found by multiplying the minor by $(-1)^{i+j}$.

►*Find the cofactor of the element A_{23} of the matrix*

$$\mathsf{A} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}.$$

Removing all the elements of the second row and third column of A and forming the determinant of the remaining terms gives the minor

$$M_{23} = \begin{vmatrix} A_{11} & A_{12} \\ A_{31} & A_{32} \end{vmatrix}.$$

Multiplying the minor by $(-1)^{2+3} = (-1)^5 = -1$ gives

$$C_{23} = -\begin{vmatrix} A_{11} & A_{12} \\ A_{31} & A_{32} \end{vmatrix}. \blacktriangleleft$$

We now define a determinant as *the sum of the products of the elements of any row or column and their corresponding cofactors*, e.g. $A_{21}C_{21} + A_{22}C_{22} + A_{23}C_{23}$ or $A_{13}C_{13} + A_{23}C_{23} + A_{33}C_{33}$. Such a sum is called a *Laplace expansion*. For example, in the first of these expansions, using the elements of the second row of the determinant defined by (8.45) and their corresponding cofactors, we write $|\mathsf{A}|$ as the Laplace expansion

$$\begin{aligned}|\mathsf{A}| &= A_{21}(-1)^{(2+1)}M_{21} + A_{22}(-1)^{(2+2)}M_{22} + A_{23}(-1)^{(2+3)}M_{23}\\ &= -A_{21}\begin{vmatrix} A_{12} & A_{13}\\ A_{32} & A_{33}\end{vmatrix} + A_{22}\begin{vmatrix} A_{11} & A_{13}\\ A_{31} & A_{33}\end{vmatrix} - A_{23}\begin{vmatrix} A_{11} & A_{12}\\ A_{31} & A_{32}\end{vmatrix}.\end{aligned}$$

We will see later that the value of the determinant is independent of the row or column chosen. Of course, we have not yet determined the value of $|\mathsf{A}|$ but, rather, written it as the weighted sum of three determinants of order 2. However, applying again the definition of a determinant, we can evaluate each of the order-2 determinants.

►*Evaluate the determinant*

$$\begin{vmatrix} A_{12} & A_{13}\\ A_{32} & A_{33}\end{vmatrix}.$$

By considering the products of the elements of the first row in the determinant, and their corresponding cofactors, we find

$$\begin{aligned}\begin{vmatrix} A_{12} & A_{13}\\ A_{32} & A_{33}\end{vmatrix} &= A_{12}(-1)^{(1+1)}|A_{33}| + A_{13}(-1)^{(1+2)}|A_{32}|\\ &= A_{12}A_{33} - A_{13}A_{32},\end{aligned}$$

where the values of the order-1 determinants $|A_{33}|$ and $|A_{32}|$ are defined to be A_{33} and A_{32} respectively. It must be remembered that the determinant is *not* the same as the modulus, e.g. $\det(-2) = |-2| = -2$, not 2. ◄

We can now combine all the above results to show that the value of the determinant (8.45) is given by

$$\begin{aligned}|\mathsf{A}| &= -A_{21}(A_{12}A_{33} - A_{13}A_{32}) + A_{22}(A_{11}A_{33} - A_{13}A_{31})\\ &\quad - A_{23}(A_{11}A_{32} - A_{12}A_{31}) \qquad (8.46)\\ &= A_{11}(A_{22}A_{33} - A_{23}A_{32}) + A_{12}(A_{23}A_{31} - A_{21}A_{33})\\ &\quad + A_{13}(A_{21}A_{32} - A_{22}A_{31}), \qquad (8.47)\end{aligned}$$

where the final expression gives the form in which the determinant is usually remembered and is the form that is obtained immediately by considering the Laplace expansion using the first row of the determinant. The last equality, which essentially rearranges a Laplace expansion using the second row into one using the first row, supports our assertion that the value of the determinant is unaffected by which row or column is chosen for the expansion.

►*Suppose the rows of a real* 3×3 *matrix* A *are interpreted as the components in a given basis of three (three-component) vectors* **a**, **b** *and* **c**. *Show that one can write the determinant of* A *as*

$$|\mathsf{A}| = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}).$$

If one writes the rows of A as the components in a given basis of three vectors **a**, **b** and **c**, we have from (8.47) that

$$|\mathsf{A}| = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = a_1(b_2c_3 - b_3c_2) + a_2(b_3c_1 - b_1c_3) + a_3(b_1c_2 - b_2c_1).$$

From expression (7.34) for the scalar triple product given in subsection 7.6.3, it follows that we may write the determinant as

$$|\mathsf{A}| = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}). \tag{8.48}$$

In other words, $|\mathsf{A}|$ is the volume of the parallelepiped defined by the vectors **a**, **b** and **c**. (One could equally well interpret the *columns* of the matrix A as the components of three vectors, and result (8.48) would still hold.) This result provides a more memorable (and more meaningful) expression than (8.47) for the value of a 3×3 determinant. Indeed, using this geometrical interpretation, we see immediately that, if the vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ are not linearly independent then the value of the determinant vanishes: $|\mathsf{A}| = 0$. ◄

The evaluation of determinants of order greater than 3 follows the same general method as that presented above, in that it relies on successively reducing the order of the determinant by writing it as a Laplace expansion. Thus, a determinant of order 4 is first written as a sum of four determinants of order 3, which are then evaluated using the above method. For higher-order determinants, one cannot write down directly a simple geometrical expression for $|\mathsf{A}|$ analogous to that given in (8.48). Nevertheless, it is still true that if the rows or columns of the $N \times N$ matrix A are interpreted as the components in a given basis of N (N-component) vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N$, then the determinant $|\mathsf{A}|$ vanishes if these vectors are not all linearly independent.

8.9.1 Properties of determinants

A number of properties of determinants follow straightforwardly from the definition of det A; their use will often reduce the labour of evaluating a determinant. We present them here without specific proofs, though they all follow readily from the alternative form for a determinant, given in equation (26.29) on page 942, and expressed in terms of the Levi–Civita symbol ϵ_{ijk} (see exercise 26.9).

(i) *Determinant of the transpose*. The transpose matrix A^{T} (which, we recall, is obtained by interchanging the rows and columns of A) has the same determinant as A itself, i.e.

$$|\mathsf{A}^{\mathrm{T}}| = |\mathsf{A}|. \tag{8.49}$$

It follows that *any* theorem established for the rows of A will apply to the columns as well, and vice versa.

(ii) *Determinant of the complex and Hermitian conjugate.* It is clear that the matrix A^* obtained by taking the complex conjugate of each element of A has the determinant $|\mathsf{A}^*| = |\mathsf{A}|^*$. Combining this result with (8.49), we find that

$$|\mathsf{A}^\dagger| = |(\mathsf{A}^*)^{\mathrm{T}}| = |\mathsf{A}^*| = |\mathsf{A}|^*. \qquad (8.50)$$

(iii) *Interchanging two rows or two columns.* If two rows (columns) of A are interchanged, its determinant changes sign but is unaltered in magnitude.

(iv) *Removing factors.* If all the elements of a single row (column) of A have a common factor, λ, then this factor may be removed; the value of the determinant is given by the product of the remaining determinant and λ. Clearly this implies that if all the elements of any row (column) are zero then $|\mathsf{A}| = 0$. It also follows that if every element of the $N \times N$ matrix A is multiplied by a constant factor λ then

$$|\lambda\mathsf{A}| = \lambda^N|\mathsf{A}|. \qquad (8.51)$$

(v) *Identical rows or columns.* If any two rows (columns) of A are identical or are multiples of one another, then it can be shown that $|\mathsf{A}| = 0$.

(vi) *Adding a constant multiple of one row (column) to another.* The determinant of a matrix is unchanged in value by adding to the elements of one row (column) any fixed multiple of the elements of another row (column).

(vii) *Determinant of a product.* If A and B are square matrices of the same order then

$$|\mathsf{AB}| = |\mathsf{A}||\mathsf{B}| = |\mathsf{BA}|. \qquad (8.52)$$

A simple extension of this property gives, for example,

$$|\mathsf{AB}\cdots\mathsf{G}| = |\mathsf{A}||\mathsf{B}|\cdots|\mathsf{G}| = |\mathsf{A}||\mathsf{G}|\cdots|\mathsf{B}| = |\mathsf{A}\cdots\mathsf{GB}|,$$

which shows that the determinant is invariant under permutation of the matrices in a multiple product.

There is no explicit procedure for using the above results in the evaluation of any given determinant, and judging the quickest route to an answer is a matter of experience. A general guide is to try to reduce all terms but one in a row or column to zero and hence in effect to obtain a determinant of smaller size. The steps taken in evaluating the determinant in the example below are certainly not the fastest, but they have been chosen in order to illustrate the use of most of the properties listed above.

►*Evaluate the determinant*

$$|\mathsf{A}| = \begin{vmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & -2 & 1 \\ 3 & -3 & 4 & -2 \\ -2 & 1 & -2 & -1 \end{vmatrix}.$$

Taking a factor 2 out of the third column and then adding the second column to the third gives

$$|\mathsf{A}| = 2\begin{vmatrix} 1 & 0 & 1 & 3 \\ 0 & 1 & -1 & 1 \\ 3 & -3 & 2 & -2 \\ -2 & 1 & -1 & -1 \end{vmatrix} = 2\begin{vmatrix} 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 1 \\ 3 & -3 & -1 & -2 \\ -2 & 1 & 0 & -1 \end{vmatrix}.$$

Subtracting the second column from the fourth gives

$$|\mathsf{A}| = 2\begin{vmatrix} 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 0 \\ 3 & -3 & -1 & 1 \\ -2 & 1 & 0 & -2 \end{vmatrix}.$$

We now note that the second row has only one non-zero element and so the determinant may conveniently be written as a Laplace expansion, i.e.

$$|\mathsf{A}| = 2 \times 1 \times (-1)^{2+2}\begin{vmatrix} 1 & 1 & 3 \\ 3 & -1 & 1 \\ -2 & 0 & -2 \end{vmatrix} = 2\begin{vmatrix} 4 & 0 & 4 \\ 3 & -1 & 1 \\ -2 & 0 & -2 \end{vmatrix},$$

where the last equality follows by adding the second row to the first. It can now be seen that the first row is minus twice the third, and so the value of the determinant is zero, by property (v) above. ◄

8.10 The inverse of a matrix

Our first use of determinants will be in defining the *inverse* of a matrix. If we were dealing with ordinary numbers we would consider the relation $\mathsf{P} = \mathsf{AB}$ as equivalent to $\mathsf{B} = \mathsf{P}/\mathsf{A}$, provided that $\mathsf{A} \neq 0$. However, if A, B and P are matrices then this notation does not have an obvious meaning. What we really want to know is whether an explicit formula for B can be obtained in terms of A and P. It will be shown that this is possible for those cases in which $|\mathsf{A}| \neq 0$. A square matrix whose determinant is zero is called a *singular* matrix; otherwise it is *non-singular*. We will show that if A is non-singular we can define a matrix, denoted by A^{-1} and called the *inverse* of A, which has the property that if $\mathsf{AB} = \mathsf{P}$ then $\mathsf{B} = \mathsf{A}^{-1}\mathsf{P}$. In words, B can be obtained by multiplying P from the left by A^{-1}. Analogously, if B is non-singular then, by multiplication from the right, $\mathsf{A} = \mathsf{PB}^{-1}$.

It is clear that

$$\mathsf{AI} = \mathsf{A} \quad \Rightarrow \quad \mathsf{I} = \mathsf{A}^{-1}\mathsf{A}, \tag{8.53}$$

where I is the unit matrix, and so $\mathsf{A}^{-1}\mathsf{A} = \mathsf{I} = \mathsf{AA}^{-1}$. These statements are equivalent to saying that if we first multiply a matrix, B say, by A and then multiply by the inverse A^{-1}, we end up with the matrix we started with, i.e.

$$\mathsf{A}^{-1}\mathsf{A}\mathsf{B} = \mathsf{B}. \tag{8.54}$$

This justifies our use of the term inverse. It is also clear that the inverse is only defined for square matrices.

So far we have only defined what we mean by the inverse of a matrix. Actually finding the inverse of a matrix A may be carried out in a number of ways. We will show that one method is to construct first the matrix C containing the cofactors of the elements of A, as discussed in the last subsection. Then the required inverse A^{-1} can be found by forming the transpose of C and dividing by the determinant of A. Thus the elements of the inverse A^{-1} are given by

$$(\mathsf{A}^{-1})_{ik} = \frac{(\mathsf{C})^{\mathrm{T}}_{ik}}{|\mathsf{A}|} = \frac{C_{ki}}{|\mathsf{A}|}. \tag{8.55}$$

That this procedure does indeed result in the inverse may be seen by considering the components of $\mathsf{A}^{-1}\mathsf{A}$, i.e.

$$(\mathsf{A}^{-1}\mathsf{A})_{ij} = \sum_k (\mathsf{A}^{-1})_{ik}(\mathsf{A})_{kj} = \sum_k \frac{C_{ki}}{|\mathsf{A}|}A_{kj} = \frac{|\mathsf{A}|}{|\mathsf{A}|}\delta_{ij}. \tag{8.56}$$

The last equality in (8.56) relies on the property

$$\sum_k C_{ki}A_{kj} = |\mathsf{A}|\delta_{ij}; \tag{8.57}$$

this can be proved by considering the matrix A' obtained from the original matrix A when the ith column of A is replaced by one of the other columns, say the jth. Thus A' is a matrix with two identical columns and so has zero determinant. However, replacing the ith column by another does not change the cofactors C_{ki} of the elements in the ith column, which are therefore the same in A and A'. Recalling the Laplace expansion of a determinant, i.e.

$$|\mathsf{A}| = \sum_k A_{ki}C_{ki},$$

we obtain

$$0 = |\mathsf{A}'| = \sum_k A'_{ki}C'_{ki} = \sum_k A_{kj}C_{ki}, \qquad i \neq j,$$

which together with the Laplace expansion itself may be summarised by (8.57).

It is immediately obvious from (8.55) that the inverse of a matrix is not defined if the matrix is singular (i.e. if $|\mathsf{A}| = 0$).

►*Find the inverse of the matrix*

$$\mathsf{A} = \begin{pmatrix} 2 & 4 & 3 \\ 1 & -2 & -2 \\ -3 & 3 & 2 \end{pmatrix}.$$

We first determine $|\mathsf{A}|$:

$$\begin{aligned} |\mathsf{A}| &= 2[-2(2) - (-2)3] + 4[(-2)(-3) - (1)(2)] + 3[(1)(3) - (-2)(-3)] \\ &= 11. \end{aligned} \tag{8.58}$$

This is non-zero and so an inverse matrix can be constructed. To do this we need the matrix of the cofactors, C, and hence C^{T}. We find

$$\mathsf{C} = \begin{pmatrix} 2 & 4 & -3 \\ 1 & 13 & -18 \\ -2 & 7 & -8 \end{pmatrix} \quad \text{and} \quad \mathsf{C}^{\mathrm{T}} = \begin{pmatrix} 2 & 1 & -2 \\ 4 & 13 & 7 \\ -3 & -18 & -8 \end{pmatrix},$$

and hence

$$\mathsf{A}^{-1} = \frac{\mathsf{C}^{\mathrm{T}}}{|\mathsf{A}|} = \frac{1}{11}\begin{pmatrix} 2 & 1 & -2 \\ 4 & 13 & 7 \\ -3 & -18 & -8 \end{pmatrix}. ◄ \tag{8.59}$$

For a 2 × 2 matrix, the inverse has a particularly simple form. If the matrix is

$$\mathsf{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

then its determinant $|\mathsf{A}|$ is given by $|\mathsf{A}| = A_{11}A_{22} - A_{12}A_{21}$, and the matrix of cofactors is

$$\mathsf{C} = \begin{pmatrix} A_{22} & -A_{21} \\ -A_{12} & A_{11} \end{pmatrix}.$$

Thus the inverse of A is given by

$$\mathsf{A}^{-1} = \frac{\mathsf{C}^{\mathrm{T}}}{|\mathsf{A}|} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}}\begin{pmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{pmatrix}. \tag{8.60}$$

It can be seen that the transposed matrix of cofactors for a 2 × 2 matrix is the same as the matrix formed by swapping the elements on the leading diagonal (A_{11} and A_{22}) and changing the signs of the other two elements (A_{12} and A_{21}). This is completely general for a 2 × 2 matrix and is easy to remember.

The following are some further useful properties related to the inverse matrix and may be straightforwardly derived.

(i) $(\mathsf{A}^{-1})^{-1} = \mathsf{A}$.
(ii) $(\mathsf{A}^{\mathrm{T}})^{-1} = (\mathsf{A}^{-1})^{\mathrm{T}}$.
(iii) $(\mathsf{A}^{\dagger})^{-1} = (\mathsf{A}^{-1})^{\dagger}$.
(iv) $(\mathsf{AB})^{-1} = \mathsf{B}^{-1}\mathsf{A}^{-1}$.
(v) $(\mathsf{AB}\cdots\mathsf{G})^{-1} = \mathsf{G}^{-1}\cdots\mathsf{B}^{-1}\mathsf{A}^{-1}$.

►*Prove the properties* (i)–(v) *stated above.*

We begin by writing down the fundamental expression defining the inverse of a non-singular square matrix A:

$$\mathsf{AA}^{-1} = \mathsf{I} = \mathsf{A}^{-1}\mathsf{A}. \tag{8.61}$$

Property (i). This follows immediately from the expression (8.61).

Property (ii). Taking the transpose of each expression in (8.61) gives

$$(\mathsf{AA}^{-1})^{\mathrm{T}} = \mathsf{I}^{\mathrm{T}} = (\mathsf{A}^{-1}\mathsf{A})^{\mathrm{T}}.$$

Using the result (8.39) for the transpose of a product of matrices and noting that $\mathsf{I}^{\mathrm{T}} = \mathsf{I}$, we find

$$(\mathsf{A}^{-1})^{\mathrm{T}}\mathsf{A}^{\mathrm{T}} = \mathsf{I} = \mathsf{A}^{\mathrm{T}}(\mathsf{A}^{-1})^{\mathrm{T}}.$$

However, from (8.61), this implies $(\mathsf{A}^{-1})^{\mathrm{T}} = (\mathsf{A}^{\mathrm{T}})^{-1}$ and hence proves result (ii) above.

Property (iii). This may be proved in an analogous way to property (ii), by replacing the transposes in (ii) by Hermitian conjugates and using the result (8.40) for the Hermitian conjugate of a product of matrices.

Property (iv). Using (8.61), we may write

$$(\mathsf{AB})(\mathsf{AB})^{-1} = \mathsf{I} = (\mathsf{AB})^{-1}(\mathsf{AB}),$$

From the left-hand equality it follows, by multiplying on the left by A^{-1}, that

$$\mathsf{A}^{-1}\mathsf{AB}(\mathsf{AB})^{-1} = \mathsf{A}^{-1}\mathsf{I} \qquad \text{and hence} \qquad \mathsf{B}(\mathsf{AB})^{-1} = \mathsf{A}^{-1}.$$

Now multiplying on the left by B^{-1} gives

$$\mathsf{B}^{-1}\mathsf{B}(\mathsf{AB})^{-1} = \mathsf{B}^{-1}\mathsf{A}^{-1},$$

and hence the stated result.

Property (v). Finally, result (iv) may extended to case (v) in a straightforward manner. For example, using result (iv) twice we find

$$(\mathsf{ABC})^{-1} = (\mathsf{BC})^{-1}\mathsf{A}^{-1} = \mathsf{C}^{-1}\mathsf{B}^{-1}\mathsf{A}^{-1}. \blacktriangleleft$$

We conclude this section by noting that the determinant $|\mathsf{A}^{-1}|$ of the inverse matrix can be expressed very simply in terms of the determinant $|\mathsf{A}|$ of the matrix itself. Again we start with the fundamental expression (8.61). Then, using the property (8.52) for the determinant of a product, we find

$$|\mathsf{AA}^{-1}| = |\mathsf{A}||\mathsf{A}^{-1}| = |\mathsf{I}|.$$

It is straightforward to show by Laplace expansion that $|\mathsf{I}| = 1$, and so we arrive at the useful result

$$|\mathsf{A}^{-1}| = \frac{1}{|\mathsf{A}|}. \tag{8.62}$$

8.11 The rank of a matrix

The *rank* of a general $M \times N$ matrix is an important concept, particularly in the solution of sets of simultaneous linear equations, to be discussed in the next section, and we now discuss it in some detail. Like the trace and determinant, the rank of matrix A is a single number (or algebraic expression) that depends on the elements of A. Unlike the trace and determinant, however, the rank of a matrix can be defined even when A is not square. As we shall see, there are two *equivalent* definitions of the rank of a general matrix.

Firstly, the rank of a matrix may be defined in terms of the *linear independence* of vectors. Suppose that the columns of an $M \times N$ matrix are interpreted as the components in a given basis of N (M-component) vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$, as follows:

$$\mathsf{A} = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \mathsf{v}_1 & \mathsf{v}_2 & \ldots & \mathsf{v}_N \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}.$$

Then the *rank* of A, denoted by rank A or by $R(\mathsf{A})$, is defined as the number of *linearly independent* vectors in the set $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$, and equals the dimension of the vector space spanned by those vectors. Alternatively, we may consider the rows of A to contain the components in a given basis of the M (N-component) vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$ as follows:

$$\mathsf{A} = \begin{pmatrix} \leftarrow & \mathsf{w}_1 & \rightarrow \\ \leftarrow & \mathsf{w}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathsf{w}_M & \rightarrow \end{pmatrix}.$$

It may then be shown[§] that the rank of A is also equal to the number of linearly independent vectors in the set $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$. From this definition it is should be clear that the rank of A is unaffected by the exchange of two rows (or two columns) or by the multiplication of a row (or column) by a constant. Furthermore, suppose that a constant multiple of one row (column) is added to another row (column): for example, we might replace the row w_i by $\mathsf{w}_i + c\mathsf{w}_j$. This also has no effect on the number of linearly independent rows and so leaves the rank of A unchanged. We may use these properties to evaluate the rank of a given matrix.

A second (equivalent) definition of the rank of a matrix may be given and uses the concept of *submatrices*. A submatrix of A is any matrix that can be formed from the elements of A by ignoring one, or more than one, row or column. It

[§] For a fuller discussion, see, for example, C. D. Cantrell, *Modern Mathematical Methods for Physicists and Engineers* (Cambridge: Cambridge University Press, 2000), chapter 6.

may be shown that the rank of a general $M \times N$ matrix is equal to the size of the largest square submatrix of A whose determinant is non-zero. Therefore, if a matrix A has an $r \times r$ submatrix S with $|\mathsf{S}| \neq 0$, but no $(r+1) \times (r+1)$ submatrix with non-zero determinant then the rank of the matrix is r. From either definition it is clear that the rank of A is less than or equal to the smaller of M and N.

►*Determine the rank of the matrix*

$$\mathsf{A} = \begin{pmatrix} 1 & 1 & 0 & -2 \\ 2 & 0 & 2 & 2 \\ 4 & 1 & 3 & 1 \end{pmatrix}.$$

The largest possible square submatrices of A must be of dimension 3×3. Clearly, A possesses four such submatrices, the determinants of which are given by

$$\begin{vmatrix} 1 & 1 & 0 \\ 2 & 0 & 2 \\ 4 & 1 & 3 \end{vmatrix} = 0, \qquad \begin{vmatrix} 1 & 1 & -2 \\ 2 & 0 & 2 \\ 4 & 1 & 1 \end{vmatrix} = 0,$$

$$\begin{vmatrix} 1 & 0 & -2 \\ 2 & 2 & 2 \\ 4 & 3 & 1 \end{vmatrix} = 0, \qquad \begin{vmatrix} 1 & 0 & -2 \\ 0 & 2 & 2 \\ 1 & 3 & 1 \end{vmatrix} = 0.$$

(In each case the determinant may be evaluated as described in subsection 8.9.1.)

The next largest square submatrices of A are of dimension 2×2. Consider, for example, the 2×2 submatrix formed by ignoring the third row and the third and fourth columns of A; this has determinant

$$\begin{vmatrix} 1 & 1 \\ 2 & 0 \end{vmatrix} = 1 \times 0 - 2 \times 1 = -2.$$

Since its determinant is non-zero, A is of rank 2 and we need not consider any other 2×2 submatrix. ◄

In the special case in which the matrix A is a *square* $N \times N$ matrix, by comparing either of the above definitions of rank with our discussion of determinants in section 8.9, we see that $|\mathsf{A}| = 0$ unless the rank of A is N. In other words, A is *singular* unless $R(\mathsf{A}) = N$.

8.12 Special types of square matrix

Matrices that are square, i.e. $N \times N$, are very common in physical applications. We now consider some special forms of square matrix that are of particular importance.

8.12.1 Diagonal matrices

The unit matrix, which we have already encountered, is an example of a *diagonal* matrix. Such matrices are characterised by having non-zero elements only on the *leading diagonal*, i.e. only elements A_{ij} with $i = j$ may be non-zero. For example,

$$\mathsf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{pmatrix},$$

is a 3×3 diagonal matrix. Such a matrix is often denoted by $\mathsf{A} = \text{diag}\,(1, 2, -3)$. By performing a Laplace expansion, it is easily shown that the determinant of an $N \times N$ diagonal matrix is equal to the product of the diagonal elements. Thus, if the matrix has the form $\mathsf{A} = \text{diag}(A_{11}, A_{22}, \ldots, A_{NN})$ then

$$|\mathsf{A}| = A_{11}A_{22}\cdots A_{NN}. \tag{8.63}$$

Moreover, it is also straightforward to show that the inverse of A is also a diagonal matrix given by

$$\mathsf{A}^{-1} = \text{diag}\left(\frac{1}{A_{11}}, \frac{1}{A_{22}}, \ldots, \frac{1}{A_{NN}}\right).$$

Finally, we note that, if two matrices A and B are *both* diagonal then they have the useful property that their product is commutative:

$$\mathsf{AB} = \mathsf{BA}.$$

This is *not* true for matrices in general.

8.12.2 Lower and upper triangular matrices

A square matrix A is called *lower triangular* if all the elements *above* the principal diagonal are zero. For example, the general form for a 3×3 lower triangular matrix is

$$\mathsf{A} = \begin{pmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & 0 \\ A_{31} & A_{32} & A_{33} \end{pmatrix},$$

where the elements A_{ij} may be zero or non-zero. Similarly an *upper triangular* square matrix is one for which all the elements *below* the principal diagonal are zero. The general 3×3 form is thus

$$\mathsf{A} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}.$$

By performing a Laplace expansion, it is straightforward to show that, in the general $N \times N$ case, the determinant of an upper or lower triangular matrix is equal to the product of its diagonal elements,

$$|\mathsf{A}| = A_{11}A_{22}\cdots A_{NN}. \tag{8.64}$$

Clearly result (8.63) for diagonal matrices is a special case of this result. Moreover, it may be shown that the inverse of a non-singular lower (upper) triangular matrix is also lower (upper) triangular.

8.12.3 Symmetric and antisymmetric matrices

A square matrix A of order N with the property $\mathsf{A} = \mathsf{A}^{\mathrm{T}}$ is said to be *symmetric*. Similarly a matrix for which $\mathsf{A} = -\mathsf{A}^{\mathrm{T}}$ is said to be *anti-* or *skew*-symmetric and its diagonal elements $a_{11}, a_{22}, \ldots, a_{NN}$ are necessarily zero. Moreover, if A is (anti-)symmetric then so too is its inverse A^{-1}. This is easily proved by noting that if $\mathsf{A} = \pm\mathsf{A}^{\mathrm{T}}$ then

$$(\mathsf{A}^{-1})^{\mathrm{T}} = (\mathsf{A}^{\mathrm{T}})^{-1} = \pm\mathsf{A}^{-1}.$$

Any $N \times N$ matrix A can be written as the sum of a symmetric and an antisymmetric matrix, since we may write

$$\mathsf{A} = \tfrac{1}{2}(\mathsf{A} + \mathsf{A}^{\mathrm{T}}) + \tfrac{1}{2}(\mathsf{A} - \mathsf{A}^{\mathrm{T}}) = \mathsf{B} + \mathsf{C},$$

where clearly $\mathsf{B} = \mathsf{B}^{\mathrm{T}}$ and $\mathsf{C} = -\mathsf{C}^{\mathrm{T}}$. The matrix B is therefore called the symmetric part of A, and C is the antisymmetric part.

►*If* A *is an* $N \times N$ *antisymmetric matrix, show that* $|\mathsf{A}| = 0$ *if* N *is odd.*

If A is antisymmetric then $\mathsf{A}^{\mathrm{T}} = -\mathsf{A}$. Using the properties of determinants (8.49) and (8.51), we have

$$|\mathsf{A}| = |\mathsf{A}^{\mathrm{T}}| = |-\mathsf{A}| = (-1)^N|\mathsf{A}|.$$

Thus, if N is odd then $|\mathsf{A}| = -|\mathsf{A}|$, which implies that $|\mathsf{A}| = 0$. ◄

8.12.4 Orthogonal matrices

A non-singular matrix with the property that its transpose is also its inverse,

$$\mathsf{A}^{\mathrm{T}} = \mathsf{A}^{-1}, \qquad (8.65)$$

is called an *orthogonal matrix*. It follows immediately that the inverse of an orthogonal matrix is also orthogonal, since

$$(\mathsf{A}^{-1})^{\mathrm{T}} = (\mathsf{A}^{\mathrm{T}})^{-1} = (\mathsf{A}^{-1})^{-1}.$$

Moreover, since for an orthogonal matrix $\mathsf{A}^{\mathrm{T}}\mathsf{A} = \mathsf{I}$, we have

$$|\mathsf{A}^{\mathrm{T}}\mathsf{A}| = |\mathsf{A}^{\mathrm{T}}||\mathsf{A}| = |\mathsf{A}|^2 = |\mathsf{I}| = 1.$$

Thus the determinant of an orthogonal matrix must be $|\mathsf{A}| = \pm 1$.

An orthogonal matrix represents, in a particular basis, a linear operator that leaves the norms (lengths) of real vectors unchanged, as we will now show.

Suppose that $\mathbf{y} = \mathcal{A}\mathbf{x}$ is represented in some coordinate system by the matrix equation $\mathsf{y} = \mathsf{A}\mathsf{x}$; then $\langle \mathbf{y}|\mathbf{y}\rangle$ is given in this coordinate system by

$$\mathsf{y}^{\mathrm{T}}\mathsf{y} = \mathsf{x}^{\mathrm{T}}\mathsf{A}^{\mathrm{T}}\mathsf{A}\mathsf{x} = \mathsf{x}^{\mathrm{T}}\mathsf{x}.$$

Hence $\langle \mathbf{y}|\mathbf{y}\rangle = \langle \mathbf{x}|\mathbf{x}\rangle$, showing that the action of a linear operator represented by an orthogonal matrix does not change the norm of a real vector.

8.12.5 Hermitian and anti-Hermitian matrices

An *Hermitian* matrix is one that satisfies $\mathsf{A} = \mathsf{A}^\dagger$, where $\mathsf{A}^\dagger$ is the Hermitian conjugate discussed in section 8.7. Similarly if $\mathsf{A}^\dagger = -\mathsf{A}$, then A is called *anti-Hermitian*. A real (anti-)symmetric matrix is a special case of an (anti-)Hermitian matrix, in which all the elements of the matrix are real. Also, if A is an (anti-)Hermitian matrix then so too is its inverse A^{-1}, since

$$(\mathsf{A}^{-1})^\dagger = (\mathsf{A}^\dagger)^{-1} = \pm\mathsf{A}^{-1}.$$

Any $N \times N$ matrix A can be written as the sum of an Hermitian matrix and an anti-Hermitian matrix, since

$$\mathsf{A} = \tfrac{1}{2}(\mathsf{A} + \mathsf{A}^\dagger) + \tfrac{1}{2}(\mathsf{A} - \mathsf{A}^\dagger) = \mathsf{B} + \mathsf{C},$$

where clearly $\mathsf{B} = \mathsf{B}^\dagger$ and $\mathsf{C} = -\mathsf{C}^\dagger$. The matrix B is called the Hermitian part of A, and C is called the anti-Hermitian part.

8.12.6 Unitary matrices

A *unitary* matrix A is defined as one for which

$$\mathsf{A}^\dagger = \mathsf{A}^{-1}. \tag{8.66}$$

Clearly, if A is real then $\mathsf{A}^\dagger = \mathsf{A}^{\mathrm{T}}$, showing that a real orthogonal matrix is a special case of a unitary matrix, one in which all the elements are real. We note that the inverse A^{-1} of a unitary is also unitary, since

$$(\mathsf{A}^{-1})^\dagger = (\mathsf{A}^\dagger)^{-1} = (\mathsf{A}^{-1})^{-1}.$$

Moreover, since for a unitary matrix $\mathsf{A}^\dagger\mathsf{A} = \mathsf{I}$, we have

$$|\mathsf{A}^\dagger\mathsf{A}| = |\mathsf{A}^\dagger||\mathsf{A}| = |\mathsf{A}|^*|\mathsf{A}| = |\mathsf{I}| = 1.$$

Thus the determinant of a unitary matrix has unit modulus.

A unitary matrix represents, in a particular basis, a linear operator that leaves the norms (lengths) of complex vectors unchanged. If $\mathbf{y} = \mathcal{A}\mathbf{x}$ is represented in some coordinate system by the matrix equation $\mathsf{y} = \mathsf{A}\mathsf{x}$ then $\langle \mathbf{y}|\mathbf{y}\rangle$ is given in this coordinate system by

$$\mathsf{y}^\dagger\mathsf{y} = \mathsf{x}^\dagger\mathsf{A}^\dagger\mathsf{A}\mathsf{x} = \mathsf{x}^\dagger\mathsf{x}.$$

Hence $\langle \mathbf{y}|\mathbf{y}\rangle = \langle \mathbf{x}|\mathbf{x}\rangle$, showing that the action of the linear operator represented by a unitary matrix does not change the norm of a complex vector. The action of a unitary matrix on a complex column matrix thus parallels that of an orthogonal matrix acting on a real column matrix.

8.12.7 Normal matrices

A final important set of special matrices consists of the *normal* matrices, for which

$$\mathsf{A}\mathsf{A}^\dagger = \mathsf{A}^\dagger\mathsf{A},$$

i.e. a normal matrix is one that commutes with its Hermitian conjugate.

We can easily show that Hermitian matrices and unitary matrices (or symmetric matrices and orthogonal matrices in the real case) are examples of normal matrices. For an Hermitian matrix, $\mathsf{A} = \mathsf{A}^\dagger$ and so

$$\mathsf{A}\mathsf{A}^\dagger = \mathsf{A}\mathsf{A} = \mathsf{A}^\dagger\mathsf{A}.$$

Similarly, for a unitary matrix, $\mathsf{A}^{-1} = \mathsf{A}^\dagger$ and so

$$\mathsf{A}\mathsf{A}^\dagger = \mathsf{A}\mathsf{A}^{-1} = \mathsf{A}^{-1}\mathsf{A} = \mathsf{A}^\dagger\mathsf{A}.$$

Finally, we note that, if A is normal then so too is its inverse A^{-1}, since

$$\mathsf{A}^{-1}(\mathsf{A}^{-1})^\dagger = \mathsf{A}^{-1}(\mathsf{A}^\dagger)^{-1} = (\mathsf{A}^\dagger\mathsf{A})^{-1} = (\mathsf{A}\mathsf{A}^\dagger)^{-1} = (\mathsf{A}^\dagger)^{-1}\mathsf{A}^{-1} = (\mathsf{A}^{-1})^\dagger\mathsf{A}^{-1}.$$

This broad class of matrices is important in the discussion of eigenvectors and eigenvalues in the next section.

8.13 Eigenvectors and eigenvalues

Suppose that a linear operator $\mathcal{A}$ transforms vectors $\mathbf{x}$ in an N-dimensional vector space into other vectors $\mathcal{A}\mathbf{x}$ in the same space. The possibility then arises that there exist vectors $\mathbf{x}$ each of which is transformed by $\mathcal{A}$ into a multiple of itself. Such vectors would have to satisfy

$$\mathcal{A}\mathbf{x} = \lambda\mathbf{x}. \tag{8.67}$$

Any non-zero vector $\mathbf{x}$ that satisfies (8.67) for some value of λ is called an *eigenvector* of the linear operator $\mathcal{A}$, and λ is called the corresponding *eigenvalue*. As will be discussed below, in general the operator $\mathcal{A}$ has N independent eigenvectors $\mathbf{x}^i$, with eigenvalues λ_i. The λ_i are not necessarily all distinct.

If we choose a particular basis in the vector space, we can write (8.67) in terms of the components of $\mathcal{A}$ and $\mathbf{x}$ with respect to this basis as the matrix equation

$$\mathsf{A}\mathsf{x} = \lambda\mathsf{x}, \tag{8.68}$$

where A is an $N \times N$ matrix. The column matrices x that satisfy (8.68) obviously represent the eigenvectors **x** of $\mathcal{A}$ in our chosen coordinate system. Conventionally, these column matrices are also referred to as the *eigenvectors of the matrix* A.[§] Clearly, if x is an eigenvector of A (with some eigenvalue λ) then any scalar multiple $\mu\mathsf{x}$ is also an eigenvector with the same eigenvalue. We therefore often use *normalised* eigenvectors, for which

$$\mathsf{x}^\dagger\mathsf{x} = 1$$

(note that $\mathsf{x}^\dagger\mathsf{x}$ corresponds to the inner product $\langle\mathbf{x}|\mathbf{x}\rangle$ in our basis). Any eigenvector x can be normalised by dividing all its components by the scalar $(\mathsf{x}^\dagger\mathsf{x})^{1/2}$.

As will be seen, the problem of finding the eigenvalues and corresponding eigenvectors of a square matrix A plays an important role in many physical investigations. Throughout this chapter we denote the ith eigenvector of a square matrix A by x^i and the corresponding eigenvalue by λ_i. This superscript notation for eigenvectors is used to avoid any confusion with components.

►*A non-singular matrix* A *has eigenvalues* λ_i *and eigenvectors* x^i. *Find the eigenvalues and eigenvectors of the inverse matrix* A^{-1}.

The eigenvalues and eigenvectors of A satisfy

$$\mathsf{A}\mathsf{x}^i = \lambda_i\mathsf{x}^i.$$

Left-multiplying both sides of this equation by A^{-1}, we find

$$\mathsf{A}^{-1}\mathsf{A}\mathsf{x}^i = \lambda_i\mathsf{A}^{-1}\mathsf{x}^i.$$

Since $\mathsf{A}^{-1}\mathsf{A} = \mathsf{I}$, on rearranging we obtain

$$\mathsf{A}^{-1}\mathsf{x}^i = \frac{1}{\lambda_i}\mathsf{x}^i.$$

Thus, we see that A^{-1} has the *same* eigenvectors $\mathbf{x}^i$ as does A, but the corresponding eigenvalues are $1/\lambda_i$. ◄

In the remainder of this section we will discuss some useful results concerning the eigenvectors and eigenvalues of certain special (though commonly occurring) square matrices. The results will be established for matrices whose elements may be complex; the corresponding properties for real matrices may be obtained as special cases.

8.13.1 Eigenvectors and eigenvalues of a normal matrix

In subsection 8.12.7 we defined a normal matrix A as one that commutes with its Hermitian conjugate, so that

$$\mathsf{A}^\dagger\mathsf{A} = \mathsf{A}\mathsf{A}^\dagger.$$

[§] In this context, when referring to linear combinations of eigenvectors x we will normally use the term 'vector'.

We also showed that both Hermitian and unitary matrices (or symmetric and orthogonal matrices in the real case) are examples of normal matrices. We now discuss the properties of the eigenvectors and eigenvalues of a normal matrix.

If x is an eigenvector of a normal matrix A with corresponding eigenvalue λ then $\mathsf{Ax} = \lambda\mathsf{x}$, or equivalently,

$$(\mathsf{A} - \lambda\mathsf{I})\mathsf{x} = \mathsf{0}. \tag{8.69}$$

Denoting $\mathsf{B} = \mathsf{A} - \lambda\mathsf{I}$, (8.69) becomes $\mathsf{Bx} = \mathsf{0}$ and, taking the Hermitian conjugate, we also have

$$(\mathsf{Bx})^\dagger = \mathsf{x}^\dagger\mathsf{B}^\dagger = \mathsf{0}. \tag{8.70}$$

From (8.69) and (8.70) we then have

$$\mathsf{x}^\dagger\mathsf{B}^\dagger\mathsf{Bx} = 0. \tag{8.71}$$

However, the product $\mathsf{B}^\dagger\mathsf{B}$ is given by

$$\mathsf{B}^\dagger\mathsf{B} = (\mathsf{A} - \lambda\mathsf{I})^\dagger(\mathsf{A} - \lambda\mathsf{I}) = (\mathsf{A}^\dagger - \lambda^*\mathsf{I})(\mathsf{A} - \lambda\mathsf{I}) = \mathsf{A}^\dagger\mathsf{A} - \lambda^*\mathsf{A} - \lambda\mathsf{A}^\dagger + \lambda\lambda^*.$$

Now since A is normal, $\mathsf{AA}^\dagger = \mathsf{A}^\dagger\mathsf{A}$ and so

$$\mathsf{B}^\dagger\mathsf{B} = \mathsf{AA}^\dagger - \lambda^*\mathsf{A} - \lambda\mathsf{A}^\dagger + \lambda\lambda^* = (\mathsf{A} - \lambda\mathsf{I})(\mathsf{A} - \lambda\mathsf{I})^\dagger = \mathsf{BB}^\dagger,$$

and hence B is also normal. From (8.71) we then find

$$\mathsf{x}^\dagger\mathsf{B}^\dagger\mathsf{Bx} = \mathsf{x}^\dagger\mathsf{BB}^\dagger\mathsf{x} = (\mathsf{B}^\dagger\mathsf{x})^\dagger\mathsf{B}^\dagger\mathsf{x} = 0,$$

from which we obtain

$$\mathsf{B}^\dagger\mathsf{x} = (\mathsf{A}^\dagger - \lambda^*\mathsf{I})\mathsf{x} = \mathsf{0}.$$

Therefore, for a normal matrix A, *the eigenvalues of* $\mathsf{A}^\dagger$ *are the complex conjugates of the eigenvalues of* A.

Let us now consider two eigenvectors x^i and x^j of a normal matrix A corresponding to two *different* eigenvalues λ_i and λ_j. We then have

$$\mathsf{Ax}^i = \lambda_i\mathsf{x}^i, \tag{8.72}$$
$$\mathsf{Ax}^j = \lambda_j\mathsf{x}^j. \tag{8.73}$$

Multiplying (8.73) on the left by $(\mathsf{x}^i)^\dagger$ we obtain

$$(\mathsf{x}^i)^\dagger\mathsf{Ax}^j = \lambda_j(\mathsf{x}^i)^\dagger\mathsf{x}^j. \tag{8.74}$$

However, on the LHS of (8.74) we have

$$(\mathsf{x}^i)^\dagger\mathsf{A} = (\mathsf{A}^\dagger\mathsf{x}^i)^\dagger = (\lambda_i^*\mathsf{x}^i)^\dagger = \lambda_i(\mathsf{x}^i)^\dagger, \tag{8.75}$$

where we have used (8.40) and the property just proved for a normal matrix to write $\mathsf{A}^\dagger \mathsf{x}^i = \lambda_i^* \mathsf{x}^i$. From (8.74) and (8.75) we have

$$(\lambda_i - \lambda_j)(\mathsf{x}^i)^\dagger \mathsf{x}^j = 0. \tag{8.76}$$

Thus, *if* $\lambda_i \neq \lambda_j$ *the eigenvectors* x^i *and* x^j *must be orthogonal*, i.e. $(\mathsf{x}^i)^\dagger \mathsf{x}^j = 0$.

It follows immediately from (8.76) that if all N eigenvalues of a normal matrix A are distinct then all N eigenvectors of A are mutually orthogonal. If, however, two or more eigenvalues are the same then further consideration is required. An eigenvalue corresponding to two or more different eigenvectors (i.e. they are not simply multiples of one another) is said to be *degenerate*. Suppose that λ_1 is k-fold degenerate, i.e.

$$\mathsf{A}\mathsf{x}^i = \lambda_1 \mathsf{x}^i \quad \text{for } i = 1, 2, \ldots, k, \tag{8.77}$$

but that it is different from any of λ_{k+1}, λ_{k+2}, etc. Then any linear combination of these x^i is also an eigenvector with eigenvalue λ_1, since, for $\mathsf{z} = \sum_{i=1}^k c_i \mathsf{x}^i$,

$$\mathsf{A}\mathsf{z} \equiv \mathsf{A} \sum_{i=1}^k c_i \mathsf{x}^i = \sum_{i=1}^k c_i \mathsf{A}\mathsf{x}^i = \sum_{i=1}^k c_i \lambda_1 \mathsf{x}^i = \lambda_1 \mathsf{z}. \tag{8.78}$$

If the x^i defined in (8.77) are not already mutually orthogonal then we can construct new eigenvectors z^i that are orthogonal by the following procedure:

$$\begin{aligned}
\mathsf{z}^1 &= \mathsf{x}^1, \\
\mathsf{z}^2 &= \mathsf{x}^2 - \left[(\hat{\mathsf{z}}^1)^\dagger \mathsf{x}^2\right] \hat{\mathsf{z}}^1, \\
\mathsf{z}^3 &= \mathsf{x}^3 - \left[(\hat{\mathsf{z}}^2)^\dagger \mathsf{x}^3\right] \hat{\mathsf{z}}^2 - \left[(\hat{\mathsf{z}}^1)^\dagger \mathsf{x}^3\right] \hat{\mathsf{z}}^1, \\
&\ \vdots \\
\mathsf{z}^k &= \mathsf{x}^k - \left[(\hat{\mathsf{z}}^{k-1})^\dagger \mathsf{x}^k\right] \hat{\mathsf{z}}^{k-1} - \cdots - \left[(\hat{\mathsf{z}}^1)^\dagger \mathsf{x}^k\right] \hat{\mathsf{z}}^1.
\end{aligned}$$

In this procedure, known as *Gram–Schmidt orthogonalisation*, each new eigenvector z^i is normalised to give the unit vector $\hat{\mathsf{z}}^i$ before proceeding to the construction of the next one (the normalisation is carried out by dividing each element of the vector z^i by $[(\mathsf{z}^i)^\dagger \mathsf{z}^i]^{1/2}$). Note that each factor in brackets $(\hat{\mathsf{z}}^m)^\dagger \mathsf{x}^n$ is a scalar product and thus only a number. It follows that, as shown in (8.78), each vector z^i so constructed is an eigenvector of A with eigenvalue λ_1 and will remain so on normalisation. It is straightforward to check that, provided the previous new eigenvectors have been normalised as prescribed, each z^i is orthogonal to all its predecessors. (In practice, however, the method is laborious and the example in subsection 8.14.1 gives a less rigorous but considerably quicker way.)

Therefore, even if A has some degenerate eigenvalues we can *by construction* obtain a set of N mutually orthogonal eigenvectors. Moreover, it may be shown (although the proof is beyond the scope of this book) that these eigenvectors are *complete* in that they form a basis for the N-dimensional vector space. As a result any arbitrary vector y can be expressed as a linear combination of the eigenvectors x^i:

$$\mathsf{y} = \sum_{i=1}^{N} a_i \mathsf{x}^i, \tag{8.79}$$

where $a_i = (\mathsf{x}^i)^\dagger \mathsf{y}$. Thus, the eigenvectors form an orthogonal basis for the vector space. By normalising the eigenvectors so that $(\mathsf{x}^i)^\dagger \mathsf{x}^i = 1$ this basis is made orthonormal.

►*Show that a normal matrix* A *can be written in terms of its eigenvalues* λ_i *and orthonormal eigenvectors* x^i *as*

$$\mathsf{A} = \sum_{i=1}^{N} \lambda_i \mathsf{x}^i (\mathsf{x}^i)^\dagger. \tag{8.80}$$

The key to proving the validity of (8.80) is to show that both sides of the expression give the same result when acting on an arbitary vector y. Since A is normal, we may expand y in terms of the eigenvectors x^i, as shown in (8.79). Thus, we have

$$\mathsf{A}\mathsf{y} = \mathsf{A} \sum_{i=1}^{N} a_i \mathsf{x}^i = \sum_{i=1}^{N} a_i \lambda_i \mathsf{x}^i.$$

Alternatively, the action of the RHS of (8.80) on y is given by

$$\sum_{i=1}^{N} \lambda_i \mathsf{x}^i (\mathsf{x}^i)^\dagger \mathsf{y} = \sum_{i=1}^{N} a_i \lambda_i \mathsf{x}^i,$$

since $a_i = (\mathsf{x}^i)^\dagger \mathsf{y}$. We see that the two expressions for the action of each side of (8.80) on y are identical, which implies that this relationship is indeed correct. ◄

8.13.2 Eigenvectors and eigenvalues of Hermitian and anti-Hermitian matrices

For a normal matrix we showed that if $\mathsf{A}\mathsf{x} = \lambda \mathsf{x}$ then $\mathsf{A}^\dagger \mathsf{x} = \lambda^* \mathsf{x}$. However, if A is also Hermitian, $\mathsf{A} = \mathsf{A}^\dagger$, it follows necessarily that $\lambda = \lambda^*$. Thus, the eigenvalues of an Hermitian matrix are real, a result which may be proved directly.

►*Prove that the eigenvalues of an Hermitian matrix are real.*

For any particular eigenvector x^i, we take the Hermitian conjugate of $\mathsf{A}\mathsf{x}^i = \lambda_i \mathsf{x}^i$ to give

$$(\mathsf{x}^i)^\dagger \mathsf{A}^\dagger = \lambda_i^* (\mathsf{x}^i)^\dagger. \tag{8.81}$$

Using $\mathsf{A}^\dagger = \mathsf{A}$, since A is Hermitian, and multiplying on the right by x^i, we obtain

$$(\mathsf{x}^i)^\dagger \mathsf{A} \mathsf{x}^i = \lambda_i^* (\mathsf{x}^i)^\dagger \mathsf{x}^i. \tag{8.82}$$

But multiplying $\mathsf{A}\mathsf{x}^i = \lambda_i \mathsf{x}^i$ through on the left by $(\mathsf{x}^i)^\dagger$ gives

$$(\mathsf{x}^i)^\dagger \mathsf{A} \mathsf{x}^i = \lambda_i (\mathsf{x}^i)^\dagger \mathsf{x}^i.$$

Subtracting this from (8.82) yields

$$0 = (\lambda_i^* - \lambda_i)(\mathsf{x}^i)^\dagger \mathsf{x}^i.$$

But $(\mathsf{x}^i)^\dagger \mathsf{x}^i$ is the modulus squared of the non-zero vector x^i and is thus non-zero. Hence λ_i^* must equal λ_i and thus be real. The same argument can be used to show that the eigenvalues of a real symmetric matrix are themselves real. ◄

The importance of the above result will be apparent to any student of quantum mechanics. In quantum mechanics the eigenvalues of operators correspond to measured values of observable quantities, e.g. energy, angular momentum, parity and so on, and these clearly must be real. If we use Hermitian operators to formulate the theories of quantum mechanics, the above property guarantees physically meaningful results.

Since an Hermitian matrix is also a normal matrix, its eigenvectors are orthogonal (or can be made so using the Gram–Schmidt orthogonalisation procedure). Alternatively we can prove the orthogonality of the eigenvectors directly.

►*Prove that the eigenvectors corresponding to different eigenvalues of an Hermitian matrix are orthogonal.*

Consider two unequal eigenvalues λ_i and λ_j and their corresponding eigenvectors satisfying

$$\mathsf{A}\mathsf{x}^i = \lambda_i \mathsf{x}^i, \tag{8.83}$$

$$\mathsf{A}\mathsf{x}^j = \lambda_j \mathsf{x}^j. \tag{8.84}$$

Taking the Hermitian conjugate of (8.83) we find $(\mathsf{x}^i)^\dagger \mathsf{A}^\dagger = \lambda_i^*(\mathsf{x}^i)^\dagger$. Multiplying this on the right by x^j we obtain

$$(\mathsf{x}^i)^\dagger \mathsf{A}^\dagger \mathsf{x}^j = \lambda_i^*(\mathsf{x}^i)^\dagger \mathsf{x}^j,$$

and similarly multiplying (8.84) through on the left by $(\mathsf{x}^i)^\dagger$ we find

$$(\mathsf{x}^i)^\dagger \mathsf{A} \mathsf{x}^j = \lambda_j(\mathsf{x}^i)^\dagger \mathsf{x}^j.$$

Then, since $\mathsf{A}^\dagger = \mathsf{A}$, the two left-hand sides are equal and, because the λ_i are real, on subtraction we obtain

$$0 = (\lambda_i - \lambda_j)(\mathsf{x}^i)^\dagger \mathsf{x}^j.$$

Finally we note that $\lambda_i \neq \lambda_j$ and so $(\mathsf{x}^i)^\dagger \mathsf{x}^j = 0$, i.e. the eigenvectors x^i and x^j are orthogonal. ◄

In the case where some of the eigenvalues are equal, further justification of the orthogonality of the eigenvectors is needed. The Gram–Schmidt orthogonalisation procedure discussed above provides a proof of, and a means of achieving, orthogonality. The general method has already been described and we will not repeat it here.

We may also consider the properties of the eigenvalues and eigenvectors of an anti-Hermitian matrix, for which $\mathsf{A}^\dagger = -\mathsf{A}$ and thus

$$\mathsf{A}\mathsf{A}^\dagger = \mathsf{A}(-\mathsf{A}) = (-\mathsf{A})\mathsf{A} = \mathsf{A}^\dagger \mathsf{A}.$$

Therefore matrices that are anti-Hermitian are also normal and so have mutually orthogonal eigenvectors. The properties of the eigenvalues are also simply deduced, since if $\mathsf{A}\mathsf{x} = \lambda \mathsf{x}$ then

$$\lambda^* \mathsf{x} = \mathsf{A}^\dagger \mathsf{x} = -\mathsf{A}\mathsf{x} = -\lambda \mathsf{x}.$$

Hence $\lambda^* = -\lambda$ and so λ must be *pure imaginary* (or *zero*). In a similar manner to that used for Hermitian matrices, these properties may be proved directly.

8.13.3 Eigenvectors and eigenvalues of a unitary matrix

A unitary matrix satisfies $\mathsf{A}^\dagger = \mathsf{A}^{-1}$ and is also a normal matrix, with mutually orthogonal eigenvectors. To investigate the eigenvalues of a unitary matrix, we note that if $\mathsf{A}\mathsf{x} = \lambda\mathsf{x}$ then

$$\mathsf{x}^\dagger\mathsf{x} = \mathsf{x}^\dagger\mathsf{A}^\dagger\mathsf{A}\mathsf{x} = \lambda^*\lambda\mathsf{x}^\dagger\mathsf{x},$$

and we deduce that $\lambda\lambda^* = |\lambda|^2 = 1$. Thus, the eigenvalues of a unitary matrix have unit modulus.

8.13.4 Eigenvectors and eigenvalues of a general square matrix

When an $N \times N$ matrix is not normal there are no general properties of its eigenvalues and eigenvectors; in general it is not possible to find any orthogonal set of N eigenvectors or even to find *pairs* of orthogonal eigenvectors (except by chance in some cases). While the N non-orthogonal eigenvectors are usually linearly independent and hence form a basis for the N-dimensional vector space, this is not necessarily so. It may be shown (although we will not prove it) that any $N \times N$ matrix with *distinct* eigenvalues has N linearly independent eigenvectors, which therefore form a basis for the N-dimensional vector space. If a general square matrix has degenerate eigenvalues, however, then it may or may not have N linearly independent eigenvectors. A matrix whose eigenvectors are not linearly independent is said to be *defective*.

8.13.5 Simultaneous eigenvectors

We may now ask under what conditions two different normal matrices can have a common set of eigenvectors. The result – that they do so if, and only if, they commute – has profound significance for the foundations of quantum mechanics.

To prove this important result let A and B be two $N \times N$ normal matrices and x^i be the ith eigenvector of A corresponding to eigenvalue λ_i, i.e.

$$\mathsf{A}\mathsf{x}^i = \lambda_i\mathsf{x}^i \quad \text{for} \quad i = 1, 2, \ldots, N.$$

For the present we assume that the eigenvalues are all different.

(i) First suppose that A and B commute. Now consider

$$\mathsf{A}\mathsf{B}\mathsf{x}^i = \mathsf{B}\mathsf{A}\mathsf{x}^i = \mathsf{B}\lambda_i\mathsf{x}^i = \lambda_i\mathsf{B}\mathsf{x}^i,$$

where we have used the commutativity for the first equality and the eigenvector property for the second. It follows that $\mathsf{A}(\mathsf{B}\mathsf{x}^i) = \lambda_i(\mathsf{B}\mathsf{x}^i)$ and thus that $\mathsf{B}\mathsf{x}^i$ is an eigenvector of A corresponding to eigenvalue λ_i. But the eigenvector solutions of $(\mathsf{A}-\lambda_i\mathsf{I})\mathsf{x}^i=\mathsf{0}$ are unique to within a scale factor, and we therefore conclude that

$$\mathsf{B}\mathsf{x}^i=\mu_i\mathsf{x}^i$$

for some scale factor μ_i. However, this is just an eigenvector equation for B and shows that x^i is an eigenvector of B, in addition to being an eigenvector of A. By reversing the roles of A and B, it also follows that every eigenvector of B is an eigenvector of A. Thus the two sets of eigenvectors are identical.

(ii) Now suppose that A and B have all their eigenvectors in common, a typical one x^i satisfying both

$$\mathsf{A}\mathsf{x}^i=\lambda_i\mathsf{x}^i \quad\text{and}\quad \mathsf{B}\mathsf{x}^i=\mu_i\mathsf{x}^i.$$

As the eigenvectors span the N-dimensional vector space, any arbitrary vector x in the space can be written as a linear combination of the eigenvectors,

$$\mathsf{x}=\sum_{i=1}^{N}c_i\mathsf{x}^i.$$

Now consider both

$$\mathsf{A}\mathsf{B}\mathsf{x}=\mathsf{A}\mathsf{B}\sum_{i=1}^{N}c_i\mathsf{x}^i=\mathsf{A}\sum_{i=1}^{N}c_i\mu_i\mathsf{x}^i=\sum_{i=1}^{N}c_i\lambda_i\mu_i\mathsf{x}^i,$$

and

$$\mathsf{B}\mathsf{A}\mathsf{x}=\mathsf{B}\mathsf{A}\sum_{i=1}^{N}c_i\mathsf{x}^i=\mathsf{B}\sum_{i=1}^{N}c_i\lambda_i\mathsf{x}^i=\sum_{i=1}^{N}c_i\mu_i\lambda_i\mathsf{x}^i.$$

It follows that $\mathsf{A}\mathsf{B}\mathsf{x}$ and $\mathsf{B}\mathsf{A}\mathsf{x}$ are the same for any arbitrary x and hence that

$$(\mathsf{A}\mathsf{B}-\mathsf{B}\mathsf{A})\mathsf{x}=\mathsf{0}$$

for all x. That is, A and B *commute*.

This completes the proof that a necessary and sufficient condition for two normal matrices to have a set of eigenvectors in common is that they commute. It should be noted that if an eigenvalue of A, say, is degenerate then not all of its possible sets of eigenvectors will also constitute a set of eigenvectors of B. However, provided that by taking linear combinations one set of joint eigenvectors can be found, the proof is still valid and the result still holds.

When extended to the case of Hermitian operators and continuous eigenfunctions (sections 17.2 and 17.3) the connection between commuting matrices and a set of common eigenvectors plays a fundamental role in the postulatory basis of quantum mechanics. It draws the distinction between commuting and non-commuting observables and sets limits on how much information about a system can be known, even in principle, at any one time.

8.14 Determination of eigenvalues and eigenvectors

The next step is to show how the eigenvalues and eigenvectors of a given $N \times N$ matrix A are found. To do this we refer to (8.68) and as in (8.69) rewrite it as

$$\mathsf{Ax} - \lambda \mathsf{Ix} = (\mathsf{A} - \lambda \mathsf{I})\mathsf{x} = \mathsf{0}. \tag{8.85}$$

The slight rearrangement used here is to write x as Ix, where I is the unit matrix of order N. The point of doing this is immediate since (8.85) now has the form of a homogeneous set of simultaneous equations, the theory of which will be developed in section 8.18. What will be proved there is that the equation $\mathsf{Bx} = \mathsf{0}$ only has a non-trivial solution x if $|\mathsf{B}| = 0$. Correspondingly, therefore, we must have in the present case that

$$|\mathsf{A} - \lambda \mathsf{I}| = 0, \tag{8.86}$$

if there are to be non-zero solutions x to (8.85).

Equation (8.86) is known as the *characteristic equation* for A and its LHS as the *characteristic* or *secular determinant* of A. The equation is a polynomial of degree N in the quantity λ. The N roots of this equation λ_i, $i = 1, 2, \ldots, N$, give the eigenvalues of A. Corresponding to each λ_i there will be a column vector x^i, which is the ith eigenvector of A and can be found by using (8.68).

It will be observed that when (8.86) is written out as a polynomial equation in λ, the coefficient of $-\lambda^{N-1}$ in the equation will be simply $A_{11} + A_{22} + \cdots + A_{NN}$ relative to the coefficient of λ^N. As discussed in section 8.8, the quantity $\sum_{i=1}^{N} A_{ii}$ is the *trace* of A and, from the ordinary theory of polynomial equations, will be equal to the sum of the roots of (8.86):

$$\sum_{i=1}^{N} \lambda_i = \mathrm{Tr}\,\mathsf{A}. \tag{8.87}$$

This can be used as one check that a computation of the eigenvalues λ_i has been done correctly. Unless equation (8.87) is satisfied by a computed set of eigenvalues, they have not been calculated correctly. However, that equation (8.87) is satisfied is a necessary, but not sufficient, condition for a correct computation. An alternative proof of (8.87) is given in section 8.16.

►*Find the eigenvalues and normalised eigenvectors of the real symmetric matrix*

$$\mathsf{A} = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 1 & -3 \\ 3 & -3 & -3 \end{pmatrix}.$$

Using (8.86),

$$\begin{vmatrix} 1-\lambda & 1 & 3 \\ 1 & 1-\lambda & -3 \\ 3 & -3 & -3-\lambda \end{vmatrix} = 0.$$

Expanding out this determinant gives

$$(1-\lambda)\,[(1-\lambda)(-3-\lambda)-(-3)(-3)]+1\,[(-3)(3)-1(-3-\lambda)]$$
$$+\,3\,[1(-3)-(1-\lambda)(3)]=0,$$

which simplifies to give

$$\begin{aligned}(1-\lambda)(\lambda^2+2\lambda-12)+(\lambda-6)+3(3\lambda-6)&=0,\\ \Rightarrow\qquad (\lambda-2)(\lambda-3)(\lambda+6)&=0.\end{aligned}$$

Hence the roots of the characteristic equation, which are the eigenvalues of A, are $\lambda_1=2$, $\lambda_2=3$, $\lambda_3=-6$. We note that, as expected,

$$\lambda_1+\lambda_2+\lambda_3=-1=1+1-3=A_{11}+A_{22}+A_{33}=\mathrm{Tr}\,\mathsf{A}.$$

For the first root, $\lambda_1=2$, a suitable eigenvector x^1, with elements x_1, x_2, x_3, must satisfy $\mathsf{A}\mathsf{x}^1=2\mathsf{x}^1$ or, equivalently,

$$\begin{aligned}x_1+x_2+3x_3&=2x_1,\\ x_1+x_2-3x_3&=2x_2,\\ 3x_1-3x_2-3x_3&=2x_3.\end{aligned}\qquad(8.88)$$

These three equations are consistent (to ensure this was the purpose in finding the particular values of λ) and yield $x_3=0$, $x_1=x_2=k$, where k is any non-zero number. A suitable eigenvector would thus be

$$\mathsf{x}^1=(k\quad k\quad 0)^{\mathrm{T}}.$$

If we apply the normalisation condition, we require $k^2+k^2+0^2=1$ or $k=1/\sqrt{2}$. Hence

$$\mathsf{x}^1=\left(\frac{1}{\sqrt{2}}\quad\frac{1}{\sqrt{2}}\quad 0\right)^{\mathrm{T}}=\frac{1}{\sqrt{2}}(1\quad 1\quad 0)^{\mathrm{T}}.$$

Repeating the last paragraph, but with the factor 2 on the RHS of (8.88) replaced successively by $\lambda_2=3$ and $\lambda_3=-6$, gives two further normalised eigenvectors

$$\mathsf{x}^2=\frac{1}{\sqrt{3}}(1\quad -1\quad 1)^{\mathrm{T}},\qquad \mathsf{x}^3=\frac{1}{\sqrt{6}}(1\quad -1\quad -2)^{\mathrm{T}}.\ ◀$$

In the above example, the three values of λ are all different and A is a real symmetric matrix. Thus we expect, and it is easily checked, that the three eigenvectors are mutually orthogonal, i.e.

$$\left(\mathsf{x}^1\right)^{\mathrm{T}}\mathsf{x}^2=\left(\mathsf{x}^1\right)^{\mathrm{T}}\mathsf{x}^3=\left(\mathsf{x}^2\right)^{\mathrm{T}}\mathsf{x}^3=0.$$

It will be apparent also that, as expected, the normalisation of the eigenvectors has no effect on their orthogonality.

8.14.1 Degenerate eigenvalues

We return now to the case of degenerate eigenvalues, i.e. those that have two or more associated eigenvectors. We have shown already that it is always possible to construct an orthogonal set of eigenvectors for a normal matrix, see subsection 8.13.1, and the following example illustrates one method for constructing such a set.

►*Construct an orthonormal set of eigenvectors for the matrix*

$$\mathsf{A} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & -2 & 0 \\ 3 & 0 & 1 \end{pmatrix}.$$

We first determine the eigenvalues using $|\mathsf{A} - \lambda \mathsf{I}| = 0$:

$$0 = \begin{vmatrix} 1-\lambda & 0 & 3 \\ 0 & -2-\lambda & 0 \\ 3 & 0 & 1-\lambda \end{vmatrix} = -(1-\lambda)^2(2+\lambda) + 3(3)(2+\lambda)$$
$$= (4-\lambda)(\lambda+2)^2.$$

Thus $\lambda_1 = 4$, $\lambda_2 = -2 = \lambda_3$. The eigenvector $\mathsf{x}^1 = (x_1 \quad x_2 \quad x_3)^{\mathrm{T}}$ is found from

$$\begin{pmatrix} 1 & 0 & 3 \\ 0 & -2 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 4 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \Rightarrow \quad \mathsf{x}^1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

A general column vector that is orthogonal to x^1 is

$$\mathsf{x} = (a \quad b \quad -a)^{\mathrm{T}}, \tag{8.89}$$

and it is easily shown that

$$\mathsf{Ax} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & -2 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ -a \end{pmatrix} = -2 \begin{pmatrix} a \\ b \\ -a \end{pmatrix} = -2\mathsf{x}.$$

Thus x is a eigenvector of A with associated eigenvalue -2. It is clear, however, that there is an infinite set of eigenvectors x all possessing the required property; the geometrical analogue is that there are an infinite number of corresponding vectors $\mathbf{x}$ lying in the plane that has $\mathbf{x}^1$ as its normal. We do require that the two remaining eigenvectors are orthogonal to one another, but this still leaves an infinite number of possibilities. For x^2, therefore, let us choose a simple form of (8.89), suitably normalised, say,

$$\mathsf{x}^2 = (0 \quad 1 \quad 0)^{\mathrm{T}}.$$

The third eigenvector is then specified (to within an arbitrary multiplicative constant) by the requirement that it must be orthogonal to x^1 and x^2; thus x^3 may be found by evaluating the vector product of x^1 and x^2 and normalising the result. This gives

$$\mathsf{x}^3 = \frac{1}{\sqrt{2}}(-1 \quad 0 \quad 1)^{\mathrm{T}},$$

to complete the construction of an orthonormal set of eigenvectors. ◄

8.15 Change of basis and similarity transformations

Throughout this chapter we have considered the vector $\mathbf{x}$ as a geometrical quantity that is independent of any basis (or coordinate system). If we introduce a basis $\mathbf{e}_i$, $i = 1, 2, \ldots, N$, into our N-dimensional vector space then we may write

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_N\mathbf{e}_N,$$

and represent **x** in this basis by the column matrix

$$\mathsf{x} = (x_1 \quad x_2 \; \cdots \; x_n)^{\mathrm{T}},$$

having components x_i. We now consider how these components change as a result of a prescribed change of basis. Let us introduce a new basis $\mathbf{e}'_i$, $i = 1, 2, \ldots, N$, which is related to the old basis by

$$\mathbf{e}'_j = \sum_{i=1}^{N} S_{ij}\mathbf{e}_i, \tag{8.90}$$

the coefficient S_{ij} being the ith component of $\mathbf{e}'_j$ with respect to the old (unprimed) basis. For an arbitrary vector **x** it follows that

$$\mathbf{x} = \sum_{i=1}^{N} x_i\mathbf{e}_i = \sum_{j=1}^{N} x'_j\mathbf{e}'_j = \sum_{j=1}^{N} x'_j \sum_{i=1}^{N} S_{ij}\mathbf{e}_i.$$

From this we derive the relationship between the components of **x** in the two coordinate systems as

$$x_i = \sum_{j=1}^{N} S_{ij}x'_j,$$

which we can write in matrix form as

$$\mathsf{x} = \mathsf{S}\mathsf{x}' \tag{8.91}$$

where S is the *transformation matrix* associated with the change of basis.

Furthermore, since the vectors $\mathbf{e}'_j$ are linearly independent, the matrix S is non-singular and so possesses an inverse S^{-1}. Multiplying (8.91) on the left by S^{-1} we find

$$\mathsf{x}' = \mathsf{S}^{-1}\mathsf{x}, \tag{8.92}$$

which relates the components of **x** in the new basis to those in the old basis. Comparing (8.92) and (8.90) we note that the components of **x** transform inversely to the way in which the basis vectors $\mathbf{e}_i$ themselves transform. This has to be so, as the vector **x** itself must remain unchanged.

We may also find the transformation law for the components of a linear operator under the same change of basis. Now, the operator equation $\mathbf{y} = \mathcal{A}\,\mathbf{x}$ (which is basis independent) can be written as a matrix equation in each of the two bases as

$$\mathsf{y} = \mathsf{A}\mathsf{x}, \qquad \mathsf{y}' = \mathsf{A}'\mathsf{x}'. \tag{8.93}$$

But, using (8.91), we may rewrite the first equation as

$$\mathsf{S}\mathsf{y}' = \mathsf{A}\mathsf{S}\mathsf{x}' \quad \Rightarrow \quad \mathsf{y}' = \mathsf{S}^{-1}\mathsf{A}\mathsf{S}\mathsf{x}'.$$

Comparing this with the second equation in (8.93) we find that the components of the linear operator $\mathcal{A}$ transform as

$$\mathsf{A}' = \mathsf{S}^{-1}\mathsf{AS}. \tag{8.94}$$

Equation (8.94) is an example of a *similarity transformation* – a transformation that can be particularly useful in converting matrices into convenient forms for computation.

Given a square matrix A, we may interpret it as representing a linear operator $\mathcal{A}$ in a given basis $\mathbf{e}_i$. From (8.94), however, we may also consider the matrix $\mathsf{A}' = \mathsf{S}^{-1}\mathsf{AS}$, for any non-singular matrix S, as representing the same linear operator $\mathcal{A}$ but in a new basis $\mathbf{e}'_j$, related to the old basis by

$$\mathbf{e}'_j = \sum_i S_{ij}\mathbf{e}_i.$$

Therefore we would expect that any property of the matrix A that represents some (basis-independent) property of the linear operator $\mathcal{A}$ will also be shared by the matrix A'. We list these properties below.

(i) If $\mathsf{A} = \mathsf{I}$ then $\mathsf{A}' = \mathsf{I}$, since, from (8.94),

$$\mathsf{A}' = \mathsf{S}^{-1}\mathsf{IS} = \mathsf{S}^{-1}\mathsf{S} = \mathsf{I}. \tag{8.95}$$

(ii) The value of the determinant is unchanged:

$$|\mathsf{A}'| = |\mathsf{S}^{-1}\mathsf{AS}| = |\mathsf{S}^{-1}||\mathsf{A}||\mathsf{S}| = |\mathsf{A}||\mathsf{S}^{-1}||\mathsf{S}| = |\mathsf{A}||\mathsf{S}^{-1}\mathsf{S}| = |\mathsf{A}|. \tag{8.96}$$

(iii) The characteristic determinant and hence the eigenvalues of A' are the same as those of A: from (8.86),

$$\begin{aligned}|\mathsf{A}' - \lambda\mathsf{I}| &= |\mathsf{S}^{-1}\mathsf{AS} - \lambda\mathsf{I}| = |\mathsf{S}^{-1}(\mathsf{A} - \lambda\mathsf{I})\mathsf{S}| \\ &= |\mathsf{S}^{-1}||\mathsf{S}||\mathsf{A} - \lambda\mathsf{I}| = |\mathsf{A} - \lambda\mathsf{I}|. \end{aligned} \tag{8.97}$$

(iv) The value of the trace is unchanged: from (8.87),

$$\begin{aligned}\operatorname{Tr}\mathsf{A}' &= \sum_i A'_{ii} = \sum_i\sum_j\sum_k (\mathsf{S}^{-1})_{ij}A_{jk}S_{ki} \\ &= \sum_i\sum_j\sum_k S_{ki}(\mathsf{S}^{-1})_{ij}A_{jk} = \sum_j\sum_k \delta_{kj}A_{jk} = \sum_j A_{jj} \\ &= \operatorname{Tr}\mathsf{A}. \end{aligned} \tag{8.98}$$

An important class of similarity transformations is that for which S is a unitary matrix; in this case $\mathsf{A}' = \mathsf{S}^{-1}\mathsf{AS} = \mathsf{S}^{\dagger}\mathsf{AS}$. Unitary transformation matrices are particularly important, for the following reason. If the original basis $\mathbf{e}_i$ is orthonormal and the transformation matrix S is unitary then

$$\begin{aligned}\langle \mathbf{e}_i'|\mathbf{e}_j'\rangle &= \Big\langle \sum_k S_{ki}\mathbf{e}_k \Big| \sum_r S_{rj}\mathbf{e}_r \Big\rangle \\ &= \sum_k S_{ki}^* \sum_r S_{rj}\langle \mathbf{e}_k|\mathbf{e}_r\rangle \\ &= \sum_k S_{ki}^* \sum_r S_{rj}\delta_{kr} = \sum_k S_{ki}^* S_{kj} = (\mathsf{S}^\dagger \mathsf{S})_{ij} = \delta_{ij},\end{aligned}$$

showing that the new basis is also orthonormal.

Furthermore, in addition to the properties of general similarity transformations, for unitary transformations the following hold.

(i) If A is Hermitian (anti-Hermitian) then A' is Hermitian (anti-Hermitian), i.e. if $\mathsf{A}^\dagger = \pm\mathsf{A}$ then

$$(\mathsf{A}')^\dagger = (\mathsf{S}^\dagger\mathsf{A}\mathsf{S})^\dagger = \mathsf{S}^\dagger\mathsf{A}^\dagger\mathsf{S} = \pm\mathsf{S}^\dagger\mathsf{A}\mathsf{S} = \pm\mathsf{A}'. \tag{8.99}$$

(ii) If A is unitary (so that $\mathsf{A}^\dagger = \mathsf{A}^{-1}$) then A' is unitary, since

$$\begin{aligned}(\mathsf{A}')^\dagger\mathsf{A}' &= (\mathsf{S}^\dagger\mathsf{A}\mathsf{S})^\dagger(\mathsf{S}^\dagger\mathsf{A}\mathsf{S}) = \mathsf{S}^\dagger\mathsf{A}^\dagger\mathsf{S}\mathsf{S}^\dagger\mathsf{A}\mathsf{S} = \mathsf{S}^\dagger\mathsf{A}^\dagger\mathsf{A}\mathsf{S} \\ &= \mathsf{S}^\dagger\mathsf{I}\mathsf{S} = \mathsf{I}.\end{aligned} \tag{8.100}$$

8.16 Diagonalisation of matrices

Suppose that a linear operator $\mathcal{A}$ is represented in some basis $\mathbf{e}_i$, $i = 1, 2, \ldots, N$, by the matrix A. Consider a new basis $\mathbf{x}^j$ given by

$$\mathbf{x}^j = \sum_{i=1}^N S_{ij}\mathbf{e}_i,$$

where the $\mathbf{x}^j$ are chosen to be the eigenvectors of the linear operator $\mathcal{A}$, i.e.

$$\mathcal{A}\,\mathbf{x}^j = \lambda_j\mathbf{x}^j. \tag{8.101}$$

In the new basis, $\mathcal{A}$ is represented by the matrix $\mathsf{A}' = \mathsf{S}^{-1}\mathsf{A}\mathsf{S}$, which has a particularly simple form, as we shall see shortly. The element S_{ij} of S is the ith component, in the old (unprimed) basis, of the jth eigenvector x^j of A, i.e. the columns of S are the eigenvectors of the matrix A:

$$\mathsf{S} = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \mathsf{x}^1 & \mathsf{x}^2 & \cdots & \mathsf{x}^N \\ \downarrow & \downarrow & & \downarrow \end{pmatrix},$$

that is, $S_{ij} = (\mathsf{x}^j)_i$. Therefore A' is given by

$$
\begin{aligned}
(\mathsf{S}^{-1}\mathsf{A}\mathsf{S})_{ij} &= \sum_k \sum_l (\mathsf{S}^{-1})_{ik} A_{kl} S_{lj} \\
&= \sum_k \sum_l (\mathsf{S}^{-1})_{ik} A_{kl} (\mathsf{x}^j)_l \\
&= \sum_k (\mathsf{S}^{-1})_{ik} \lambda_j (\mathsf{x}^j)_k \\
&= \sum_k \lambda_j (\mathsf{S}^{-1})_{ik} S_{kj} = \lambda_j \delta_{ij}.
\end{aligned}
$$

So the matrix A' is diagonal with the eigenvalues of $\mathcal{A}$ as the diagonal elements, i.e.

$$
\mathsf{A}' = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_N \end{pmatrix}.
$$

Therefore, given a matrix A, if we construct the matrix S that has the eigenvectors of A as its columns then the matrix $\mathsf{A}' = \mathsf{S}^{-1}\mathsf{A}\mathsf{S}$ is diagonal and has the eigenvalues of A as its diagonal elements. Since we require S to be non-singular ($|\mathsf{S}| \neq 0$), the N eigenvectors of A must be linearly independent and form a basis for the N-dimensional vector space. It may be shown that *any matrix with distinct eigenvalues* can be diagonalised by this procedure. If, however, a general square matrix has degenerate eigenvalues then it may, or may not, have N linearly independent eigenvectors. If it does not then it *cannot* be diagonalised.

For normal matrices (which include Hermitian, anti-Hermitian and unitary matrices) the N eigenvectors are indeed linearly independent. Moreover, when normalised, these eigenvectors form an *orthonormal* set (or can be made to do so). Therefore the matrix S with these normalised eigenvectors as columns, i.e. whose elements are $S_{ij} = (\mathsf{x}^j)_i$, has the property

$$
(\mathsf{S}^\dagger\mathsf{S})_{ij} = \sum_k (\mathsf{S}^\dagger)_{ik}(\mathsf{S})_{kj} = \sum_k S_{ki}^* S_{kj} = \sum_k (\mathsf{x}^i)_k^* (\mathsf{x}^j)_k = (\mathsf{x}^i)^\dagger \mathsf{x}^j = \delta_{ij}.
$$

Hence S is unitary ($\mathsf{S}^{-1} = \mathsf{S}^\dagger$) and the original matrix A can be diagonalised by

$$
\mathsf{A}' = \mathsf{S}^{-1}\mathsf{A}\mathsf{S} = \mathsf{S}^\dagger\mathsf{A}\mathsf{S}.
$$

Therefore, any normal matrix A can be diagonalised by a similarity transformation using a *unitary* transformation matrix S.

►*Diagonalise the matrix*

$$\mathsf{A} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & -2 & 0 \\ 3 & 0 & 1 \end{pmatrix}.$$

The matrix A is symmetric and so may be diagonalised by a transformation of the form $\mathsf{A}' = \mathsf{S}^\dagger \mathsf{A} \mathsf{S}$, where S has the normalised eigenvectors of A as its columns. We have already found these eigenvectors in subsection 8.14.1, and so we can write straightaway

$$\mathsf{S} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & -1 \\ 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

We note that although the eigenvalues of A are degenerate, its three eigenvectors are linearly independent and so A can still be diagonalised. Thus, calculating $\mathsf{S}^\dagger \mathsf{A} \mathsf{S}$ we obtain

$$\begin{aligned} \mathsf{S}^\dagger \mathsf{A} \mathsf{S} &= \frac{1}{2} \begin{pmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 \\ 0 & -2 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 \\ 0 & \sqrt{2} & 0 \\ 1 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 4 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \end{aligned}$$

which is diagonal, as required, and has as its diagonal elements the eigenvalues of A. ◄

If a matrix A is diagonalised by the similarity transformation $\mathsf{A}' = \mathsf{S}^{-1} \mathsf{A} \mathsf{S}$, so that $\mathsf{A}' = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$, then we have immediately

$$\mathrm{Tr}\, \mathsf{A}' = \mathrm{Tr}\, \mathsf{A} = \sum_{i=1}^{N} \lambda_i, \tag{8.102}$$

$$|\mathsf{A}'| = |\mathsf{A}| = \prod_{i=1}^{N} \lambda_i, \tag{8.103}$$

since the eigenvalues of the matrix are unchanged by the transformation. Moreover, these results may be used to prove the rather useful *trace formula*

$$|\exp \mathsf{A}| = \exp(\mathrm{Tr}\, \mathsf{A}), \tag{8.104}$$

where the exponential of a matrix is as defined in (8.38).

►*Prove the trace formula (8.104).*

At the outset, we note that for the similarity transformation $\mathsf{A}' = \mathsf{S}^{-1} \mathsf{A} \mathsf{S}$, we have

$$(\mathsf{A}')^n = (\mathsf{S}^{-1} \mathsf{A} \mathsf{S})(\mathsf{S}^{-1} \mathsf{A} \mathsf{S}) \cdots (\mathsf{S}^{-1} \mathsf{A} \mathsf{S}) = \mathsf{S}^{-1} \mathsf{A}^n \mathsf{S}.$$

Thus, from (8.38), we obtain $\exp \mathsf{A}' = \mathsf{S}^{-1}(\exp \mathsf{A})\mathsf{S}$, from which it follows that $|\exp \mathsf{A}'| =$

$|\exp \mathsf{A}|$. Moreover, by choosing the similarity transformation so that it diagonalises A, we have $\mathsf{A}' = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$, and so

$$|\exp \mathsf{A}| = |\exp \mathsf{A}'| = |\exp[\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)]| = |\text{diag}(\exp \lambda_1, \exp \lambda_2, \ldots, \exp \lambda_N)| = \prod_{i=1}^{N} \exp \lambda_i.$$

Rewriting the final product of exponentials of the eigenvalues as the exponential of the sum of the eigenvalues, we find

$$|\exp\ \mathsf{A}| = \prod_{i=1}^{N} \exp\ \lambda_i = \exp\left(\sum_{i=1}^{N} \lambda_i\right) = \exp(\text{Tr}\ \mathsf{A}),$$

which gives the trace formula (8.104). ◄

8.17 Quadratic and Hermitian forms

Let us now introduce the concept of quadratic forms (and their complex analogues, Hermitian forms). A quadratic form Q is a scalar function of a real vector $\mathbf{x}$ given by

$$Q(\mathbf{x}) = \langle \mathbf{x} | \mathcal{A}\, \mathbf{x} \rangle, \tag{8.105}$$

for some real linear operator $\mathcal{A}$. In any given basis (coordinate system) we can write (8.105) in matrix form as

$$Q(\mathsf{x}) = \mathsf{x}^{\mathrm{T}} \mathsf{A} \mathsf{x}, \tag{8.106}$$

where A is a real matrix. In fact, as will be explained below, we need only consider the case where A is symmetric, i.e. $\mathsf{A} = \mathsf{A}^{\mathrm{T}}$. As an example in a three-dimensional space,

$$\begin{aligned} Q = \mathsf{x}^{\mathrm{T}} \mathsf{A} \mathsf{x} &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 3 \\ 1 & 1 & -3 \\ 3 & -3 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x_1^2 + x_2^2 - 3x_3^2 + 2x_1x_2 + 6x_1x_3 - 6x_2x_3. \end{aligned} \tag{8.107}$$

It is reasonable to ask whether a quadratic form $Q = \mathsf{x}^{\mathrm{T}} \mathsf{M} \mathsf{x}$, where M is any (possibly non-symmetric) real square matrix, is a more general definition. That this is not the case may be seen by expressing M in terms of a symmetric matrix $\mathsf{A} = \frac{1}{2}(\mathsf{M} + \mathsf{M}^{\mathrm{T}})$ and an antisymmetric matrix $\mathsf{B} = \frac{1}{2}(\mathsf{M} - \mathsf{M}^{\mathrm{T}})$ such that $\mathsf{M} = \mathsf{A} + \mathsf{B}$. We then have

$$Q = \mathsf{x}^{\mathrm{T}} \mathsf{M} \mathsf{x} = \mathsf{x}^{\mathrm{T}} \mathsf{A} \mathsf{x} + \mathsf{x}^{\mathrm{T}} \mathsf{B} \mathsf{x}. \tag{8.108}$$

However, Q is a scalar quantity and so

$$Q = Q^{\mathrm{T}} = (\mathsf{x}^{\mathrm{T}} \mathsf{A} \mathsf{x})^{\mathrm{T}} + (\mathsf{x}^{\mathrm{T}} \mathsf{B} \mathsf{x})^{\mathrm{T}} = \mathsf{x}^{\mathrm{T}} \mathsf{A}^{\mathrm{T}} \mathsf{x} + \mathsf{x}^{\mathrm{T}} \mathsf{B}^{\mathrm{T}} \mathsf{x} = \mathsf{x}^{\mathrm{T}} \mathsf{A} \mathsf{x} - \mathsf{x}^{\mathrm{T}} \mathsf{B} \mathsf{x}. \tag{8.109}$$

Comparing (8.108) and (8.109) shows that $\mathsf{x}^{\mathrm{T}} \mathsf{B} \mathsf{x} = 0$, and hence $\mathsf{x}^{\mathrm{T}} \mathsf{M} \mathsf{x} = \mathsf{x}^{\mathrm{T}} \mathsf{A} \mathsf{x}$, i.e. Q is unchanged by considering only the symmetric part of M. Hence, with no loss of generality, we may assume $\mathsf{A} = \mathsf{A}^{\mathrm{T}}$ in (8.106).

From its definition (8.105), Q is clearly a basis- (i.e. coordinate-) independent quantity. Let us therefore consider a new basis related to the old one by an orthogonal transformation matrix S, the components in the two bases of any vector $\mathbf{x}$ being related (as in (8.91)) by $\mathsf{x} = \mathsf{S}\mathsf{x}'$ or, equivalently, by $\mathsf{x}' = \mathsf{S}^{-1}\mathsf{x} = \mathsf{S}^{\mathrm{T}}\mathsf{x}$. We then have

$$Q = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} = (\mathsf{x}')^{\mathrm{T}}\mathsf{S}^{\mathrm{T}}\mathsf{A}\mathsf{S}\mathsf{x}' = (\mathsf{x}')^{\mathrm{T}}\mathsf{A}'\mathsf{x}',$$

where (as expected) the matrix describing the linear operator $\mathcal{A}$ in the new basis is given by $\mathsf{A}' = \mathsf{S}^{\mathrm{T}}\mathsf{A}\mathsf{S}$ (since $\mathsf{S}^{\mathrm{T}} = \mathsf{S}^{-1}$). But, from the last section, if we choose as S the matrix whose columns are the *normalised* eigenvectors of A then $\mathsf{A}' = \mathsf{S}^{\mathrm{T}}\mathsf{A}\mathsf{S}$ is diagonal with the eigenvalues of A as the diagonal elements. (Since A is symmetric, its normalised eigenvectors are orthogonal, or can be made so, and hence S is orthogonal with $\mathsf{S}^{-1} = \mathsf{S}^{\mathrm{T}}$.)

In the new basis

$$Q = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} = (\mathsf{x}')^{\mathrm{T}}\Lambda\mathsf{x}' = \lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \cdots + \lambda_N x_N'^2, \tag{8.110}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and the λ_i are the eigenvalues of A. It should be noted that Q contains no cross-terms of the form $x_1'x_2'$.

►*Find an orthogonal transformation that takes the quadratic form (8.107) into the form*

$$\lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \lambda_3 x_3'^2.$$

The required transformation matrix S has the *normalised* eigenvectors of A as its columns. We have already found these in section 8.14, and so we can write immediately

$$\mathsf{S} = \frac{1}{\sqrt{6}}\begin{pmatrix} \sqrt{3} & \sqrt{2} & 1 \\ \sqrt{3} & -\sqrt{2} & -1 \\ 0 & \sqrt{2} & -2 \end{pmatrix},$$

which is easily verified as being orthogonal. Since the eigenvalues of A are $\lambda = 2$, 3, and -6, the general result already proved shows that the transformation $\mathsf{x} = \mathsf{S}\mathsf{x}'$ will carry (8.107) into the form $2x_1'^2 + 3x_2'^2 - 6x_3'^2$. This may be verified most easily by writing out the inverse transformation $\mathsf{x}' = \mathsf{S}^{-1}\mathsf{x} = \mathsf{S}^{\mathrm{T}}\mathsf{x}$ and substituting. The inverse equations are

$$\begin{aligned} x_1' &= (x_1 + x_2)/\sqrt{2}, \\ x_2' &= (x_1 - x_2 + x_3)/\sqrt{3}, \\ x_3' &= (x_1 - x_2 - 2x_3)/\sqrt{6}. \end{aligned} \tag{8.111}$$

If these are substituted into the form $Q = 2x_1'^2 + 3x_2'^2 - 6x_3'^2$ then the original expression (8.107) is recovered. ◄

In the definition of Q it was assumed that the components x_1, x_2, x_3 and the matrix A were real. It is clear that in this case the quadratic form $Q \equiv \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}$ is real also. Another, rather more general, expression that is also real is the *Hermitian form*

$$H(\mathsf{x}) \equiv \mathsf{x}^\dagger \mathsf{A} \mathsf{x}, \tag{8.112}$$

where A is Hermitian (i.e. $\mathsf{A}^\dagger = \mathsf{A}$) and the components of x may now be complex. It is straightforward to show that H is real, since

$$H^* = (H^{\mathrm{T}})^* = \mathsf{x}^\dagger \mathsf{A}^\dagger \mathsf{x} = \mathsf{x}^\dagger \mathsf{A} \mathsf{x} = H.$$

With suitable generalisation, the properties of quadratic forms apply also to Hermitian forms, but to keep the presentation simple we will restrict our discussion to quadratic forms.

A special case of a quadratic (Hermitian) form is one for which $Q = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}$ is greater than zero for all column matrices x. By choosing as the basis the eigenvectors of A we have Q in the form

$$Q = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2.$$

The requirement that $Q > 0$ for all x means that all the eigenvalues λ_i of A must be positive. A symmetric (Hermitian) matrix A with this property is called *positive definite*. If, instead, $Q \geq 0$ for all x then it is possible that some of the eigenvalues are zero, and A is called *positive semi-definite*.

8.17.1 The stationary properties of the eigenvectors

Consider a quadratic form, such as $Q(\mathbf{x}) = \langle \mathbf{x} | \mathcal{A}\,\mathbf{x} \rangle$, equation (8.105), in a fixed basis. As the vector **x** is varied, through changes in its three components x_1, x_2 and x_3, the value of the quantity Q also varies. Because of the homogeneous form of Q we may restrict any investigation of these variations to vectors of unit length (since multiplying any vector **x** by any scalar k simply multiplies the value of Q by a factor k^2).

Of particular interest are any vectors **x** that make the value of the quadratic form a maximum or minimum. A necessary, but not sufficient, condition for this is that Q is stationary with respect to small variations $\Delta\mathbf{x}$ in **x**, whilst $\langle \mathbf{x} | \mathbf{x} \rangle$ is maintained at a constant value (unity).

In the chosen basis the quadratic form is given by $Q = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}$ and, using Lagrange undetermined multipliers to incorporate the variational constraints, we are led to seek solutions of

$$\Delta[\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} - \lambda(\mathsf{x}^{\mathrm{T}}\mathsf{x} - 1)] = 0. \tag{8.113}$$

This may be used directly, together with the fact that $(\Delta\mathsf{x}^{\mathrm{T}})\mathsf{A}\mathsf{x} = \mathsf{x}^{\mathrm{T}}\mathsf{A}\,\Delta\mathsf{x}$, since A is symmetric, to obtain

$$\mathsf{A}\mathsf{x} = \lambda\mathsf{x} \tag{8.114}$$

as the necessary condition that x must satisfy. If (8.114) is satisfied for some eigenvector x then the value of $Q(\mathsf{x})$ is given by

$$Q = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} = \mathsf{x}^{\mathrm{T}}\lambda\mathsf{x} = \lambda. \tag{8.115}$$

However, if x and y are eigenvectors corresponding to different eigenvalues then they are (or can be chosen to be) orthogonal. Consequently the expression $\mathsf{y}^{\mathrm{T}}\mathsf{A}\mathsf{x}$ is necessarily zero, since

$$\mathsf{y}^{\mathrm{T}}\mathsf{A}\mathsf{x} = \mathsf{y}^{\mathrm{T}}\lambda\mathsf{x} = \lambda\mathsf{y}^{\mathrm{T}}\mathsf{x} = 0. \tag{8.116}$$

Summarising, those column matrices x of unit magnitude that make the quadratic form Q stationary are eigenvectors of the matrix A, and the stationary value of Q is then equal to the corresponding eigenvalue. It is straightforward to see from the proof of (8.114) that, conversely, any eigenvector of A makes Q stationary.

Instead of maximising or minimising $Q = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}$ subject to the constraint $\mathsf{x}^{\mathrm{T}}\mathsf{x} = 1$, an equivalent procedure is to extremise the function

$$\lambda(\mathsf{x}) = \frac{\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{x}}.$$

►*Show that if $\lambda(\mathsf{x})$ is stationary then x is an eigenvector of A and $\lambda(\mathsf{x})$ is equal to the corresponding eigenvalue.*

We require $\Delta\lambda(\mathsf{x}) = 0$ with respect to small variations in x. Now

$$\begin{aligned}\Delta\lambda &= \frac{1}{(\mathsf{x}^{\mathrm{T}}\mathsf{x})^2}\left[(\mathsf{x}^{\mathrm{T}}\mathsf{x})\left(\Delta\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} + \mathsf{x}^{\mathrm{T}}\mathsf{A}\,\Delta\mathsf{x}\right) - \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}\left(\Delta\mathsf{x}^{\mathrm{T}}\mathsf{x} + \mathsf{x}^{\mathrm{T}}\Delta\mathsf{x}\right)\right]\\ &= \frac{2\Delta\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{x}} - 2\left(\frac{\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{x}}\right)\frac{\Delta\mathsf{x}^{\mathrm{T}}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{x}},\end{aligned}$$

since $\mathsf{x}^{\mathrm{T}}\mathsf{A}\,\Delta\mathsf{x} = (\Delta\mathsf{x}^{\mathrm{T}})\mathsf{A}\mathsf{x}$ and $\mathsf{x}^{\mathrm{T}}\Delta\mathsf{x} = (\Delta\mathsf{x}^{\mathrm{T}})\mathsf{x}$. Thus

$$\Delta\lambda = \frac{2}{\mathsf{x}^{\mathrm{T}}\mathsf{x}}\Delta\mathsf{x}^{\mathrm{T}}[\mathsf{A}\mathsf{x} - \lambda(\mathsf{x})\mathsf{x}].$$

Hence, if $\Delta\lambda = 0$ then $\mathsf{A}\mathsf{x} = \lambda(\mathsf{x})\mathsf{x}$, i.e. x is an eigenvector of A with eigenvalue $\lambda(\mathsf{x})$. ◄

Thus the eigenvalues of a symmetric matrix A are the values of the function

$$\lambda(\mathsf{x}) = \frac{\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{x}}$$

at its stationary points. The eigenvectors of A lie along those directions in space for which the quadratic form $Q = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}$ has stationary values, given a fixed magnitude for the vector x. Similar results hold for Hermitian matrices.

8.17.2 Quadratic surfaces

The results of the previous subsection may be turned round to state that the surface given by

$$\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} = \text{constant} = 1 \text{ (say)} \tag{8.117}$$

and called a *quadratic surface*, has stationary values of its radius (i.e. origin–surface distance) in those directions that are along the eigenvectors of A. More specifically, in three dimensions the quadratic surface $\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x} = 1$ has its principal axes along the three mutually perpendicular eigenvectors of A, and the squares of the corresponding principal radii are given by λ_i^{-1}, $i = 1, 2, 3$. As well as having this stationary property of the radius, a *principal axis* is characterised by the fact that any section of the surface perpendicular to it has some degree of symmetry about it. If the eigenvalues corresponding to any two principal axes are degenerate then the quadratic surface has rotational symmetry about the third principal axis and the choice of a pair of axes perpendicular to that axis is not uniquely defined.

►*Find the shape of the quadratic surface*

$$x_1^2 + x_2^2 - 3x_3^2 + 2x_1x_2 + 6x_1x_3 - 6x_2x_3 = 1.$$

If, instead of expressing the quadratic surface in terms of x_1, x_2, x_3, as in (8.107), we were to use the new variables x_1', x_2', x_3' defined in (8.111), for which the coordinate axes are along the three mutually perpendicular eigenvector directions $(1, 1, 0)$, $(1, -1, 1)$ and $(1, -1, -2)$, then the equation of the surface would take the form (see (8.110))

$$\frac{x_1'^{\,2}}{(1/\sqrt{2})^2} + \frac{x_2'^{\,2}}{(1/\sqrt{3})^2} - \frac{x_3'^{\,2}}{(1/\sqrt{6})^2} = 1.$$

Thus, for example, a section of the quadratic surface in the plane $x_3' = 0$, i.e. $x_1 - x_2 - 2x_3 = 0$, is an ellipse, with semi-axes $1/\sqrt{2}$ and $1/\sqrt{3}$. Similarly a section in the plane $x_1' = x_1 + x_2 = 0$ is a hyperbola. ◄

Clearly the simplest three-dimensional situation to visualise is that in which all the eigenvalues are positive, since then the quadratic surface is an ellipsoid.

8.18 Simultaneous linear equations

In physical applications we often encounter sets of simultaneous linear equations. In general we may have M equations in N unknowns $x_1, x_2, \ldots, x_N$ of the form

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1N}x_N &= b_1, \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2N}x_N &= b_2, \\ &\ \ \vdots \\ A_{M1}x_1 + A_{M2}x_2 + \cdots + A_{MN}x_N &= b_M, \end{aligned} \tag{8.118}$$

where the A_{ij} and b_i have known values. If all the b_i are zero then the system of equations is called *homogeneous*, otherwise it is *inhomogeneous*. Depending on the given values, this set of equations for the N unknowns x_1, x_2, $\dots$, x_N may have either a unique solution, no solution or infinitely many solutions. Matrix analysis may be used to distinguish between the possibilities. The set of equations may be expressed as a single matrix equation $\mathsf{A}\mathsf{x} = \mathsf{b}$, or, written out in full, as

$$\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{pmatrix}.$$

8.18.1 The range and null space of a matrix

As we discussed in section 8.2, we may interpret the matrix equation $\mathsf{A}\mathsf{x} = \mathsf{b}$ as representing, in some basis, the linear transformation $\mathcal{A}\,\mathbf{x} = \mathbf{b}$ of a vector $\mathbf{x}$ in an N-dimensional vector space V into a vector $\mathbf{b}$ in some other (in general different) M-dimensional vector space W.

In general the operator $\mathcal{A}$ will map *any* vector in V into some particular *subspace* of W, which may be the entire space. This subspace is called the *range* of $\mathcal{A}$ (or A) and its dimension is equal to the *rank* of A. Moreover, if $\mathcal{A}$ (and hence A) is *singular* then there exists some subspace of V that is mapped onto the zero vector $\mathbf{0}$ in W; that is, any vector $\mathbf{y}$ that lies in the subspace satisfies $\mathcal{A}\,\mathbf{y} = \mathbf{0}$. This subspace is called the *null space* of A and the dimension of this null space is called the *nullity* of A. We note that the matrix A *must* be singular if $M \neq N$ and *may* be singular even if $M = N$.

The dimensions of the range and the null space of a matrix are related through the fundamental relationship

$$\text{rank}\ \mathsf{A} + \text{nullity}\ \mathsf{A} = N, \tag{8.119}$$

where N is the number of original unknowns $x_1, x_2, \dots, x_N$.

►*Prove the relationship (8.119).*

As discussed in section 8.11, if the columns of an $M \times N$ matrix A are interpreted as the components, in a given basis, of N (M-component) vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$ then rank A is equal to the number of linearly independent vectors in this set (this number is also equal to the dimension of the vector space spanned by these vectors). Writing (8.118) in terms of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$, we have

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \cdots + x_N\mathbf{v}_N = \mathbf{b}. \tag{8.120}$$

From this expression, we immediately deduce that the range of A is merely the span of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$ and hence has dimension $r = \text{rank}\,\mathsf{A}$.

If a vector $\mathbf{y}$ lies in the null space of A then $\mathcal{A}\,\mathbf{y} = \mathbf{0}$, which we may write as

$$y_1\mathbf{v}_1 + y_2\mathbf{v}_2 + \cdots + y_N\mathbf{v}_N = \mathbf{0}. \tag{8.121}$$

As just shown above, however, only r ($\leq N$) of these vectors are linearly independent. By renumbering, if necessary, we may assume that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ form a linearly independent set; the remaining vectors, $\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_N$, can then be written as a linear superposition of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$. We are therefore free to choose the $N - r$ coefficients $y_{r+1}, y_{r+2}, \ldots, y_N$ arbitrarily and (8.121) will still be satisfied for some set of r coefficients $y_1, y_2, \ldots, y_r$ (which are not all zero). The dimension of the null space is therefore $N - r$, and this completes the proof of (8.119). ◂

Equation (8.119) has far-reaching consequences for the existence of solutions to sets of simultaneous linear equations such as (8.118). As mentioned previously, these equations may have *no solution*, a *unique solution* or *infinitely many solutions*. We now discuss these three cases in turn.

No solution

The system of equations possesses no solution unless $\mathbf{b}$ lies in the range of $\mathcal{A}$; in this case (8.120) will be satisfied for some $x_1, x_2, \ldots, x_N$. This in turn requires the set of vectors $\mathbf{b}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ to have the same span (see (8.8)) as $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$. In terms of matrices, this is equivalent to the requirement that the matrix A and the *augmented matrix*

$$\mathsf{M} = \begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1N} & b_1 \\ A_{21} & A_{22} & \ldots & A_{2N} & b_1 \\ \vdots & & \ddots & & \vdots \\ A_{M1} & A_{M2} & \ldots & A_{MN} & b_M \end{pmatrix}$$

have the *same* rank r. If this condition is satisfied then $\mathbf{b}$ does lie in the range of $\mathcal{A}$, and the set of equations (8.118) will have either a unique solution or infinitely many solutions. If, however, A and M have different ranks then there will be no solution.

A unique solution

If $\mathbf{b}$ lies in the range of $\mathcal{A}$ and if $r = N$ then all the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ in (8.120) are linearly independent and the equation has a *unique solution* $x_1, x_2, \ldots, x_N$.

Infinitely many solutions

If $\mathbf{b}$ lies in the range of $\mathcal{A}$ and if $r < N$ then only r of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ in (8.120) are linearly independent. We may therefore choose the coefficients of $n - r$ vectors in an arbitrary way, while still satisfying (8.120) for some set of coefficients $x_1, x_2, \ldots, x_N$. There are therefore *infinitely many solutions*, which span an $(n - r)$-dimensional vector space. We may also consider this space of solutions in terms of the null space of A: if $\mathbf{x}$ is some vector satisfying $\mathcal{A}\,\mathbf{x} = \mathbf{b}$ and $\mathbf{y}$ is *any* vector in the null space of $\mathcal{A}$ (i.e. $\mathcal{A}\,\mathbf{y} = \mathbf{0}$) then

$$\mathcal{A}(\mathbf{x} + \mathbf{y}) = \mathcal{A}\,\mathbf{x} + \mathcal{A}\,\mathbf{y} = \mathcal{A}\,\mathbf{x} + \mathbf{0} = \mathbf{b},$$

and so $\mathbf{x} + \mathbf{y}$ is also a solution. Since the null space is $(n - r)$-dimensional, so too is the space of solutions.

We may use the above results to investigate the special case of the solution of a *homogeneous* set of linear equations, for which $\mathbf{b} = \mathbf{0}$. Clearly the set *always* has the trivial solution $x_1 = x_2 = \cdots = x_n = 0$, and if $r = N$ this will be the only solution. If $r < N$, however, there are infinitely many solutions; they form the null space of A, which has dimension $n - r$. In particular, we note that if $M < N$ (i.e. there are fewer equations than unknowns) then $r < N$ automatically. Hence a set of *homogeneous* linear equations with fewer equations than unknowns *always* has infinitely many solutions.

8.18.2 N simultaneous linear equations in N unknowns

A special case of (8.118) occurs when $M = N$. In this case the matrix A is *square* and we have the same number of equations as unknowns. Since A is square, the condition $r = N$ corresponds to $|\mathsf{A}| \neq 0$ and the matrix A is *non-singular*. The case $r < N$ corresponds to $|\mathsf{A}| = 0$, in which case A is *singular*.

As mentioned above, the equations will have a solution provided b lies in the range of A. If this is true then the equations will possess a unique solution when $|\mathsf{A}| \neq 0$ or infinitely many solutions when $|\mathsf{A}| = 0$. There exist several methods for obtaining the solution(s). Perhaps the most elementary method is *Gaussian elimination*; this method is discussed in subsection 27.3.1, where we also address numerical subtleties such as equation interchange (pivoting). In this subsection, we will outline three further methods for solving a square set of simultaneous linear equations.

Direct inversion

Since A is square it will possess an inverse, provided $|\mathsf{A}| \neq 0$. Thus, if A is non-singular, we immediately obtain

$$\mathsf{x} = \mathsf{A}^{-1}\mathsf{b} \tag{8.122}$$

as the unique solution to the set of equations. However, if $\mathsf{b} = \mathsf{0}$ then we see immediately that the set of equations possesses only the trivial solution $\mathsf{x} = \mathsf{0}$. The direct inversion method has the advantage that, once A^{-1} has been calculated, one may obtain the solutions x corresponding to different vectors b_1, b_2, ... on the RHS, with little further work.

►*Show that the set of simultaneous equations*

$$\begin{aligned} 2x_1 + 4x_2 + 3x_3 &= 4,\\ x_1 - 2x_2 - 2x_3 &= 0,\\ -3x_1 + 3x_2 + 2x_3 &= -7, \end{aligned} \tag{8.123}$$

has a unique solution, and find that solution.

The simultaneous equations can be represented by the matrix equation $\mathsf{Ax} = \mathsf{b}$, i.e.

$$\begin{pmatrix} 2 & 4 & 3 \\ 1 & -2 & -2 \\ -3 & 3 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ -7 \end{pmatrix}.$$

As we have already shown that A^{-1} exists and have calculated it, see (8.59), it follows that $\mathsf{x} = \mathsf{A}^{-1}\mathsf{b}$ or, more explicitly, that

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \frac{1}{11}\begin{pmatrix} 2 & 1 & -2 \\ 4 & 13 & 7 \\ -3 & -18 & -8 \end{pmatrix}\begin{pmatrix} 4 \\ 0 \\ -7 \end{pmatrix} = \begin{pmatrix} 2 \\ -3 \\ 4 \end{pmatrix}. \tag{8.124}$$

Thus the unique solution is $x_1 = 2$, $x_2 = -3$, $x_3 = 4$. ◄

LU decomposition

Although conceptually simple, finding the solution by calculating A^{-1} can be computationally demanding, especially when N is large. In fact, as we shall now show, it is not necessary to perform the full inversion of A in order to solve the simultaneous equations $\mathsf{Ax} = \mathsf{b}$. Rather, we can perform a *decomposition* of the matrix into the product of a square *lower triangular* matrix L and a square *upper triangular* matrix U, which are such that

$$\mathsf{A} = \mathsf{LU}, \tag{8.125}$$

and then use the fact that triangular systems of equations can be solved very simply.

We must begin, therefore, by finding the matrices L and U such that (8.125) is satisfied. This may be achieved straightforwardly by writing out (8.125) in component form. For illustration, let us consider the 3×3 case. It is, in fact, always possible, and convenient, to take the diagonal elements of L as unity, so we have

$$\begin{aligned} \mathsf{A} &= \begin{pmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{pmatrix}\begin{pmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{pmatrix} \\ &= \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ L_{21}U_{11} & L_{21}U_{12} + U_{22} & L_{21}U_{13} + U_{23} \\ L_{31}U_{11} & L_{31}U_{12} + L_{32}U_{22} & L_{31}U_{13} + L_{32}U_{23} + U_{33} \end{pmatrix} \end{aligned} \tag{8.126}$$

The nine unknown elements of L and U can now be determined by equating the nine elements of (8.126) to those of the 3×3 matrix A. This is done in the particular order illustrated in the example below.

Once the matrices L and U have been determined, one can use the decomposition to solve the set of equations $\mathsf{Ax} = \mathsf{b}$ in the following way. From (8.125), we have $\mathsf{LUx} = \mathsf{b}$, but this can be written as *two* triangular sets of equations

$$\mathsf{Ly} = \mathsf{b} \qquad \text{and} \qquad \mathsf{Ux} = \mathsf{y},$$

where y is another column matrix to be determined. One may easily solve the first triangular set of equations for y, which is then substituted into the second set. The required solution x is then obtained readily from the second triangular set of equations. We note that, as with direct inversion, once the *LU* decomposition has been determined, one can solve for various RHS column matrices $\mathsf{b}_1, \mathsf{b}_2, \ldots$, with little extra work.

►*Use LU decomposition to solve the set of simultaneous equations (8.123).*

We begin the determination of the matrices L and U by equating the elements of the matrix in (8.126) with those of the matrix

$$\mathsf{A} = \begin{pmatrix} 2 & 4 & 3 \\ 1 & -2 & -2 \\ -3 & 3 & 2 \end{pmatrix}.$$

This is performed in the following order:

1st row:	$U_{11} = 2,$	$U_{12} = 4,$	$U_{13} = 3$
1st column:	$L_{21}U_{11} = 1,$	$L_{31}U_{11} = -3$	$\Rightarrow L_{21} = \frac{1}{2},\ L_{31} = -\frac{3}{2}$
2nd row:	$L_{21}U_{12} + U_{22} = -2$	$L_{21}U_{13} + U_{23} = -2$	$\Rightarrow U_{22} = -4,\ U_{23} = -\frac{7}{2}$
2nd column:	$L_{31}U_{12} + L_{32}U_{22} = 3$		$\Rightarrow L_{32} = -\frac{9}{4}$
3rd row:	$L_{31}U_{13} + L_{32}U_{23} + U_{33} = 2$		$\Rightarrow U_{33} = -\frac{11}{8}$

Thus we may write the matrix A as

$$\mathsf{A} = \mathsf{LU} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{3}{2} & -\frac{9}{4} & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 & 3 \\ 0 & -4 & -\frac{7}{2} \\ 0 & 0 & -\frac{11}{8} \end{pmatrix}.$$

We must now solve the set of equations $\mathsf{Ly} = \mathsf{b}$, which read

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{3}{2} & -\frac{9}{4} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ -7 \end{pmatrix}.$$

Since this set of equations is triangular, we quickly find

$$y_1 = 4, \quad y_2 = 0 - (\tfrac{1}{2})(4) = -2, \quad y_3 = -7 - (-\tfrac{3}{2})(4) - (-\tfrac{9}{4})(-2) = -\tfrac{11}{2}.$$

These values must then be substituted into the equations $\mathsf{Ux} = \mathsf{y}$, which read

$$\begin{pmatrix} 2 & 4 & 3 \\ 0 & -4 & -\frac{7}{2} \\ 0 & 0 & -\frac{11}{8} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ -2 \\ -\frac{11}{2} \end{pmatrix}.$$

This set of equations is also triangular, and we easily find the solution

$$x_1 = 2, \quad x_2 = -3, \quad x_3 = 4,$$

which agrees with the result found above by direct inversion. ◀

We note, in passing, that one can calculate both the inverse and the determinant of A from its LU decomposition. To find the inverse A^{-1}, one solves the system of equations $\mathsf{Ax} = \mathsf{b}$ repeatedly for the N different RHS column matrices $\mathsf{b} = \mathsf{e}_i$, $i = 1, 2, \ldots, N$, where e_i is the column matrix with its ith element equal to unity and the others equal to zero. The solution x in each case gives the corresponding column of A^{-1}. Evaluation of the determinant $|\mathsf{A}|$ is much simpler. From (8.125), we have

$$|\mathsf{A}| = |\mathsf{LU}| = |\mathsf{L}||\mathsf{U}|. \tag{8.127}$$

Since L and U are triangular, however, we see from (8.64) that their determinants are equal to the products of their diagonal elements. Since $L_{ii} = 1$ for all i, we thus find

$$|\mathsf{A}| = U_{11}U_{22}\cdots U_{NN} = \prod_{i=1}^{N} U_{ii}.$$

As an illustration, in the above example we find $|\mathsf{A}| = (2)(-4)(-11/8) = 11$, which, as it must, agrees with our earlier calculation (8.58).

Finally, we note that if the matrix A is symmetric and positive semi-definite then we can decompose it as

$$\mathsf{A} = \mathsf{LL}^\dagger, \tag{8.128}$$

where L is a lower triangular matrix whose diagonal elements are *not*, in general, equal to unity. This is known as a *Cholesky decomposition* (in the special case where A is real, the decomposition becomes $\mathsf{A} = \mathsf{LL}^{\mathrm{T}}$). The reason that we cannot set the diagonal elements of L equal to unity in this case is that we require the same number of independent elements in L as in A. The requirement that the matrix be positive semi-definite is easily derived by considering the Hermitian form (or quadratic form in the real case)

$$\mathsf{x}^\dagger\mathsf{Ax} = \mathsf{x}^\dagger\mathsf{LL}^\dagger\mathsf{x} = (\mathsf{L}^\dagger\mathsf{x})^\dagger(\mathsf{L}^\dagger\mathsf{x}).$$

Denoting the column matrix $\mathsf{L}^\dagger\mathsf{x}$ by y, we see that the last term on the RHS is $\mathsf{y}^\dagger\mathsf{y}$, which must be greater than or equal to zero. Thus, we require $\mathsf{x}^\dagger\mathsf{Ax} \geq 0$ for any arbitrary column matrix x, and so A must be positive semi-definite (see section 8.17).

We recall that the requirement that a matrix be positive semi-definite is equivalent to demanding that all the eigenvalues of A are positive or zero. If one of the eigenvalues of A is zero, however, then from (8.103) we have $|\mathsf{A}| = 0$ and so A is *singular*. Thus, if A is a non-singular matrix, it must be *positive definite* (rather than just positive semi-definite) in order to perform the Cholesky decomposition (8.128). In fact, in this case, the inability to find a matrix L that satisfies (8.128) implies that A cannot be positive definite.

The Cholesky decomposition can be applied in an analogous way to the *LU* decomposition discussed above, but we shall not explore it further.

Cramer's rule

An alternative method of solution is to use *Cramer's rule*, which also provides some insight into the nature of the solutions in the various cases. To illustrate this method let us consider a set of three equations in three unknowns,

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 &= b_1, \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 &= b_2, \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 &= b_3, \end{aligned} \tag{8.129}$$

which may be represented by the matrix equation $\mathsf{Ax} = \mathsf{b}$. We wish either to find the solution(s) x to these equations or to establish that there are no solutions. From result (vi) of subsection 8.9.1, the determinant $|\mathsf{A}|$ is unchanged by adding to its first column the combination

$$\frac{x_2}{x_1} \times (\text{second column of } |\mathsf{A}|) + \frac{x_3}{x_1} \times (\text{third column of } |\mathsf{A}|).$$

We thus obtain

$$|\mathsf{A}| = \begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} = \begin{vmatrix} A_{11} + (x_2/x_1)A_{12} + (x_3/x_1)A_{13} & A_{12} & A_{13} \\ A_{21} + (x_2/x_1)A_{22} + (x_3/x_1)A_{23} & A_{22} & A_{23} \\ A_{31} + (x_2/x_1)A_{32} + (x_3/x_1)A_{33} & A_{32} & A_{33} \end{vmatrix},$$

which, on substituting b_i/x_1 for the ith entry in the first column, yields

$$|\mathsf{A}| = \frac{1}{x_1}\begin{vmatrix} b_1 & A_{12} & A_{13} \\ b_2 & A_{22} & A_{23} \\ b_3 & A_{32} & A_{33} \end{vmatrix} = \frac{1}{x_1}\Delta_1.$$

The determinant Δ_1 is known as a *Cramer determinant*. Similar manipulations of the second and third columns of $|\mathsf{A}|$ yield x_2 and x_3, and so the full set of results reads

$$x_1 = \frac{\Delta_1}{|\mathsf{A}|}, \qquad x_2 = \frac{\Delta_2}{|\mathsf{A}|}, \qquad x_3 = \frac{\Delta_3}{|\mathsf{A}|}, \tag{8.130}$$

where

$$\Delta_1 = \begin{vmatrix} b_1 & A_{12} & A_{13} \\ b_2 & A_{22} & A_{23} \\ b_3 & A_{32} & A_{33} \end{vmatrix}, \quad \Delta_2 = \begin{vmatrix} A_{11} & b_1 & A_{13} \\ A_{21} & b_2 & A_{23} \\ A_{31} & b_3 & A_{33} \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} A_{11} & A_{12} & b_1 \\ A_{21} & A_{22} & b_2 \\ A_{31} & A_{32} & b_3 \end{vmatrix}.$$

It can be seen that each Cramer determinant Δ_i is simply $|\mathsf{A}|$ but with column i replaced by the RHS of the original set of equations. If $|\mathsf{A}| \neq 0$ then (8.130) gives the unique solution. The proof given here appears to fail if any of the solutions x_i is zero, but it can be shown that result (8.130) is valid even in such a case.

►*Use Cramer's rule to solve the set of simultaneous equations (8.123).*

Let us again represent these simultaneous equations by the matrix equation $\mathsf{A}\mathsf{x} = \mathsf{b}$, i.e.

$$\begin{pmatrix} 2 & 4 & 3 \\ 1 & -2 & -2 \\ -3 & 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \\ -7 \end{pmatrix}.$$

From (8.58), the determinant of A is given by $|\mathsf{A}| = 11$. Following the discussion given above, the three Cramer determinants are

$$\Delta_1 = \begin{vmatrix} 4 & 4 & 3 \\ 0 & -2 & -2 \\ -7 & 3 & 2 \end{vmatrix}, \quad \Delta_2 = \begin{vmatrix} 2 & 4 & 3 \\ 1 & 0 & -2 \\ -3 & -7 & 2 \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} 2 & 4 & 4 \\ 1 & -2 & 0 \\ -3 & 3 & -7 \end{vmatrix}.$$

These may be evaluated using the properties of determinants listed in subsection 8.9.1 and we find $\Delta_1 = 22$, $\Delta_2 = -33$ and $\Delta_3 = 44$. From (8.130) the solution to the equations (8.123) is given by

$$x_1 = \frac{22}{11} = 2, \qquad x_2 = \frac{-33}{11} = -3, \qquad x_3 = \frac{44}{11} = 4,$$

which agrees with the solution found in the previous example. ◄

At this point it is useful to consider each of the three equations (8.129) as representing a plane in three-dimensional Cartesian coordinates. Using result (7.42) of chapter 7, the sets of components of the vectors normal to the planes are (A_{11}, A_{12}, A_{13}), (A_{21}, A_{22}, A_{23}) and (A_{31}, A_{32}, A_{33}), and using (7.46) the perpendicular distances of the planes from the origin are given by

$$d_i = \frac{b_i}{\left(A_{i1}^2 + A_{i2}^2 + A_{i3}^2\right)^{1/2}} \qquad \text{for } i = 1, 2, 3.$$

Finding the solution(s) to the simultaneous equations above corresponds to finding the point(s) of intersection of the planes.

If there is a unique solution the planes intersect at only a single point. This happens if their normals are linearly independent vectors. Since the rows of A represent the directions of these normals, this requirement is equivalent to $|\mathsf{A}| \neq 0$. If $\mathsf{b} = (0 \ \ 0 \ \ 0)^{\mathrm{T}} = \mathbf{0}$ then all the planes pass through the origin and, since there is only a single solution to the equations, the origin is that solution.

Let us now turn to the cases where $|\mathsf{A}| = 0$. The simplest such case is that in which all three planes are parallel; this implies that the normals are all parallel and so A is of rank 1. Two possibilities exist:

(i) the planes are coincident, i.e. $d_1 = d_2 = d_3$, in which case there is an infinity of solutions;

(ii) the planes are not all coincident, i.e. $d_1 \neq d_2$ and/or $d_1 \neq d_3$ and/or $d_2 \neq d_3$, in which case there are no solutions.

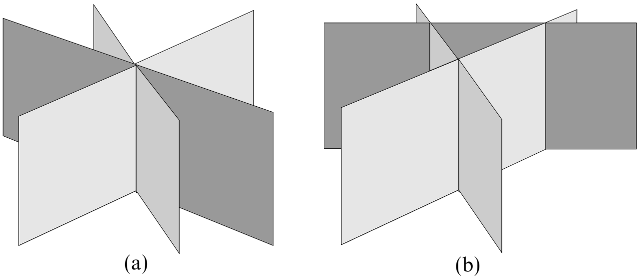


Figure 8.1 The two possible cases when A is of rank 2. In both cases all the normals lie in a horizontal plane but in (a) the planes all intersect on a single line (corresponding to an infinite number of solutions) whilst in (b) there are no common intersection points (no solutions).

It is apparent from (8.130) that case (i) occurs when all the Cramer determinants are zero and case (ii) occurs when at least one Cramer determinant is non-zero.

The most complicated cases with $|\mathsf{A}| = 0$ are those in which the normals to the planes themselves lie in a plane but are not parallel. In this case A has rank 2. Again two possibilities exist and these are shown in figure 8.1. Just as in the rank-1 case, if all the Cramer determinants are zero then we get an infinity of solutions (this time on a line). Of course, in the special case in which $\mathsf{b} = \mathsf{0}$ (and the system of equations is homogeneous), the planes all pass through the origin and so they must intersect on a line through it. If at least one of the Cramer determinants is non-zero, we get no solution.

These rules may be summarised as follows.

(i) $|\mathsf{A}| \neq 0$, $\mathsf{b} \neq \mathsf{0}$: The three planes intersect at a single point that is not the origin, and so there is only one solution, given by both (8.122) and (8.130).

(ii) $|\mathsf{A}| \neq 0$, $\mathsf{b} = \mathsf{0}$: The three planes intersect at the origin only and there is only the trivial solution, $\mathsf{x} = 0$.

(iii) $|\mathsf{A}| = 0$, $\mathsf{b} \neq \mathsf{0}$, Cramer determinants all zero: There is an infinity of solutions either on a line if A is rank 2, i.e. the cofactors are not all zero, or on a plane if A is rank 1, i.e. the cofactors are all zero.

(iv) $|\mathsf{A}| = 0$, $\mathsf{b} \neq \mathsf{0}$, Cramer determinants not all zero: No solutions.

(v) $|\mathsf{A}| = 0$, $\mathsf{b} = \mathsf{0}$: The three planes intersect on a line through the origin giving an infinity of solutions.

8.18.3 Singular value decomposition

There exists a very powerful technique for dealing with a simultaneous set of linear equations $\mathsf{Ax} = \mathsf{b}$, such as (8.118), which may be applied *whether or not* the number of simultaneous equations M is equal to the number of unknowns N. This technique is known as *singular value decomposition* (SVD) and is the method of choice in analysing *any* set of simultaneous linear equations.

We will consider the general case, in which A is an $M \times N$ (complex) matrix. Let us suppose we can write A as the product[§]

$$\mathsf{A} = \mathsf{U}\mathsf{S}\mathsf{V}^\dagger, \tag{8.131}$$

where the matrices U, S and V have the following properties.

(i) The square matrix U has dimensions $M \times M$ and is *unitary*.
(ii) The matrix S has dimensions $M \times N$ (the same dimensions as those of A) and is *diagonal* in the sense that $S_{ij} = 0$ if $i \neq j$. We denote its diagonal elements by s_i for $i = 1, 2, \ldots, p$, where $p = \min(M, N)$; these elements are termed the *singular values* of A.
(iii) The square matrix V has dimensions $N \times N$ and is *unitary*.

We must now determine the elements of these matrices in terms of the elements of A. From the matrix A, we can construct two square matrices: $\mathsf{A}^\dagger\mathsf{A}$ with dimensions $N \times N$ and $\mathsf{A}\mathsf{A}^\dagger$ with dimensions $M \times M$. Both are clearly *Hermitian*. From (8.131), and using the fact that U and V are unitary, we find

$$\mathsf{A}^\dagger\mathsf{A} = \mathsf{V}\mathsf{S}^\dagger\mathsf{U}^\dagger\mathsf{U}\mathsf{S}\mathsf{V}^\dagger = \mathsf{V}\mathsf{S}^\dagger\mathsf{S}\mathsf{V}^\dagger \tag{8.132}$$

$$\mathsf{A}\mathsf{A}^\dagger = \mathsf{U}\mathsf{S}\mathsf{V}^\dagger\mathsf{V}\mathsf{S}^\dagger\mathsf{U}^\dagger = \mathsf{U}\mathsf{S}\mathsf{S}^\dagger\mathsf{U}^\dagger, \tag{8.133}$$

where $\mathsf{S}^\dagger\mathsf{S}$ and $\mathsf{S}\mathsf{S}^\dagger$ are diagonal matrices with dimensions $N \times N$ and $M \times M$ respectively. The first p elements of each diagonal matrix are s_i^2, $i = 1, 2, \ldots, p$, where $p = \min(M, N)$, and the rest (where they exist) are zero.

These two equations imply that both $\mathsf{V}^{-1}\mathsf{A}^\dagger\mathsf{A}\mathsf{V}$ $\left(= \mathsf{V}^{-1}\mathsf{A}^\dagger\mathsf{A}(\mathsf{V}^\dagger)^{-1}\right)$ and, by a similar argument, $\mathsf{U}^{-1}\mathsf{A}\mathsf{A}^\dagger\mathsf{U}$, must be diagonal. From our discussion of the diagonalisation of Hermitian matrices in section 8.16, we see that the columns of V must therefore be the normalised eigenvectors v^i, $i = 1, 2, \ldots, N$, of the matrix $\mathsf{A}^\dagger\mathsf{A}$ and the columns of U must be the normalised eigenvectors u^j, $j = 1, 2, \ldots, M$, of the matrix $\mathsf{A}\mathsf{A}^\dagger$. Moreover, the singular values s_i must satisfy $s_i^2 = \lambda_i$, where the λ_i are the eigenvalues of the smaller of $\mathsf{A}^\dagger\mathsf{A}$ and $\mathsf{A}\mathsf{A}^\dagger$. Clearly, the λ_i are also some of the eigenvalues of the larger of these two matrices, the remaining ones being equal to zero. Since each matrix is Hermitian, the λ_i are real and the singular values s_i may be taken as real and non-negative. Finally, to make the decomposition (8.131) unique, it is customary to arrange the singular values in decreasing order of their values, so that $s_1 \geq s_2 \geq \cdots \geq s_p$.

[§] The proof that such a decomposition always exists is beyond the scope of this book. For a full account of SVD one might consult, for example, G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd edn (Baltimore MD: Johns Hopkins University Press, 1996).

►*Show that, for* $i = 1, 2, \ldots, p$, $\mathsf{A}\mathsf{v}^i = s_i\mathsf{u}^i$ *and* $\mathsf{A}^\dagger\mathsf{u}^i = s_i\mathsf{v}^i$, *where* $p = \min(M, N)$.

Post-multiplying both sides of (8.131) by V, and using the fact that V is unitary, we obtain

$$\mathsf{AV} = \mathsf{US}.$$

Since the columns of V and U consist of the vectors v^i and u^j respectively and S has only diagonal non-zero elements, we find immediately that, for $i = 1, 2, \ldots, p$,

$$\mathsf{A}\mathsf{v}^i = s_i\mathsf{u}^i. \qquad (8.134)$$

Moreover, we note that $\mathsf{A}\mathsf{v}^i = 0$ for $i = p+1, p+2, \ldots, N$.

Taking the Hermitian conjugate of both sides of (8.131) and post-multiplying by U, we obtain

$$\mathsf{A}^\dagger\mathsf{U} = \mathsf{V}\mathsf{S}^\dagger = \mathsf{V}\mathsf{S}^{\mathrm{T}},$$

where we have used the fact that U is unitary and S is real. We then see immediately that, for $i = 1, 2, \ldots, p$,

$$\mathsf{A}^\dagger\mathsf{u}^i = s_i\mathsf{v}^i. \qquad (8.135)$$

We also note that $\mathsf{A}^\dagger\mathsf{u}^i = 0$ for $i = p+1, p+2, \ldots, M$. Results (8.134) and (8.135) are useful for investigating the properties of the SVD. ◄

The decomposition (8.131) has some advantageous features for the analysis of sets of simultaneous linear equations. These are best illustrated by writing the decomposition (8.131) in terms of the vectors u^i and v^i as

$$\mathsf{A} = \sum_{i=1}^{p} s_i\mathsf{u}^i(\mathsf{v}^i)^\dagger,$$

where $p = \min(M, N)$. It may be, however, that some of the singular values s_i are *zero*, as a result of degeneracies in the set of M linear equations $\mathsf{Ax} = \mathsf{b}$. Let us suppose that there are r non-zero singular values. Since our convention is to arrange the singular values in order of decreasing size, the non-zero singular values are s_i, $i = 1, 2, \ldots, r$, and the zero singular values are $s_{r+1}, s_{r+2}, \ldots, s_p$. Therefore we can write A as

$$\mathsf{A} = \sum_{i=1}^{r} s_i\mathsf{u}^i(\mathsf{v}^i)^\dagger. \qquad (8.136)$$

Let us consider the action of (8.136) on an arbitrary vector x. This is given by

$$\mathsf{Ax} = \sum_{i=1}^{r} s_i\mathsf{u}^i(\mathsf{v}^i)^\dagger\mathsf{x}.$$

Since $(\mathsf{v}^i)^\dagger\mathsf{x}$ is just a number, we see immediately that the vectors u^i, $i = 1, 2, \ldots, r$, must span the *range* of the matrix A; moreover, these vectors form an orthonormal basis for the range. Further, since this subspace is r-dimensional, we have rank $\mathsf{A} = r$, i.e. the rank of A is equal to the number of non-zero singular values.

The SVD is also useful in characterising the null space of A. From (8.119), we already know that the null space must have dimension $N - r$; so, if A has r

non-zero singular values s_i, $i = 1, 2, \ldots, r$, then from the worked example above we have

$$\mathsf{A}\mathsf{v}^i = 0 \qquad \text{for } i = r+1, r+2, \ldots, N.$$

Thus, the $N - r$ vectors v^i, $i = r+1, r+2, \ldots, N$, form an orthonormal basis for the null space of A.

► *Find the singular value decompostion of the matrix*

$$\mathsf{A} = \begin{pmatrix} 2 & 2 & 2 & 2 \\ \frac{17}{10} & \frac{1}{10} & -\frac{17}{10} & -\frac{1}{10} \\ \frac{3}{5} & \frac{9}{5} & -\frac{3}{5} & -\frac{9}{5} \end{pmatrix}. \tag{8.137}$$

The matrix A has dimension 3×4 (i.e. $M = 3$, $N = 4$), and so we may construct from it the 3×3 matrix $\mathsf{A}\mathsf{A}^\dagger$ and the 4×4 matrix $\mathsf{A}^\dagger\mathsf{A}$ (in fact, since A is real, the Hermitian conjugates are just transposes). We begin by finding the eigenvalues λ_i and eigenvectors u^i of the smaller matrix $\mathsf{A}\mathsf{A}^\dagger$. This matrix is easily found to be given by

$$\mathsf{A}\mathsf{A}^\dagger = \begin{pmatrix} 16 & 0 & 0 \\ 0 & \frac{29}{5} & \frac{12}{5} \\ 0 & \frac{12}{5} & \frac{36}{5} \end{pmatrix},$$

and its characteristic equation reads

$$\begin{vmatrix} 16-\lambda & 0 & 0 \\ 0 & \frac{29}{5}-\lambda & \frac{12}{5} \\ 0 & \frac{12}{5} & \frac{36}{5}-\lambda \end{vmatrix} = (16-\lambda)(36 - 13\lambda + \lambda^2) = 0.$$

Thus, the eigenvalues are $\lambda_1 = 16$, $\lambda_2 = 9$, $\lambda_3 = 4$. Since the singular values of A are given by $s_i = \sqrt{\lambda_i}$ and the matrix S in (8.131) has the same dimensions as A, we have

$$\mathsf{S} = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \end{pmatrix}, \tag{8.138}$$

where we have arranged the singular values in order of decreasing size. Now the matrix U has as its columns the normalised eigenvectors u^i of the 3×3 matrix $\mathsf{A}\mathsf{A}^\dagger$. These normalised eigenvectors correspond to the eigenvalues of $\mathsf{A}\mathsf{A}^\dagger$ as follows:

$$\begin{aligned} \lambda_1 = 16 \quad &\Rightarrow \quad \mathsf{u}^1 = (1 \quad 0 \quad 0)^{\mathrm{T}} \\ \lambda_2 = 9 \quad &\Rightarrow \quad \mathsf{u}^2 = (0 \quad \tfrac{3}{5} \quad \tfrac{4}{5})^{\mathrm{T}} \\ \lambda_3 = 4 \quad &\Rightarrow \quad \mathsf{u}^3 = (0 \quad -\tfrac{4}{5} \quad \tfrac{3}{5})^{\mathrm{T}}, \end{aligned}$$

and so we obtain the matrix

$$\mathsf{U} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{5} & -\frac{4}{5} \\ 0 & \frac{4}{5} & \frac{3}{5} \end{pmatrix}. \tag{8.139}$$

The columns of the matrix V in (8.131) are the normalised eigenvectors of the 4×4 matrix $\mathsf{A}^\dagger\mathsf{A}$, which is given by

$$\mathsf{A}^\dagger\mathsf{A} = \frac{1}{4}\begin{pmatrix} 29 & 21 & 3 & 11 \\ 21 & 29 & 11 & 3 \\ 3 & 11 & 29 & 21 \\ 11 & 3 & 21 & 29 \end{pmatrix}.$$

We already know from the above discussion, however, that the non-zero eigenvalues of this matrix are *equal* to those of $\mathsf{AA}^\dagger$ found above, and that the remaining eigenvalue is *zero*. The corresponding normalised eigenvectors are easily found:

$$\begin{aligned}
\lambda_1 &= 16 &\Rightarrow\quad \mathsf{v}^1 &= \tfrac{1}{2}(1 \quad 1 \quad 1 \quad 1)^{\mathrm{T}} \\
\lambda_2 &= 9 &\Rightarrow\quad \mathsf{v}^2 &= \tfrac{1}{2}(1 \quad 1 \quad -1 \quad -1)^{\mathrm{T}} \\
\lambda_3 &= 4 &\Rightarrow\quad \mathsf{v}^3 &= \tfrac{1}{2}(-1 \quad 1 \quad 1 \quad -1)^{\mathrm{T}} \\
\lambda_4 &= 0 &\Rightarrow\quad \mathsf{v}^4 &= \tfrac{1}{2}(1 \quad -1 \quad 1 \quad -1)^{\mathrm{T}}
\end{aligned}$$

and so the matrix V is given by

$$\mathsf{V} = \frac{1}{2}\begin{pmatrix} 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix}. \tag{8.140}$$

Alternatively, we could have found the first three columns of V by using the relation (8.135) to obtain

$$\mathsf{v}^i = \frac{1}{s_i}\mathsf{A}^\dagger \mathsf{u}^i \qquad \text{for } i = 1, 2, 3.$$

The fourth eigenvector could then be found using the Gram–Schmidt orthogonalisation procedure. We note that if there were more than one eigenvector corresponding to a zero eigenvalue then we would need to use this procedure to orthogonalise these eigenvectors before constructing the matrix V.

Collecting our results together, we find the SVD of the matrix A:

$$\mathsf{A} = \mathsf{USV}^\dagger = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{5} & -\frac{4}{5} \\ 0 & \frac{4}{5} & \frac{3}{5} \end{pmatrix}\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix};$$

this can be verified by direct multiplication. ◀

Let us now consider the use of SVD in solving a set of M simultaneous linear equations in N unknowns, which we write again as $\mathsf{Ax} = \mathsf{b}$. Firstly, consider the solution of a homogeneous set of equations, for which $\mathsf{b} = \mathsf{0}$. As mentioned previously, if A is square and non-singular (and so possesses no zero singular values) then the equations have the unique trivial solution $\mathsf{x} = \mathsf{0}$. Otherwise, *any* of the vectors v^i, $i = r+1, r+2, \ldots, N$, or any linear combination of them, will be a solution.

In the inhomogeneous case, where b is not a zero vector, the set of equations will possess solutions if b lies in the range of A. To investigate these solutions, it is convenient to introduce the $N \times M$ matrix $\overline{\mathsf{S}}$, which is constructed by taking the transpose of S in (8.131) and replacing each non-zero singular value s_i on the diagonal by $1/s_i$. It is clear that, with this construction, $\mathsf{S}\overline{\mathsf{S}}$ is an $M \times M$ diagonal matrix with diagonal entries that equal unity for those values of j for which $s_j \neq 0$, and zero otherwise.

Now consider the vector

$$\hat{\mathsf{x}} = \mathsf{V}\overline{\mathsf{S}}\mathsf{U}^\dagger\mathsf{b}. \tag{8.141}$$

Using the unitarity of the matrices U and V, we find that

$$\mathsf{A}\hat{\mathsf{x}} - \mathsf{b} = \mathsf{U}\mathsf{S}\overline{\mathsf{S}}\mathsf{U}^\dagger\mathsf{b} - \mathsf{b} = \mathsf{U}(\mathsf{S}\overline{\mathsf{S}} - \mathsf{I})\mathsf{U}^\dagger\mathsf{b}. \qquad (8.142)$$

The matrix $(\mathsf{S}\overline{\mathsf{S}} - \mathsf{I})$ is diagonal and the jth element on its leading diagonal is non-zero (and equal to -1) only when $s_j = 0$. However, the jth element of the vector $\mathsf{U}^\dagger\mathsf{b}$ is given by the scalar product $(\mathsf{u}^j)^\dagger\mathsf{b}$; if b lies in the range of A, this scalar product can be non-zero only if $s_j \neq 0$. Thus the RHS of (8.142) must equal zero, and so $\hat{\mathsf{x}}$ given by (8.141) is a solution to the equations $\mathsf{A}\mathsf{x} = \mathsf{b}$. We may, however, add to this solution *any* linear combination of the $N-r$ vectors v^i, $i = r+1, r+2, \ldots, N$, that form an orthonormal basis for the null space of A; thus, in general, there exists an infinity of solutions (although it is straightforward to show that (8.141) is the solution vector of shortest length). The only way in which the solution (8.141) can be *unique* is if the rank r equals N, so that the matrix A does not possess a null space; this only occurs if A is square and non-singular.

If b does not lie in the range of A then the set of equations $\mathsf{A}\mathsf{x} = \mathsf{b}$ does not have a solution. Nevertheless, the vector (8.141) provides the closest possible 'solution' in a least-squares sense. In other words, although the vector (8.141) does not exactly solve $\mathsf{A}\mathsf{x} = \mathsf{b}$, it is the vector that minimises the *residual*

$$\epsilon = |\mathsf{A}\mathsf{x} - \mathsf{b}|,$$

where here the vertical lines denote the absolute value of the quantity they contain, not the determinant. This is proved as follows.

Suppose we were to add some arbitrary vector x' to the vector $\hat{\mathsf{x}}$ in (8.141). This would result in the addition of the vector $\mathsf{b}' = \mathsf{A}\mathsf{x}'$ to $\mathsf{A}\hat{\mathsf{x}} - \mathsf{b}$; b' is clearly in the range of A since any part of x' belonging to the null space of A contributes nothing to $\mathsf{A}\mathsf{x}'$. We would then have

$$\begin{aligned} |\mathsf{A}\hat{\mathsf{x}} - \mathsf{b} + \mathsf{b}'| &= |(\mathsf{U}\mathsf{S}\overline{\mathsf{S}}\mathsf{U}^\dagger - \mathsf{I})\mathsf{b} + \mathsf{b}'| \\ &= |\mathsf{U}[(\mathsf{S}\overline{\mathsf{S}} - \mathsf{I})\mathsf{U}^\dagger\mathsf{b} + \mathsf{U}^\dagger\mathsf{b}']| \\ &= |(\mathsf{S}\overline{\mathsf{S}} - \mathsf{I})\mathsf{U}^\dagger\mathsf{b} + \mathsf{U}^\dagger\mathsf{b}'|; \qquad (8.143) \end{aligned}$$

in the last line we have made use of the fact that the length of a vector is left unchanged by the action of the unitary matrix U. Now, the jth component of the vector $(\mathsf{S}\overline{\mathsf{S}} - \mathsf{I})\mathsf{U}^\dagger\mathsf{b}$ will only be non-zero when $s_j = 0$. However, the jth element of the vector $\mathsf{U}^\dagger\mathsf{b}'$ is given by the scalar product $(\mathsf{u}^j)^\dagger\mathsf{b}'$, which is non-zero only if $s_j \neq 0$, since b' lies in the range of A. Thus, as these two terms only contribute to (8.143) for two disjoint sets of j-values, its minimum value, as x' is varied, occurs when $\mathsf{b}' = \mathsf{0}$; this requires $\mathsf{x}' = \mathsf{0}$.

►*Find the solution(s) to the set of simultaneous linear equations* $\mathsf{A}\mathsf{x} = \mathsf{b}$, *where* A *is given by (8.137) and* $\mathsf{b} = (1 \quad 0 \quad 0)^{\mathrm{T}}$.

To solve the set of equations, we begin by calculating the vector given in (8.141),

$$\mathsf{x} = \mathsf{V}\overline{\mathsf{S}}\mathsf{U}^\dagger\mathsf{b},$$

where U and V are given by (8.139) and (8.140) respectively and $\overline{\mathsf{S}}$ is obtained by taking the transpose of S in (8.138) and replacing all the non-zero singular values s_i by $1/s_i$. Thus, $\overline{\mathsf{S}}$ reads

$$\overline{\mathsf{S}} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix}.$$

Substituting the appropriate matrices into the expression for x we find

$$\mathsf{x} = \tfrac{1}{8}(1 \quad 1 \quad 1 \quad 1)^{\mathrm{T}}. \tag{8.144}$$

It is straightforward to show that this solves the set of equations $\mathsf{Ax} = \mathsf{b}$ exactly, and so the vector $\mathsf{b} = (1 \quad 0 \quad 0)^{\mathrm{T}}$ must lie in the range of A. This is, in fact, immediately clear, since $\mathsf{b} = \mathsf{u}^1$. The solution (8.144) is *not*, however, unique. There are three non-zero singular values, but $N = 4$. Thus, the matrix A has a one-dimensional null space, which is 'spanned' by v^4, the fourth column of V, given in (8.140). The solutions to our set of equations, consisting of the sum of the exact solution and *any* vector in the null space of A, therefore lie along the line

$$\mathsf{x} = \tfrac{1}{8}(1 \quad 1 \quad 1 \quad 1)^{\mathrm{T}} + \alpha(1 \quad -1 \quad 1 \quad -1)^{\mathrm{T}},$$

where the parameter α can take any real value. We note that (8.144) is the point on this line that is closest to the origin. ◀

8.19 Exercises

8.1 Which of the following statements about linear vector spaces are true? Where a statement is false, give a counter-example to demonstrate this.

(a) Non-singular $N \times N$ matrices form a vector space of dimension N^2.
(b) Singular $N \times N$ matrices form a vector space of dimension N^2.
(c) Complex numbers form a vector space of dimension 2.
(d) Polynomial functions of x form an infinite-dimensional vector space.
(e) Series $\{a_0, a_1, a_2, \ldots, a_N\}$ for which $\sum_{n=0}^{N} |a_n|^2 = 1$ form an N-dimensional vector space.
(f) Absolutely convergent series form an infinite-dimensional vector space.
(g) Convergent series with terms of alternating sign form an infinite-dimensional vector space.

8.2 Evaluate the determinants

$$\text{(a)} \begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix}, \qquad \text{(b)} \begin{vmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & -2 & 1 \\ 3 & -3 & 4 & -2 \\ -2 & 1 & -2 & 1 \end{vmatrix}$$

and

$$\text{(c)} \begin{vmatrix} gc & ge & a+ge & gb+ge \\ 0 & b & b & b \\ c & e & e & b+e \\ a & b & b+f & b+d \end{vmatrix}.$$

8.3 Using the properties of determinants, solve with a minimum of calculation the following equations for x:

$$\text{(a)}\quad \begin{vmatrix} x & a & a & 1 \\ a & x & b & 1 \\ a & b & x & 1 \\ a & b & c & 1 \end{vmatrix} = 0, \qquad \text{(b)}\quad \begin{vmatrix} x+2 & x+4 & x-3 \\ x+3 & x & x+5 \\ x-2 & x-1 & x+1 \end{vmatrix} = 0.$$

8.4 Consider the matrices

$$\text{(a)}\ \mathsf{B} = \begin{pmatrix} 0 & -i & i \\ i & 0 & -i \\ -i & i & 0 \end{pmatrix}, \qquad \text{(b)}\ \mathsf{C} = \frac{1}{\sqrt{8}} \begin{pmatrix} \sqrt{3} & -\sqrt{2} & -\sqrt{3} \\ 1 & \sqrt{6} & -1 \\ 2 & 0 & 2 \end{pmatrix}.$$

Are they (i) real, (ii) diagonal, (iii) symmetric, (iv) antisymmetric, (v) singular, (vi) orthogonal, (vii) Hermitian, (viii) anti-Hermitian, (ix) unitary, (x) normal?

8.5 By considering the matrices

$$\mathsf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathsf{B} = \begin{pmatrix} 0 & 0 \\ 3 & 4 \end{pmatrix},$$

show that $\mathsf{AB} = \mathsf{0}$ does *not* imply that either A or B is the zero matrix, but that it does imply that at least one of them is singular.

8.6 This exercise considers a crystal whose unit cell has base vectors that are not necessarily mutually orthogonal.

(a) The basis vectors of the unit cell of a crystal, with the origin O at one corner, are denoted by $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$. The matrix G has elements G_{ij}, where $G_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$ and H_{ij} are the elements of the matrix $\mathsf{H} \equiv \mathsf{G}^{-1}$. Show that the vectors $\mathbf{f}_i = \sum_j H_{ij}\mathbf{e}_j$ are the reciprocal vectors and that $H_{ij} = \mathbf{f}_i \cdot \mathbf{f}_j$.

(b) If the vectors $\mathbf{u}$ and $\mathbf{v}$ are given by

$$\mathbf{u} = \sum_i u_i \mathbf{e}_i, \qquad \mathbf{v} = \sum_i v_i \mathbf{f}_i,$$

obtain expressions for $|\mathbf{u}|$, $|\mathbf{v}|$, and $\mathbf{u} \cdot \mathbf{v}$.

(c) If the basis vectors are each of length a and the angle between each pair is $\pi/3$, write down G and hence obtain H.

(d) Calculate (i) the length of the normal from O onto the plane containing the points $p^{-1}\mathbf{e}_1$, $q^{-1}\mathbf{e}_2$, $r^{-1}\mathbf{e}_3$, and (ii) the angle between this normal and $\mathbf{e}_1$.

8.7 Prove the following results involving Hermitian matrices:

(a) If A is Hermitian and U is unitary then $\mathsf{U}^{-1}\mathsf{AU}$ is Hermitian.

(b) If A is anti-Hermitian then $i\mathsf{A}$ is Hermitian.

(c) The product of two Hermitian matrices A and B is Hermitian if and only if A and B commute.

(d) If S is a real antisymmetric matrix then $\mathsf{A} = (\mathsf{I} - \mathsf{S})(\mathsf{I} + \mathsf{S})^{-1}$ is orthogonal. If A is given by

$$\mathsf{A} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

then find the matrix S that is needed to express A in the above form.

(e) If K is skew-hermitian, i.e. $\mathsf{K}^{\dagger} = -\mathsf{K}$, then $\mathsf{V} = (\mathsf{I} + \mathsf{K})(\mathsf{I} - \mathsf{K})^{-1}$ is unitary.

8.8 A and B are real non-zero 3×3 matrices and satisfy the equation

$$(\mathsf{AB})^{\mathrm{T}} + \mathsf{B}^{-1}\mathsf{A} = \mathsf{0}.$$

(a) Prove that if B is orthogonal then A is antisymmetric.

(b) Without assuming that B is orthogonal, prove that A is singular.

8.9 The *commutator* $[\mathsf{X}, \mathsf{Y}]$ of two matrices is defined by the equation

$$[\mathsf{X}, \mathsf{Y}] = \mathsf{XY} - \mathsf{YX}.$$

Two anticommuting matrices A and B satisfy

$$\mathsf{A}^2 = \mathsf{I}, \qquad \mathsf{B}^2 = \mathsf{I}, \qquad [\mathsf{A}, \mathsf{B}] = 2i\mathsf{C}.$$

(a) Prove that $\mathsf{C}^2 = \mathsf{I}$ and that $[\mathsf{B}, \mathsf{C}] = 2i\mathsf{A}$.

(b) Evaluate $[[[\mathsf{A}, \mathsf{B}], [\mathsf{B}, \mathsf{C}]], [\mathsf{A}, \mathsf{B}]]$.

8.10 The four matrices S_x, S_y, S_z and I are defined by

$$\mathsf{S}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathsf{S}_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},$$
$$\mathsf{S}_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathsf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where $i^2 = -1$. Show that $\mathsf{S}_x^2 = \mathsf{I}$ and $\mathsf{S}_x\mathsf{S}_y = i\mathsf{S}_z$, and obtain similar results by permutting x, y and z. Given that $\mathbf{v}$ is a vector with Cartesian components (v_x, v_y, v_z), the matrix $\mathsf{S}(\mathbf{v})$ is defined as

$$\mathsf{S}(\mathbf{v}) = v_x\mathsf{S}_x + v_y\mathsf{S}_y + v_z\mathsf{S}_z.$$

Prove that, for general non-zero vectors $\mathbf{a}$ and $\mathbf{b}$,

$$\mathsf{S}(\mathbf{a})\mathsf{S}(\mathbf{b}) = \mathbf{a} \cdot \mathbf{b}\,\mathsf{I} + i\,\mathsf{S}(\mathbf{a} \times \mathbf{b}).$$

Without further calculation, deduce that $\mathsf{S}(\mathbf{a})$ and $\mathsf{S}(\mathbf{b})$ commute if and only if $\mathbf{a}$ and $\mathbf{b}$ are parallel vectors.

8.11 A general triangle has angles α, β and γ and corresponding opposite sides a, b and c. Express the length of each side in terms of the lengths of the other two sides and the relevant cosines, writing the relationships in matrix and vector form, using the vectors having components a, b, c and $\cos\alpha, \cos\beta, \cos\gamma$. Invert the matrix and hence deduce the cosine-law expressions involving α, β and γ.

8.12 Given a matrix

$$\mathsf{A} = \begin{pmatrix} 1 & \alpha & 0 \\ \beta & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where α and β are non-zero complex numbers, find its eigenvalues and eigenvectors. Find the respective conditions for (a) the eigenvalues to be real and (b) the eigenvectors to be orthogonal. Show that the conditions are jointly satisfied if and only if A is Hermitian.

8.13 Using the Gram–Schmidt procedure:

(a) construct an orthonormal set of vectors from the following:

$$\mathsf{x}_1 = (0 \quad 0 \quad 1 \quad 1)^{\mathrm{T}}, \qquad \mathsf{x}_2 = (1 \quad 0 \quad -1 \quad 0)^{\mathrm{T}},$$
$$\mathsf{x}_3 = (1 \quad 2 \quad 0 \quad 2)^{\mathrm{T}}, \qquad \mathsf{x}_4 = (2 \quad 1 \quad 1 \quad 1)^{\mathrm{T}};$$

(b) find an orthonormal basis, within a four-dimensional Euclidean space, for the subspace spanned by the three vectors $(1\ \ 2\ \ 0\ \ 0)^{\mathrm{T}}$, $(3\ \ -1\ \ 2\ \ 0)^{\mathrm{T}}$ and $(0\ \ 0\ \ 2\ \ 1)^{\mathrm{T}}$.

8.14 If a unitary matrix U is written as $\mathsf{A} + i\mathsf{B}$, where A and B are Hermitian with non-degenerate eigenvalues, show the following:

(a) A and B commute;
(b) $\mathsf{A}^2 + \mathsf{B}^2 = \mathsf{I}$;
(c) The eigenvectors of A are also eigenvectors of B;
(d) The eigenvalues of U have unit modulus (as is necessary for any unitary matrix).

8.15 Determine which of the matrices below are mutually commuting, and, for those that are, demonstrate that they have a complete set of eigenvectors in common:

$$\mathsf{A} = \begin{pmatrix} 6 & -2 \\ -2 & 9 \end{pmatrix}, \qquad \mathsf{B} = \begin{pmatrix} 1 & 8 \\ 8 & -11 \end{pmatrix},$$
$$\mathsf{C} = \begin{pmatrix} -9 & -10 \\ -10 & 5 \end{pmatrix}, \qquad \mathsf{D} = \begin{pmatrix} 14 & 2 \\ 2 & 11 \end{pmatrix}.$$

8.16 Find the eigenvalues and a set of eigenvectors of the matrix

$$\begin{pmatrix} 1 & 3 & -1 \\ 3 & 4 & -2 \\ -1 & -2 & 2 \end{pmatrix}.$$

Verify that its eigenvectors are mutually orthogonal.

8.17 Find three real orthogonal column matrices, each of which is a simultaneous eigenvector of

$$\mathsf{A} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathsf{B} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

8.18 Use the results of the first worked example in section 8.14 to evaluate, without repeated matrix multiplication, the expression $\mathsf{A}^6\mathsf{x}$, where $\mathsf{x} = (2\ \ 4\ \ -1)^{\mathrm{T}}$ and A is the matrix given in the example.

8.19 Given that A is a real symmetric matrix with normalised eigenvectors e^i, obtain the coefficients α_i involved when column matrix x, which is the solution of

$$\mathsf{A}\mathsf{x} - \mu\mathsf{x} = \mathsf{v}, \qquad (*)$$

is expanded as $\mathsf{x} = \sum_i \alpha_i \mathsf{e}^i$. Here μ is a given constant and v is a given column matrix.

(a) Solve $(*)$ when

$$\mathsf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

$\mu = 2$ and $\mathsf{v} = (1\ \ 2\ \ 3)^{\mathrm{T}}$.

(b) Would $(*)$ have a solution if $\mu = 1$ and (i) $\mathsf{v} = (1\ \ 2\ \ 3)^{\mathrm{T}}$, (ii) $\mathsf{v} = (2\ \ 2\ \ 3)^{\mathrm{T}}$?

8.20 Demonstrate that the matrix

$$\mathsf{A} = \begin{pmatrix} 2 & 0 & 0 \\ -6 & 4 & 4 \\ 3 & -1 & 0 \end{pmatrix}$$

is defective, i.e. does not have three linearly independent eigenvectors, by showing the following:

(a) its eigenvalues are degenerate and, in fact, all equal;
(b) any eigenvector has the form $(\mu \quad (3\mu - 2\nu) \quad \nu)^{\mathrm{T}}$.
(c) if two pairs of values, μ_1, ν_1 and μ_2, ν_2, define two independent eigenvectors v_1 and v_2, then *any* third similarly defined eigenvector v_3 can be written as a linear combination of v_1 and v_2, i.e.

$$\mathsf{v}_3 = a\mathsf{v}_1 + b\mathsf{v}_2,$$

where

$$a = \frac{\mu_3\nu_2 - \mu_2\nu_3}{\mu_1\nu_2 - \mu_2\nu_1} \quad \text{and} \quad b = \frac{\mu_1\nu_3 - \mu_3\nu_1}{\mu_1\nu_2 - \mu_2\nu_1}.$$

Illustrate (c) using the example $(\mu_1, \nu_1) = (1, 1)$, $(\mu_2, \nu_2) = (1, 2)$ and $(\mu_3, \nu_3) = (0, 1)$.

Show further that any matrix of the form

$$\begin{pmatrix} 2 & 0 & 0 \\ 6n-6 & 4-2n & 4-4n \\ 3-3n & n-1 & 2n \end{pmatrix}$$

is defective, with the same eigenvalues and eigenvectors as A.

8.21 By finding the eigenvectors of the Hermitian matrix

$$\mathsf{H} = \begin{pmatrix} 10 & 3i \\ -3i & 2 \end{pmatrix},$$

construct a unitary matrix U such that $\mathsf{U}^{\dagger}\mathsf{H}\mathsf{U} = \Lambda$, where Λ is a real diagonal matrix.

8.22 Use the stationary properties of quadratic forms to determine the maximum and minimum values taken by the expression

$$Q = 5x^2 + 4y^2 + 4z^2 + 2xz + 2xy$$

on the unit sphere, $x^2 + y^2 + z^2 = 1$. For what values of x, y and z do they occur?

8.23 Given that the matrix

$$\mathsf{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

has two eigenvectors of the form $(1 \quad y \quad 1)^{\mathrm{T}}$, use the stationary property of the expression $J(\mathsf{x}) = \mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}/(\mathsf{x}^{\mathrm{T}}\mathsf{x})$ to obtain the corresponding eigenvalues. Deduce the third eigenvalue.

8.24 Find the lengths of the semi-axes of the ellipse

$$73x^2 + 72xy + 52y^2 = 100,$$

and determine its orientation.

8.25 The equation of a particular conic section is

$$Q \equiv 8x_1^2 + 8x_2^2 - 6x_1x_2 = 110.$$

Determine the type of conic section this represents, the orientation of its principal axes, and relevant lengths in the directions of these axes.

8.26 Show that the quadratic surface

$$5x^2 + 11y^2 + 5z^2 - 10yz + 2xz - 10xy = 4$$

is an ellipsoid with semi-axes of lengths 2, 1 and 0.5. Find the direction of its longest axis.

8.27 Find the direction of the axis of symmetry of the quadratic surface

$$7x^2 + 7y^2 + 7z^2 - 20yz - 20xz + 20xy = 3.$$

8.28 For the following matrices, find the eigenvalues and sufficient of the eigenvectors to be able to describe the quadratic surfaces associated with them:

$$\text{(a)} \begin{pmatrix} 5 & 1 & -1 \\ 1 & 5 & 1 \\ -1 & 1 & 5 \end{pmatrix}, \quad \text{(b)} \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}, \quad \text{(c)} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}.$$

8.29 This exercise demonstrates the reverse of the usual procedure of diagonalising a matrix.

(a) Rearrange the result $\mathsf{A}' = \mathsf{S}^{-1}\mathsf{A}\mathsf{S}$ of section 8.16 to express the original matrix A in terms of the unitary matrix S and the diagonal matrix A'. Hence show how to construct a matrix A that has given eigenvalues and given (orthogonal) column matrices as its eigenvectors.

(b) Find the matrix that has as eigenvectors $(1 \quad 2 \quad 1)^{\mathrm{T}}$, $(1 \quad -1 \quad 1)^{\mathrm{T}}$ and $(1 \quad 0 \quad -1)^{\mathrm{T}}$, with corresponding eigenvalues λ, μ and ν.

(c) Try a particular case, say $\lambda = 3$, $\mu = -2$ and $\nu = 1$, and verify by explicit solution that the matrix so found does have these eigenvalues.

8.30 Find an orthogonal transformation that takes the quadratic form

$$Q \equiv -x_1^2 - 2x_2^2 - x_3^2 + 8x_2x_3 + 6x_1x_3 + 8x_1x_2$$

into the form

$$\mu_1 y_1^2 + \mu_2 y_2^2 - 4y_3^2,$$

and determine μ_1 and μ_2 (see section 8.17).

8.31 One method of determining the nullity (and hence the rank) of an $M \times N$ matrix A is as follows.

- Write down an augmented transpose of A, by adding on the right an $N \times N$ unit matrix and thus producing an $N \times (M + N)$ array B.
- Subtract a suitable multiple of the first row of B from each of the other lower rows so as to make $B_{i1} = 0$ for $i > 1$.
- Subtract a suitable multiple of the second row (or the uppermost row that does not start with M zero values) from each of the other lower rows so as to make $B_{i2} = 0$ for $i > 2$.
- Continue in this way until all remaining rows have zeros in the first M places. The number of such rows is equal to the nullity of A, and the N rightmost entries of these rows are the components of vectors that span the null space. They can be made orthogonal if they are not so already.

Use this method to show that the nullity of

$$\mathsf{A} = \begin{pmatrix} -1 & 3 & 2 & 7 \\ 3 & 10 & -6 & 17 \\ -1 & -2 & 2 & -3 \\ 2 & 3 & -4 & 4 \\ 4 & 0 & -8 & -4 \end{pmatrix}$$

is 2 and that an orthogonal base for the null space of A is provided by any two column matrices of the form $(2+\alpha_i \quad -2\alpha_i \quad 1 \quad \alpha_i)^{\mathrm{T}}$, for which the α_i $(i=1,2)$ are real and satisfy $6\alpha_1\alpha_2 + 2(\alpha_1+\alpha_2) + 5 = 0$.

8.32 Do the following sets of equations have non-zero solutions? If so, find them.

(a) $3x + 2y + z = 0, \qquad x - 3y + 2z = 0, \qquad 2x + y + 3z = 0.$
(b) $2x = b(y + z), \qquad x = 2a(y - z), \qquad x = (6a - b)y - (6a + b)z.$

8.33 Solve the simultaneous equations

$$\begin{aligned} 2x + 3y + z &= 11,\\ x + y + z &= 6,\\ 5x - y + 10z &= 34. \end{aligned}$$

8.34 Solve the following simultaneous equations for x_1, x_2 and x_3, using matrix methods:

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 1,\\ 3x_1 + 4x_2 + 5x_3 &= 2,\\ x_1 + 3x_2 + 4x_3 &= 3. \end{aligned}$$

8.35 Show that the following equations have solutions only if $\eta = 1$ or 2, and find them in these cases:

$$\begin{aligned} x + y + z &= 1,\\ x + 2y + 4z &= \eta,\\ x + 4y + 10z &= \eta^2. \end{aligned}$$

8.36 Find the condition(s) on α such that the simultaneous equations

$$\begin{aligned} x_1 + \alpha x_2 &= 1,\\ x_1 - x_2 + 3x_3 &= -1,\\ 2x_1 - 2x_2 + \alpha x_3 &= -2 \end{aligned}$$

have (a) exactly one solution, (b) no solutions, or (c) an infinite number of solutions; give all solutions where they exist.

8.37 Make an LU decomposition of the matrix

$$\mathsf{A} = \begin{pmatrix} 3 & 6 & 9\\ 1 & 0 & 5\\ 2 & -2 & 16 \end{pmatrix}$$

and hence solve $\mathsf{Ax} = \mathsf{b}$, where (i) $\mathsf{b} = (21 \quad 9 \quad 28)^{\mathrm{T}}$, (ii) $\mathsf{b} = (21 \quad 7 \quad 22)^{\mathrm{T}}$.

8.38 Make an LU decomposition of the matrix

$$\mathsf{A} = \begin{pmatrix} 2 & -3 & 1 & 3\\ 1 & 4 & -3 & -3\\ 5 & 3 & -1 & -1\\ 3 & -6 & -3 & 1 \end{pmatrix}.$$

Hence solve $\mathsf{Ax} = \mathsf{b}$ for (i) $\mathsf{b} = (-4 \quad 1 \quad 8 \quad -5)^{\mathrm{T}}$, (ii) $\mathsf{b} = (-10 \quad 0 \quad -3 \quad -24)^{\mathrm{T}}$. Deduce that $\det \mathsf{A} = -160$ and confirm this by direct calculation.

8.39 Use the Cholesky separation method to determine whether the following matrices are positive definite. For each that is, determine the corresponding lower diagonal matrix L:

$$\mathsf{A} = \begin{pmatrix} 2 & 1 & 3\\ 1 & 3 & -1\\ 3 & -1 & 1 \end{pmatrix}, \qquad \mathsf{B} = \begin{pmatrix} 5 & 0 & \sqrt{3}\\ 0 & 3 & 0\\ \sqrt{3} & 0 & 3 \end{pmatrix}.$$

8.40 Find the equation satisfied by the squares of the singular values of the matrix associated with the following over-determined set of equations:

$$\begin{aligned} 2x + 3y + z &= 0 \\ x - y - z &= 1 \\ 2x + y &= 0 \\ 2y + z &= -2. \end{aligned}$$

Show that one of the singular values is close to zero. Determine the two larger singular values by an appropriate iteration process and the smallest one by indirect calculation.

8.41 Find the SVD of

$$\mathsf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 1 \\ -1 & 0 \end{pmatrix},$$

showing that the singular values are $\sqrt{3}$ and 1.

8.42 Find the SVD form of the matrix

$$\mathsf{A} = \begin{pmatrix} 22 & 28 & -22 \\ 1 & -2 & -19 \\ 19 & -2 & -1 \\ -6 & 12 & 6 \end{pmatrix}.$$

Use it to determine the best solution x of the equation $\mathsf{Ax} = \mathsf{b}$ when (i) $\mathsf{b} = (6 \quad -39 \quad 15 \quad 18)^{\mathrm{T}}$, (ii) $\mathsf{b} = (9 \quad -42 \quad 15 \quad 15)^{\mathrm{T}}$, showing that (i) has an exact solution, but that the best solution to (ii) has a residual of $\sqrt{18}$.

8.43 Four experimental measurements of particular combinations of three physical variables, x, y and z, gave the following inconsistent results:

$$\begin{aligned} 13x + 22y - 13z &= 4, \\ 10x - 8y - 10z &= 44, \\ 10x - 8y - 10z &= 47, \\ 9x - 18y - 9z &= 72. \end{aligned}$$

Find the SVD best values for x, y and z. Identify the null space of A and hence obtain the general SVD solution.

8.20 Hints and answers

8.1 (a) False. O_N, the $N \times N$ null matrix, is *not* non-singular.

(b) False. Consider the sum of $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$.

(c) True.

(d) True.

(e) False. Consider $b_n = a_n + a_n$ for which $\sum_{n=0}^{N} |b_n|^2 = 4 \neq 1$, or note that there is no zero vector with unit norm.

(f) True.

(g) False. Consider the two series defined by

$$a_0 = \tfrac{1}{2}, \qquad a_n = 2(-\tfrac{1}{2})^n \quad \text{for} \quad n \geq 1; \qquad b_n = -(-\tfrac{1}{2})^n \quad \text{for} \quad n \geq 0.$$

The series that is the sum of $\{a_n\}$ and $\{b_n\}$ does not have alternating signs and so closure does not hold.

8.3 (a) $x = a$, b or c; (b) $x = -1$; the equation is linear in x.

8.5 Use the property of the determinant of a matrix product.

8.7 (d) $\mathsf{S} = \begin{pmatrix} 0 & -\tan(\theta/2) \\ \tan(\theta/2) & 0 \end{pmatrix}$.

(e) Note that $(\mathsf{I}+\mathsf{K})(\mathsf{I}-\mathsf{K}) = \mathsf{I}-\mathsf{K}^2 = (\mathsf{I}-\mathsf{K})(\mathsf{I}+\mathsf{K})$.

8.9 (b) $32i\mathsf{A}$.

8.11 $a = b\cos\gamma + c\cos\beta$, and cyclic permutations; $a^2 = b^2 + c^2 - 2bc\cos\alpha$, and cyclic permutations.

8.13 (a) $2^{-1/2}(0 \;\; 0 \;\; 1 \;\; 1)^{\mathrm{T}}$, $6^{-1/2}(2 \;\; 0 \;\; -1 \;\; 1)^{\mathrm{T}}$, $39^{-1/2}(-1 \;\; 6 \;\; -1 \;\; 1)^{\mathrm{T}}$, $13^{-1/2}(2 \;\; 1 \;\; 2 \;\; -2)^{\mathrm{T}}$.
(b) $5^{-1/2}(1 \;\; 2 \;\; 0 \;\; 0)^{\mathrm{T}}$, $(345)^{-1/2}(14 \;\; -7 \;\; 10 \;\; 0)^{\mathrm{T}}$, $(18\,285)^{-1/2}(-56 \;\; 28 \;\; 98 \;\; 69)^{\mathrm{T}}$.

8.15 C does not commute with the others; A, B and D have $(1 \;\; -2)^{\mathrm{T}}$ and $(2 \;\; 1)^{\mathrm{T}}$ as common eigenvectors.

8.17 For A : $(1 \;\; 0 \;\; -1)^{\mathrm{T}}$, $(1 \;\; \alpha_1 \;\; 1)^{\mathrm{T}}$, $(1 \;\; \alpha_2 \;\; 1)^{\mathrm{T}}$.
For B : $(1 \;\; 1 \;\; 1)^{\mathrm{T}}$, $(\beta_1 \;\; \gamma_1 \;\; -\beta_1-\gamma_1)^{\mathrm{T}}$, $(\beta_2 \;\; \gamma_2 \;\; -\beta_2-\gamma_2)^{\mathrm{T}}$.
The α_i, β_i and γ_i are arbitrary.
Simultaneous and orthogonal: $(1 \;\; 0 \;\; -1)^{\mathrm{T}}$, $(1 \;\; 1 \;\; 1)^{\mathrm{T}}$, $(1 \;\; -2 \;\; 1)^{\mathrm{T}}$.

8.19 $\alpha_j = (\mathsf{v}\cdot\mathsf{e}^{j*})/(\lambda_j - \mu)$, where λ_j is the eigenvalue corresponding to e^j.

(a) $\mathsf{x} = (2 \;\; 1 \;\; 3)^{\mathrm{T}}$.
(b) Since μ is equal to one of A's eigenvalues λ_j, the equation only has a solution if $\mathsf{v}\cdot\mathsf{e}^{j*} = 0$; (i) no solution; (ii) $\mathsf{x} = (1 \;\; 1 \;\; 3/2)^{\mathrm{T}}$.

8.21 $\mathsf{U} = (10)^{-1/2}(1, 3i; 3i, 1)$, $\Lambda = (1, 0; 0, 11)$.

8.23 $J = (2y^2 - 4y + 4)/(y^2 + 2)$, with stationary values at $y = \pm\sqrt{2}$ and corresponding eigenvalues $2 \mp \sqrt{2}$. From the trace property of A, the third eigenvalue equals 2.

8.25 Ellipse; $\theta = \pi/4$, $a = \sqrt{22}$; $\theta = 3\pi/4$, $b = \sqrt{10}$.

8.27 The direction of the eigenvector having the unrepeated eigenvalue is $(1, 1, -1)/\sqrt{3}$.

8.29 (a) $\mathsf{A} = \mathsf{S}\mathsf{A}'\mathsf{S}^{\dagger}$, where S is the matrix whose columns are the eigenvectors of the matrix A to be constructed, and $\mathsf{A}' = \mathrm{diag}\,(\lambda, \mu, \nu)$.
(b) $\mathsf{A} = (\lambda + 2\mu + 3\nu,\ 2\lambda - 2\mu,\ \lambda + 2\mu - 3\nu;\ 2\lambda - 2\mu,\ 4\lambda + 2\mu,\ 2\lambda - 2\mu;$
$\lambda + 2\mu - 3\nu,\ 2\lambda - 2\mu,\ \lambda + 2\mu + 3\nu)$.
(c) $\frac{1}{3}(1, 5, -2; 5, 4, 5; -2, 5, 1)$.

8.31 The null space is spanned by $(2 \;\; 0 \;\; 1 \;\; 0)^{\mathrm{T}}$ and $(1 \;\; -2 \;\; 0 \;\; 1)^{\mathrm{T}}$.

8.33 $x = 3$, $y = 1$, $z = 2$.

8.35 First show that A is singular. $\eta = 1$, $x = 1 + 2z$, $y = -3z$; $\eta = 2$, $x = 2z$, $y = 1 - 3z$.

8.37 $L = (1, 0, 0; \frac{1}{3}, 1, 0; \frac{2}{3}, 3, 1)$, $\quad U = (3, 6, 9; 0, -2, 2; 0, 0, 4)$.

(i) $\mathsf{x} = (-1 \;\; 1 \;\; 2)^{\mathrm{T}}$. (ii) $\mathsf{x} = (-3 \;\; 2 \;\; 2)^{\mathrm{T}}$.

8.39 A is not positive definite, as L_{33} is calculated to be $\sqrt{-6}$.
$\mathsf{B} = \mathsf{L}\mathsf{L}^{\mathrm{T}}$, where the non-zero elements of L are
$L_{11} = \sqrt{5}$, $L_{31} = \sqrt{3/5}$, $L_{22} = \sqrt{3}$, $L_{33} = \sqrt{12/5}$.

8.41

$$\mathsf{A}^{\dagger}\mathsf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \ \mathsf{U} = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 & \sqrt{3} & \sqrt{2} \\ 2 & 0 & \sqrt{2} \\ -1 & -\sqrt{3} & \sqrt{2} \end{pmatrix}, \ \mathsf{V} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

8.43 The singular values are $12\sqrt{6}, 0, 18\sqrt{3}$ and the calculated best solution is $x = 1.71, y = -1.94, z = -1.71$. The null space is the line $x = z, y = 0$ and the general SVD solution is $x = 1.71 + \lambda$, $y = -1.94$, $z = -1.71 + \lambda$.

9

Normal modes

Any student of the physical sciences will encounter the subject of oscillations on many occasions and in a wide variety of circumstances, for example the voltage and current oscillations in an electric circuit, the vibrations of a mechanical structure and the internal motions of molecules. The matrices studied in the previous chapter provide a particularly simple way to approach what may appear, at first glance, to be difficult physical problems.

We will consider only systems for which a position-dependent potential exists, i.e., the potential energy of the system in any particular configuration depends upon the coordinates of the configuration, which need not be be lengths, however; the potential must *not* depend upon the time derivatives (generalised velocities) of these coordinates. So, for example, the potential $-q\mathbf{v}\cdot\mathbf{A}$ used in the Lagrangian description of a charged particle in an electromagnetic field is excluded. A further restriction that we place is that the potential has a local minimum at the equilibrium point; physically, this is a necessary and sufficient condition for stable equilibrium. By suitably defining the origin of the potential, we may take its value at the equilibrium point as zero.

We denote the coordinates chosen to describe a configuration of the system by q_i, $i = 1, 2, \ldots, N$. The q_i need not be distances; some could be angles, for example. For convenience we can define the q_i so that they are all zero at the equilibrium point. The instantaneous velocities of various parts of the system will depend upon the time derivatives of the q_i, denoted by $\dot{q}_i$. For small oscillations the velocities will be linear in the $\dot{q}_i$ and consequently the total kinetic energy T will be quadratic in them – and will include cross terms of the form $\dot{q}_i\dot{q}_j$ with $i \neq j$. The general expression for T can be written as the quadratic form

$$T = \sum_i \sum_j a_{ij}\dot{q}_i\dot{q}_j = \dot{\mathsf{q}}^{\mathrm{T}}\mathsf{A}\dot{\mathsf{q}}, \tag{9.1}$$

where $\dot{\mathsf{q}}$ is the column vector $(\dot{q}_1 \quad \dot{q}_2 \quad \cdots \quad \dot{q}_N)^{\mathrm{T}}$ and the $N \times N$ matrix A is real and may be chosen to be symmetric. Furthermore, A, like any matrix corresponding to a kinetic energy, is positive definite; that is, whatever non-zero real values the $\dot{q}_i$ take, the quadratic form (9.1) has a value > 0.

Turning now to the potential energy, we may write its value for a configuration q by means of a Taylor expansion about the origin $\mathsf{q} = \mathsf{0}$,

$$V(\mathsf{q}) = V(\mathsf{0}) + \sum_i \frac{\partial V(\mathsf{0})}{\partial q_i} q_i + \frac{1}{2} \sum_i \sum_j \frac{\partial^2 V(\mathsf{0})}{\partial q_i \partial q_j} q_i q_j + \cdots .$$

However, we have chosen $V(\mathsf{0}) = 0$ and, since the origin is an equilibrium point, there is no force there and $\partial V(\mathsf{0})/\partial q_i = 0$. Consequently, to second order in the q_i we also have a quadratic form, but in the coordinates rather than in their time derivatives:

$$V = \sum_i \sum_j b_{ij} q_i q_j = \mathsf{q}^{\mathrm{T}} \mathsf{B} \mathsf{q}, \tag{9.2}$$

where B is, or can be made, symmetric. In this case, and in general, the requirement that the potential is a minimum means that the potential matrix B, like the kinetic energy matrix A, is real and positive definite.

9.1 Typical oscillatory systems

We now introduce particular examples, although the results of this section are general, given the above restrictions, and the reader will find it easy to apply the results to many other instances.

Consider first a uniform rod of mass M and length l, attached by a light string also of length l to a fixed point P and executing small oscillations in a vertical plane. We choose as coordinates the angles θ_1 and θ_2 shown, with exaggerated magnitude, in figure 9.1. In terms of these coordinates the centre of gravity of the rod has, to *first order* in the θ_i, a velocity component in the x-direction equal to $l\dot{\theta}_1 + \frac{1}{2} l\dot{\theta}_2$ and in the y-direction equal to zero. Adding in the rotational kinetic energy of the rod about its centre of gravity we obtain, to second order in the $\dot{\theta}_i$,

$$\begin{aligned} T &\approx \tfrac{1}{2} M l^2 (\dot{\theta}_1^2 + \tfrac{1}{4}\dot{\theta}_2^2 + \dot{\theta}_1\dot{\theta}_2) + \tfrac{1}{24} M l^2 \dot{\theta}_2^2 \\ &= \tfrac{1}{6} M l^2 \left(3\dot{\theta}_1^2 + 3\dot{\theta}_1\dot{\theta}_2 + \dot{\theta}_2^2\right) = \tfrac{1}{12} M l^2 \dot{\mathsf{q}}^{\mathrm{T}} \begin{pmatrix} 6 & 3 \\ 3 & 2 \end{pmatrix} \dot{\mathsf{q}}, \end{aligned} \tag{9.3}$$

where $\dot{\mathsf{q}}^{\mathrm{T}} = (\dot{\theta}_1 \quad \dot{\theta}_2)$. The potential energy is given by

$$V = Mlg\left[(1 - \cos\theta_1) + \tfrac{1}{2}(1 - \cos\theta_2)\right] \tag{9.4}$$

so that

$$V \approx \tfrac{1}{4} Mlg(2\theta_1^2 + \theta_2^2) = \tfrac{1}{12} Mlg \mathsf{q}^{\mathrm{T}} \begin{pmatrix} 6 & 0 \\ 0 & 3 \end{pmatrix} \mathsf{q}, \tag{9.5}$$

where g is the acceleration due to gravity and $\mathsf{q} = (\theta_1 \quad \theta_2)^{\mathrm{T}}$; (9.5) is valid to second order in the θ_i.

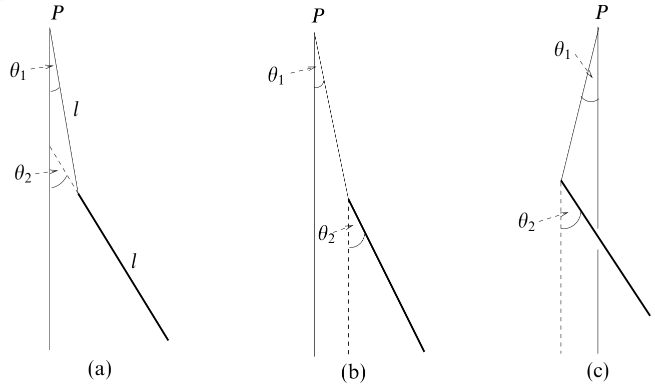


Figure 9.1 A uniform rod of length l attached to the fixed point P by a light string of the same length: (a) the general coordinate system; (b) approximation to the normal mode with lower frequency; (c) approximation to the mode with higher frequency.

With these expressions for T and V we now apply the conservation of energy,

$$\frac{d}{dt}(T+V)=0, \tag{9.6}$$

assuming that there are no external forces other than gravity. In matrix form (9.6) becomes

$$\frac{d}{dt}(\dot{\mathsf{q}}^{\mathrm{T}}\mathsf{A}\dot{\mathsf{q}}+\mathsf{q}^{\mathrm{T}}\mathsf{B}\mathsf{q})=\ddot{\mathsf{q}}^{\mathrm{T}}\mathsf{A}\dot{\mathsf{q}}+\dot{\mathsf{q}}^{\mathrm{T}}\mathsf{A}\ddot{\mathsf{q}}+\dot{\mathsf{q}}^{\mathrm{T}}\mathsf{B}\mathsf{q}+\mathsf{q}^{\mathrm{T}}\mathsf{B}\dot{\mathsf{q}}=0,$$

which, using $\mathsf{A}=\mathsf{A}^{\mathrm{T}}$ and $\mathsf{B}=\mathsf{B}^{\mathrm{T}}$, gives

$$2\dot{\mathsf{q}}^{\mathrm{T}}(\mathsf{A}\ddot{\mathsf{q}}+\mathsf{B}\mathsf{q})=0.$$

We will assume, although it is not clear that this gives the only possible solution, that the above equation implies that the coefficient of each $\dot{q}_i$ is separately zero. Hence

$$\mathsf{A}\ddot{\mathsf{q}}+\mathsf{B}\mathsf{q}=0. \tag{9.7}$$

For a rigorous derivation Lagrange's equations should be used, as in chapter 22.

Now we search for sets of coordinates q that *all* oscillate with the same period, i.e. the total motion repeats itself *exactly* after a *finite* interval. Solutions of this form will satisfy

$$\mathsf{q}=\mathsf{x}\cos\omega t; \tag{9.8}$$

the relative values of the elements of x in such a solution will indicate how each coordinate is involved in this special motion. In general there will be N values of ω if the matrices A and B are $N \times N$ and these values are known as *normal frequencies* or *eigenfrequencies.*

Putting (9.8) into (9.7) yields

$$-\omega^2\mathsf{A}\mathsf{x} + \mathsf{B}\mathsf{x} = (\mathsf{B} - \omega^2\mathsf{A})\mathsf{x} = \mathsf{0}. \tag{9.9}$$

Our work in section 8.18 showed that this can have non-trivial solutions only if

$$|\mathsf{B} - \omega^2\mathsf{A}| = 0. \tag{9.10}$$

This is a form of characteristic equation for B, except that the unit matrix I has been replaced by A. It has the more familiar form if a choice of coordinates is made in which the kinetic energy T is a simple sum of squared terms, i.e. it has been diagonalised, and the scale of the new coordinates is then chosen to make each diagonal element unity.

However, even in the present case, (9.10) can be solved to yield ω_k^2 for $k = 1, 2, \ldots, N$, where N is the order of A and B. The values of ω_k can be used with (9.9) to find the corresponding column vector x^k and the initial (stationary) physical configuration that, on release, will execute motion with period $2\pi/\omega_k$.

In equation (8.76) we showed that the eigenvectors of a real symmetric matrix were, except in the case of degeneracy of the eigenvalues, mutually orthogonal. In the present situation an analogous, but not identical, result holds. It is shown in section 9.3 that if x^1 and x^2 are two eigenvectors satisfying (9.9) for different values of ω^2 then they are orthogonal in the sense that

$$(\mathsf{x}^2)^{\mathrm{T}}\mathsf{A}\mathsf{x}^1 = 0 \qquad \text{and} \qquad (\mathsf{x}^2)^{\mathrm{T}}\mathsf{B}\mathsf{x}^1 = 0.$$

The direct ‘scalar product’ $(\mathsf{x}^2)^{\mathrm{T}}\mathsf{x}^1$, formally equal to $(\mathsf{x}^2)^{\mathrm{T}}\mathsf{I}\,\mathsf{x}^1$, is not, in general, equal to zero.

Returning to the suspended rod, we find from (9.10)

$$\left|\frac{Mlg}{12}\begin{pmatrix} 6 & 0 \\ 0 & 3 \end{pmatrix} - \frac{\omega^2 Ml^2}{12}\begin{pmatrix} 6 & 3 \\ 3 & 2 \end{pmatrix}\right| = 0.$$

Writing $\omega^2 l/g = \lambda$, this becomes

$$\begin{vmatrix} 6-6\lambda & -3\lambda \\ -3\lambda & 3-2\lambda \end{vmatrix} = 0 \qquad \Rightarrow \qquad \lambda^2 - 10\lambda + 6 = 0,$$

which has roots $\lambda = 5 \pm \sqrt{19}$. Thus we find that the two normal frequencies are given by $\omega_1 = (0.641g/l)^{1/2}$ and $\omega_2 = (9.359g/l)^{1/2}$. Putting the lower of the two values for ω^2, namely $(5 - \sqrt{19})g/l$, into (9.9) shows that for this mode

$$x_1 : x_2 = 3(5 - \sqrt{19}) : 6(\sqrt{19} - 4) = 1.923 : 2.153.$$

This corresponds to the case where the rod and string are almost straight out, i.e. they almost form a simple pendulum. Similarly it may be shown that the higher frequency corresponds to a solution where the string and rod are moving with opposite phase and $x_1 : x_2 = 9.359 : -16.718$. The two situations are shown in figure 9.1.

In connection with quadratic forms it was shown in section 8.17 how to make a change of coordinates such that the matrix for a particular form becomes diagonal. In exercise 9.6 a method is developed for diagonalising simultaneously two quadratic forms (though the transformation matrix may not be orthogonal). If this process is carried out for A and B in a general system undergoing stable oscillations, the kinetic and potential energies in the new variables η_i take the forms

$$T = \sum_i \mu_i \dot{\eta}_i^2 = \dot{\mathsf{\eta}}^{\mathrm{T}} \mathsf{M} \dot{\mathsf{\eta}}, \quad \mathsf{M} = \mathrm{diag}\,(\mu_1, \mu_2, \ldots, \mu_N), \tag{9.11}$$

$$V = \sum_i \nu_i \eta_i^2 = \mathsf{\eta}^{\mathrm{T}} \mathsf{N} \mathsf{\eta}, \quad \mathsf{N} = \mathrm{diag}\,(\nu_1, \nu_2 \ldots, \nu_N), \tag{9.12}$$

and the equations of motion are the *uncoupled* equations

$$\mu_i \ddot{\eta}_i + \nu_i \eta_i = 0, \qquad i = 1, 2, \ldots, N. \tag{9.13}$$

Clearly a simple renormalisation of the η_i can be made that reduces all the μ_i in (9.11) to unity. When this is done the variables so formed are called *normal coordinates* and equations (9.13) the *normal equations.*

When a system is executing one of these simple harmonic motions it is said to be in a *normal mode*, and once started in such a mode it will repeat its motion exactly after each interval of $2\pi/\omega_i$. Any arbitrary motion of the system may be written as a superposition of the normal modes, and each component mode will execute harmonic motion with the corresponding eigenfrequency; however, unless by chance the eigenfrequencies are in integer relationship, the system will never return to its initial configuration after any finite time interval.

As a second example we will consider a number of masses coupled together by springs. For this type of situation the potential and kinetic energies are automatically quadratic functions of the coordinates and their derivatives, provided the elastic limits of the springs are not exceeded, and the oscillations do not have to be vanishingly small for the analysis to be valid.

►*Find the normal frequencies and modes of oscillation of three particles of masses m, μ m, m connected in that order in a straight line by two equal light springs of force constant k. This arrangement could serve as a model for some linear molecules, e.g.* $\mathrm{CO_2}$.

The situation is shown in figure 9.2; the coordinates of the particles, x_1, x_2, x_3, are measured from their equilibrium positions, at which the springs are neither extended nor compressed.

The kinetic energy of the system is simply

$$T = \tfrac{1}{2} m \left(\dot{x}_1^2 + \mu \dot{x}_2^2 + \dot{x}_3^2 \right),$$

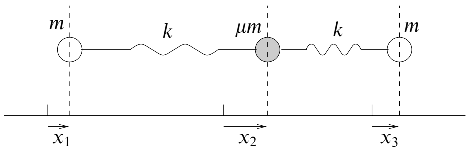


Figure 9.2 Three masses m, μm and m connected by two equal light springs of force constant k.

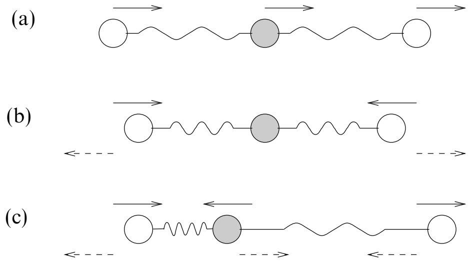


Figure 9.3 The normal modes of the masses and springs of a linear molecule such as CO_2. (a) $\omega^2 = 0$; (b) $\omega^2 = k/m$; (c) $\omega^2 = [(\mu + 2)/\mu](k/m)$.

whilst the potential energy stored in the springs is

$$V = \tfrac{1}{2}k\left[(x_2 - x_1)^2 + (x_3 - x_2)^2\right].$$

The kinetic- and potential-energy symmetric matrices are thus

$$\mathsf{A} = \frac{m}{2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathsf{B} = \frac{k}{2}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

From (9.10), to find the normal frequencies we have to solve $|\mathsf{B} - \omega^2\mathsf{A}| = 0$. Thus, writing $m\omega^2/k = \lambda$, we have

$$\begin{vmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\mu\lambda & -1 \\ 0 & -1 & 1-\lambda \end{vmatrix} = 0,$$

which leads to $\lambda = 0$, 1 or $1 + 2/\mu$. The corresponding eigenvectors are respectively

$$\mathsf{x}^1 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathsf{x}^2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \qquad \mathsf{x}^3 = \frac{1}{\sqrt{2 + (4/\mu^2)}}\begin{pmatrix} 1 \\ -2/\mu \\ 1 \end{pmatrix}.$$

The physical motions associated with these normal modes are illustrated in figure 9.3. The first, with $\lambda = \omega = 0$ and all the x_i equal, merely describes bodily translation of the whole system, with no (i.e. zero-frequency) internal oscillations.

In the second solution the central particle remains stationary, $x_2 = 0$, whilst the other two oscillate with equal amplitudes in antiphase with each other. This motion, which has frequency $\omega = (k/m)^{1/2}$, is illustrated in figure 9.3(b).

The final and most complicated of the three normal modes has angular frequency $\omega = \{[(\mu + 2)/\mu](k/m)\}^{1/2}$, and involves a motion of the central particle which is in antiphase with that of the two outer ones and which has an amplitude $2/\mu$ times as great. In this motion (see figure 9.3(c)) the two springs are compressed and extended in turn. We also note that in the second and third normal modes the centre of mass of the molecule remains stationary. ◄

9.2 Symmetry and normal modes

It will have been noticed that the system in the above example has an obvious symmetry under the interchange of coordinates 1 and 3: the matrices A and B, the equations of motion and the normal modes illustrated in figure 9.3 are all unaltered by the interchange of x_1 and $-x_3$. This reflects the more general result that for each physical symmetry possessed by a system, there is at least one normal mode with the same symmetry.

The general question of the relationship between the symmetries possessed by a physical system and those of its normal modes will be taken up more formally in chapter 29 where the representation theory of groups is considered. However, we can show here how an appreciation of a system's symmetry properties will sometimes allow its normal modes to be guessed (and then verified), something that is particularly helpful if the number of coordinates involved is greater than two and the corresponding eigenvalue equation (9.10) is a cubic or higher-degree polynomial equation.

Consider the problem of determining the normal modes of a system consisting of four equal masses M at the corners of a square of side $2L$, each pair of masses being connected by a light spring of modulus k that is unstretched in the equilibrium situation. As shown in figure 9.4, we introduce Cartesian coordinates x_n, y_n, with $n = 1, 2, 3, 4$, for the positions of the masses and denote their displacements from their equilibrium positions $\mathbf{R}_n$ by $\mathbf{q}_n = x_n\mathbf{i} + y_n\mathbf{j}$. Thus

$$\mathbf{r}_n = \mathbf{R}_n + \mathbf{q}_n \quad \text{with} \quad \mathbf{R}_n = \pm L\mathbf{i} \pm L\mathbf{j}.$$

The coordinates for the system are thus $x_1, y_1, x_2, \ldots, y_4$ and the kinetic energy matrix A is given trivially by $M\mathsf{I}_8$, where I_8 is the 8×8 identity matrix.

The potential energy matrix B is much more difficult to calculate and involves, for each pair of values m, n, evaluating the quadratic approximation to the expression

$$b_{mn} = \tfrac{1}{2}k\left(|\mathbf{r}_m - \mathbf{r}_n| - |\mathbf{R}_m - \mathbf{R}_n|\right)^2.$$

Expressing each $\mathbf{r}_i$ in terms of $\mathbf{q}_i$ and $\mathbf{R}_i$ and making the normal assumption that

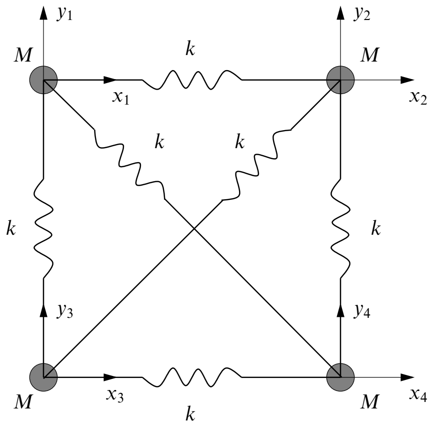


Figure 9.4 The arrangement of four equal masses and six equal springs discussed in the text. The coordinate systems x_n, y_n for $n = 1, 2, 3, 4$ measure the displacements of the masses from their equilibrium positions.

$|\mathbf{R}_m - \mathbf{R}_n| \gg |\mathbf{q}_m - \mathbf{q}_n|$, we obtain b_{mn} $(= b_{nm})$:

$$\begin{aligned}
b_{mn} &= \tfrac{1}{2}k\left[|(\mathbf{R}_m - \mathbf{R}_n) + (\mathbf{q}_m - \mathbf{q}_n)| - |\mathbf{R}_m - \mathbf{R}_n|\right]^2 \\
&= \tfrac{1}{2}k\left\{\left[|\mathbf{R}_m - \mathbf{R}_n|^2 + 2(\mathbf{q}_m - \mathbf{q}_n)\cdot(\mathbf{R}_M - \mathbf{R}_n) + |\mathbf{q}_m - \mathbf{q}_n)|^2\right]^{1/2} - |\mathbf{R}_m - \mathbf{R}_n|\right\}^2 \\
&= \tfrac{1}{2}k|\mathbf{R}_m - \mathbf{R}_n|^2\left\{\left[1 + \frac{2(\mathbf{q}_m - \mathbf{q}_n)\cdot(\mathbf{R}_M - \mathbf{R}_n)}{|\mathbf{R}_m - \mathbf{R}_n|^2} + \cdots\right]^{1/2} - 1\right\}^2 \\
&\approx \tfrac{1}{2}k\left\{\frac{(\mathbf{q}_m - \mathbf{q}_n)\cdot(\mathbf{R}_M - \mathbf{R}_n)}{|\mathbf{R}_m - \mathbf{R}_n|}\right\}^2.
\end{aligned}$$

This final expression is readily interpretable as the potential energy stored in the spring when it is extended by an amount equal to the component, along the equilibrium direction of the spring, of the relative displacement of its two ends.

Applying this result to each spring in turn gives the following expressions for the elements of the potential matrix.

m	n	$2b_{mn}/k$
1	2	$(x_1 - x_2)^2$
1	3	$(y_1 - y_3)^2$
1	4	$\frac{1}{2}(-x_1 + x_4 + y_1 - y_4)^2$
2	3	$\frac{1}{2}(x_2 - x_3 + y_2 - y_3)^2$
2	4	$(y_2 - y_4)^2$
3	4	$(x_3 - x_4)^2$.

The potential matrix is thus constructed as

$$\mathsf{B} = \frac{k}{4}\begin{pmatrix} 3 & -1 & -2 & 0 & 0 & 0 & -1 & 1 \\ -1 & 3 & 0 & 0 & 0 & -2 & 1 & -1 \\ -2 & 0 & 3 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 3 & -1 & -1 & 0 & -2 \\ 0 & 0 & -1 & -1 & 3 & 1 & -2 & 0 \\ 0 & -2 & -1 & -1 & 1 & 3 & 0 & 0 \\ -1 & 1 & 0 & 0 & -2 & 0 & 3 & -1 \\ 1 & -1 & 0 & -2 & 0 & 0 & -1 & 3 \end{pmatrix}.$$

To solve the eigenvalue equation $|\mathsf{B} - \lambda\mathsf{A}| = 0$ directly would mean solving an eigth-degree polynomial equation. Fortunately, we can exploit intuition and the symmetries of the system to obtain the eigenvectors and corresponding eigenvalues without such labour.

Firstly, we know that bodily translation of the whole system, without any internal vibration, must be possible and that there will be two independent solutions of this form, corresponding to translations in the x- and y- directions. The eigenvector for the first of these (written in row form to save space) is

$$\mathsf{x}^{(1)} = (1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0)^{\mathrm{T}}.$$

Evaluation of $\mathsf{Bx}^{(1)}$ gives

$$\mathsf{Bx}^{(1)} = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)^{\mathrm{T}},$$

showing that $\mathsf{x}^{(1)}$ is a solution of $(\mathsf{B} - \omega^2\mathsf{A})\mathsf{x} = \mathsf{0}$ corresponding to the eigenvalue $\omega^2 = 0$, whatever form Ax may take. Similarly,

$$\mathsf{x}^{(2)} = (0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1)^{\mathrm{T}}$$

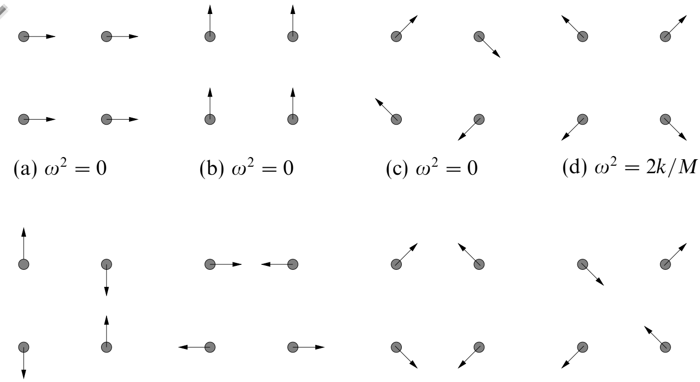
is a second eigenvector corresponding to the eigenvalue $\omega^2 = 0$.

The next intuitive solution, again involving no internal vibrations, and, therefore, expected to correspond to $\omega^2 = 0$, is pure rotation of the whole system about its centre. In this mode each mass moves perpendicularly to the line joining its position to the centre, and so the relevant eigenvector is

$$\mathsf{x}^{(3)} = \frac{1}{\sqrt{2}}(1 \quad 1 \quad 1 \quad -1 \quad -1 \quad 1 \quad -1 \quad -1)^{\mathrm{T}}.$$

It is easily verified that $\mathsf{Bx}^{(3)} = \mathsf{0}$ thus confirming both the eigenvector and the corresponding eigenvalue. The three non-oscillatory normal modes are illustrated in diagrams (a)–(c) of figure 9.5.

We now come to solutions that do involve real internal oscillations, and, because of the four-fold symmetry of the system, we expect one of them to be a mode in which all the masses move along radial lines – the so-called ‘breathing



(a) $\omega^2 = 0$ (b) $\omega^2 = 0$ (c) $\omega^2 = 0$ (d) $\omega^2 = 2k/M$

(e) $\omega^2 = k/M$ (f) $\omega^2 = k/M$ (g) $\omega^2 = k/M$ (h) $\omega^2 = k/M$

Figure 9.5 The displacements and frequencies of the eight normal modes of the system shown in figure 9.4. Modes (a), (b) and (c) are not true oscillations: (a) and (b) are purely translational whilst (c) is a mode of bodily rotation. Mode (d), the 'breathing mode', has the highest frequency and the remaining four, (e)–(h), of lower frequency, are degenerate.

mode'. Expressing this motion in coordinate form gives as the fourth eigenvector

$$\mathsf{x}^{(4)} = \frac{1}{\sqrt{2}}(-1 \quad 1 \quad 1 \quad 1 \quad -1 \quad -1 \quad 1 \quad -1)^{\mathrm{T}}.$$

Evaluation of $\mathsf{Bx}^{(4)}$ yields

$$\mathsf{Bx}^{(4)} = \frac{k}{4\sqrt{2}}(-8 \quad 8 \quad 8 \quad 8 \quad -8 \quad -8 \quad 8 \quad -8)^{\mathrm{T}} = 2k\mathsf{x}^{(4)},$$

i.e. a multiple of $\mathsf{x}^{(4)}$, confirming that it is indeed an eigenvector. Further, since $\mathsf{Ax}^{(4)} = M\mathsf{x}^{(4)}$, it follows from $(\mathsf{B} - \omega^2\mathsf{A})\mathsf{x} = \mathsf{0}$ that $\omega^2 = 2k/M$ for this normal mode. Diagram (d) of the figure illustrates the corresponding motions of the four masses.

As the next step in exploiting the symmetry properties of the system we note that, because of its reflection symmetry in the x-axis, the system is invariant under the double interchange of y_1 with $-y_3$ and y_2 with $-y_4$. This leads us to try an eigenvector of the form

$$\mathsf{x}^{(5)} = (0 \quad \alpha \quad 0 \quad \beta \quad 0 \quad -\alpha \quad 0 \quad -\beta)^{\mathrm{T}}.$$

Substituting this trial vector into $(\mathsf{B} - \omega^2\mathsf{A})\mathsf{x} = \mathsf{0}$ gives, of course, eight simultaneous equations for α and β, but they are all equivalent to just two, namely

$$\alpha + \beta = 0,$$
$$5\alpha + \beta = \frac{4M\omega^2}{k}\alpha;$$

these have the solution $\alpha = -\beta$ and $\omega^2 = k/M$. The latter thus gives the frequency of the mode with eigenvector

$$\mathsf{x}^{(5)} = (0 \quad 1 \quad 0 \quad -1 \quad 0 \quad -1 \quad 0 \quad 1)^{\mathrm{T}}.$$

Note that, in this mode, when the spring joining masses 1 and 3 is most stretched, the one joining masses 2 and 4 is at its most compressed. Similarly, based on reflection symmetry in the y-axis,

$$\mathsf{x}^{(6)} = (1 \quad 0 \quad -1 \quad 0 \quad -1 \quad 0 \quad 1 \quad 0)^{\mathrm{T}}$$

can be shown to be an eigenvector corresponding to the same frequency. These two modes are shown in diagrams (e) and (f) of figure 9.5.

This accounts for six of the expected eight modes, and the other two could be found by considering motions that are symmetric about both diagonals of the square or are invariant under successive reflections in the x- and y- axes. However, since A is a multiple of the unit matrix, and since we know that $(\mathsf{x}^{(j)})^{\mathrm{T}}\mathsf{A}\mathsf{x}^{(i)} = 0$ if $i \neq j$, we can find the two remaining eigenvectors more easily by requiring them to be orthogonal to each of those found so far.

Let us take the next (seventh) eigenvector, $\mathsf{x}^{(7)}$, to be given by

$$\mathsf{x}^{(7)} = (a \quad b \quad c \quad d \quad e \quad f \quad g \quad h)^{\mathrm{T}}.$$

Then orthogonality with each of the $\mathsf{x}^{(n)}$ for $n = 1, 2, \ldots, 6$ yields six equations satisfied by the unknowns $a, b, \ldots, h$. As the reader may verify, they can be reduced to the six simple equations

$$a + g = 0, \quad d + f = 0, \quad a + f = d + g,$$
$$b + h = 0, \quad c + e = 0, \quad b + c = e + h.$$

With six homogeneous equations for eight unknowns, effectively separated into two groups of four, we may pick one in each group arbitrarily. Taking $a = b = 1$ gives $d = e = 1$ and $c = f = g = h = -1$ as a solution. Substitution of

$$\mathsf{x}^{(7)} = (1 \quad 1 \quad -1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1)^{\mathrm{T}}.$$

into the eigenvalue equation checks that it is an eigenvector and shows that the corresponding eigenfrequency is given by $\omega^2 = k/M$.

We now have the eigenvectors for seven of the eight normal modes and the eighth can be found by making it simultaneously orthogonal to each of the other seven. It is left to the reader to show (or verify) that the final solution is

$$\mathsf{x}^{(8)} = (1 \quad -1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1 \quad 1)^{\mathrm{T}}$$

and that this mode has the same frequency as three of the other modes. The general topic of the degeneracy of normal modes is discussed in chapter 29. The movements associated with the final two modes are shown in diagrams (g) and (h) of figure 9.5; this figure summarises all eight normal modes and frequencies.

Although this example has been lengthy to write out, we have seen that the actual calculations are quite simple and provide the full solution to what is formally a matrix eigenvalue equation involving 8 × 8 matrices. It should be noted that our exploitation of the intrinsic symmetries of the system played a crucial part in finding the correct eigenvectors for the various normal modes.

9.3 Rayleigh–Ritz method

We conclude this chapter with a discussion of the Rayleigh–Ritz method for estimating the eigenfrequencies of an oscillating system. We recall from the introduction to the chapter that for a system undergoing small oscillations the potential and kinetic energy are given by

$$V = \mathsf{q}^{\mathrm{T}}\mathsf{B}\mathsf{q} \qquad \text{and} \qquad T = \dot{\mathsf{q}}^{\mathrm{T}}\mathsf{A}\dot{\mathsf{q}},$$

where the components of q are the coordinates chosen to represent the configuration of the system and A and B are symmetric matrices (or may be chosen to be such). We also recall from (9.9) that the normal modes x^i and the eigenfrequencies ω_i are given by

$$(\mathsf{B} - \omega_i^2\mathsf{A})\mathsf{x}^i = \mathsf{0}. \tag{9.14}$$

It may be shown that the eigenvectors x^i corresponding to different normal modes are linearly independent and so form a complete set. Thus, any coordinate vector q can be written $\mathsf{q} = \sum_j c_j\mathsf{x}^j$. We now consider the value of the generalised quadratic form

$$\lambda(\mathsf{x}) = \frac{\mathsf{x}^{\mathrm{T}}\mathsf{B}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}} = \frac{\sum_m (\mathsf{x}^m)^{\mathrm{T}} c_m^* \mathsf{B} \sum_i c_i \mathsf{x}^i}{\sum_j (\mathsf{x}^j)^{\mathrm{T}} c_j^* \mathsf{A} \sum_k c_k \mathsf{x}^k},$$

which, since both numerator and denominator are positive definite, is itself non-negative. Equation (9.14) can be used to replace $\mathsf{B}\mathsf{x}^i$, with the result that

$$\begin{aligned}\lambda(\mathsf{x}) &= \frac{\sum_m (\mathsf{x}^m)^{\mathrm{T}} c_m^* \mathsf{A} \sum_i \omega_i^2 c_i \mathsf{x}^i}{\sum_j (\mathsf{x}^j)^{\mathrm{T}} c_j^* \mathsf{A} \sum_k c_k \mathsf{x}^k} \\ &= \frac{\sum_m (\mathsf{x}^m)^{\mathrm{T}} c_m^* \sum_i \omega_i^2 c_i \mathsf{A}\mathsf{x}^i}{\sum_j (\mathsf{x}^j)^{\mathrm{T}} c_j^* \mathsf{A} \sum_k c_k \mathsf{x}^k}.\end{aligned} \tag{9.15}$$

Now the eigenvectors x^i obtained by solving $(\mathsf{B} - \omega^2\mathsf{A})\mathsf{x} = \mathsf{0}$ are not mutually orthogonal unless either A or B is a multiple of the unit matrix. However, it may be shown that they do possess the desirable properties

$$(\mathsf{x}^j)^{\mathrm{T}}\mathsf{A}\mathsf{x}^i = 0 \quad \text{and} \quad (\mathsf{x}^j)^{\mathrm{T}}\mathsf{B}\mathsf{x}^i = 0 \quad \text{if } i \neq j. \tag{9.16}$$

This result is proved as follows. From (9.14) it is clear that, for general i and j,

$$(\mathsf{x}^j)^{\mathrm{T}}(\mathsf{B} - \omega_i^2\mathsf{A})\mathsf{x}^i = 0. \tag{9.17}$$

But, by taking the transpose of (9.14) with i replaced by j and recalling that A and B are real and symmetric, we obtain

$$(\mathsf{x}^j)^{\mathrm{T}}(\mathsf{B} - \omega_j^2\mathsf{A}) = 0.$$

Forming the scalar product of this with x^i and subtracting the result from (9.17) gives

$$(\omega_j^2 - \omega_i^2)(\mathsf{x}^j)^{\mathrm{T}}\mathsf{A}\mathsf{x}^i = 0.$$

Thus, for $i \neq j$ and non-degenerate eigenvalues ω_i^2 and ω_j^2, we have that $(\mathsf{x}^j)^{\mathrm{T}}\mathsf{A}\mathsf{x}^i = 0$, and substituting this into (9.17) immediately establishes the corresponding result for $(\mathsf{x}^j)^{\mathrm{T}}\mathsf{B}\mathsf{x}^i$. Clearly, if either A or B is a multiple of the unit matrix then the eigenvectors are mutually orthogonal in the normal sense. The orthogonality relations (9.16) are derived again, and extended, in exercise 9.6.

Using the first of the relationships (9.16) to simplify (9.15), we find that

$$\lambda(\mathsf{x}) = \frac{\sum_i |c_i|^2\omega_i^2(\mathsf{x}^i)^{\mathrm{T}}\mathsf{A}\mathsf{x}^i}{\sum_k |c_k|^2(\mathsf{x}^k)^{\mathrm{T}}\mathsf{A}\mathsf{x}^k}. \tag{9.18}$$

Now, if ω_0^2 is the lowest eigenfrequency then $\omega_i^2 \geq \omega_0^2$ for all i and, further, since $(\mathsf{x}^i)^{\mathrm{T}}\mathsf{A}\mathsf{x}^i \geq 0$ for all i the numerator of (9.18) is $\geq \omega_0^2 \sum_i |c_i|^2(\mathsf{x}^i)^{\mathrm{T}}\mathsf{A}\mathsf{x}^i$. Hence

$$\lambda(\mathsf{x}) \equiv \frac{\mathsf{x}^{\mathrm{T}}\mathsf{B}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}} \geq \omega_0^2, \tag{9.19}$$

for any x whatsoever (whether x is an eigenvector or not). Thus we are able to estimate the lowest eigenfrequency of the system by evaluating λ for a variety of vectors x, the components of which, it will be recalled, give the ratios of the coordinate amplitudes. This is sometimes a useful approach if many coordinates are involved and direct solution for the eigenvalues is not possible.

An additional result is that the maximum eigenfrequency ω_{m}^2 may also be estimated. It is obvious that if we replace the statement '$\omega_i^2 \geq \omega_0^2$ for all i' by '$\omega_i^2 \leq \omega_{\mathrm{m}}^2$ for all i', then $\lambda(\mathsf{x}) \leq \omega_{\mathrm{m}}^2$ for any x. Thus $\lambda(\mathsf{x})$ always lies between the lowest and highest eigenfrequencies of the system. Furthermore, $\lambda(\mathsf{x})$ has a *stationary* value, equal to ω_k^2, when x is the kth eigenvector (see subsection 8.17.1).

►*Estimate the eigenfrequencies of the oscillating rod of section 9.1.*

Firstly we recall that

$$\mathsf{A} = \frac{Ml^2}{12}\begin{pmatrix} 6 & 3 \\ 3 & 2 \end{pmatrix} \qquad \text{and} \qquad \mathsf{B} = \frac{Mlg}{12}\begin{pmatrix} 6 & 0 \\ 0 & 3 \end{pmatrix}.$$

Physical intuition suggests that the slower mode will have a configuration approximating that of a simple pendulum (figure 9.1), in which $\theta_1 = \theta_2$, and so we use this as a *trial vector*. Taking $\mathsf{x} = (\theta \quad \theta)^{\mathrm{T}}$,

$$\lambda(\mathsf{x}) = \frac{\mathsf{x}^{\mathrm{T}}\mathsf{B}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{A}\mathsf{x}} = \frac{3Mlg\theta^2/4}{7Ml^2\theta^2/6} = \frac{9g}{14l} = 0.643\frac{g}{l},$$

and we conclude from (9.19) that the lower (angular) frequency is $\leq (0.643g/l)^{1/2}$. We have already seen on p. 319 that the true answer is $(0.641g/l)^{1/2}$ and so we have come very close to it.

Next we turn to the higher frequency. Here, a typical pattern of oscillation is not so obvious but, rather preempting the answer, we try $\theta_2 = -2\theta_1$; we then obtain $\lambda = 9g/l$ and so conclude that the higher eigenfrequency $\geq (9g/l)^{1/2}$. We have already seen that the exact answer is $(9.359g/l)^{1/2}$ and so again we have come close to it. ◄

A simplified version of the Rayleigh–Ritz method may be used to estimate the eigenvalues of a symmetric (or in general Hermitian) matrix B, the eigenvectors of which will be mutually orthogonal. By repeating the calculations leading to (9.18), A being replaced by the unit matrix I, it is easily verified that if

$$\lambda(\mathsf{x}) = \frac{\mathsf{x}^{\mathrm{T}}\mathsf{B}\mathsf{x}}{\mathsf{x}^{\mathrm{T}}\mathsf{x}}$$

is evaluated for *any* vector x then

$$\lambda_1 \leq \lambda(\mathsf{x}) \leq \lambda_{\mathrm{m}},$$

where $\lambda_1, \lambda_2 \ldots, \lambda_{\mathrm{m}}$ are the eigenvalues of B in order of increasing size. A similar result holds for Hermitian matrices.

9.4 Exercises

9.1 Three coupled pendulums swing perpendicularly to the horizontal line containing their points of suspension, and the following equations of motion are satisfied:

$$\begin{aligned} -m\ddot{x}_1 &= cmx_1 + d(x_1 - x_2), \\ -M\ddot{x}_2 &= cMx_2 + d(x_2 - x_1) + d(x_2 - x_3), \\ -m\ddot{x}_3 &= cmx_3 + d(x_3 - x_2), \end{aligned}$$

where x_1, x_2 and x_3 are measured from the equilibrium points; m, M and m are the masses of the pendulum bobs; and c and d are positive constants. Find the normal frequencies of the system and sketch the corresponding patterns of oscillation. What happens as $d \to 0$ or $d \to \infty$?

9.2 A double pendulum, smoothly pivoted at A, consists of two light rigid rods, AB and BC, each of length l, which are smoothly jointed at B and carry masses m and αm at B and C respectively. The pendulum makes small oscillations in one plane under gravity. At time t, AB and BC make angles $\theta(t)$ and $\phi(t)$, respectively, with the downward vertical. Find quadratic expressions for the kinetic and potential energies of the system and hence show that the normal modes have angular frequencies given by

$$\omega^2 = \frac{g}{l}\left[1+\alpha \pm \sqrt{\alpha(1+\alpha)}\right].$$

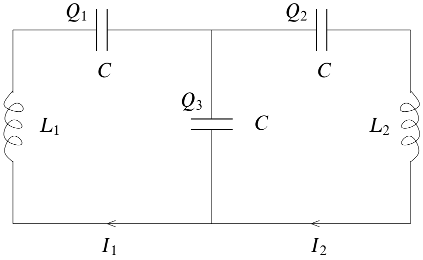
For $\alpha = 1/3$, show that in one of the normal modes the mid-point of BC does not move during the motion.

9.3 Continue the worked example, modelling a linear molecule, discussed at the end of section 9.1, for the case in which $\mu = 2$.

(a) Show that the eigenvectors derived there have the expected orthogonality properties with respect to both A and B.

(b) For the situation in which the atoms are released from rest with initial displacements $x_1 = 2\epsilon$, $x_2 = -\epsilon$ and $x_3 = 0$, determine their subsequent motions and maximum displacements.

9.4 Consider the circuit consisting of three equal capacitors and two different inductors shown in the figure. For charges Q_i on the capacitors and currents I_i through the components, write down Kirchhoff's law for the total voltage change around each of two complete circuit loops. Note that, to within an unimportant constant, the conservation of current implies that $Q_3 = Q_1 - Q_2$. Express the loop equations in the form given in (9.7), namely



$$\mathsf{A}\ddot{\mathsf{Q}} + \mathsf{B}\mathsf{Q} = \mathsf{0}.$$

Use this to show that the normal frequencies of the circuit are given by

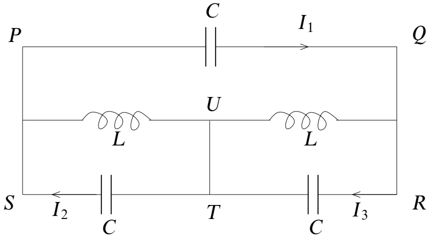
$$\omega^2 = \frac{1}{CL_1L_2}\left[L_1 + L_2 \pm (L_1^2 + L_2^2 - L_1L_2)^{1/2}\right].$$

Obtain the same matrices and result by finding the total energy stored in the various capacitors (typically $Q^2/(2C)$) and in the inductors (typically $LI^2/2$).

For the special case $L_1 = L_2 = L$ determine the relevant eigenvectors and so describe the patterns of current flow in the circuit.

9.5 It is shown in physics and engineering textbooks that circuits containing capacitors and inductors can be analysed by replacing a capacitor of capacitance C by a 'complex impedance' $1/(i\omega C)$ and an inductor of inductance L by an impedance $i\omega L$, where ω is the angular frequency of the currents flowing and $i^2 = -1$.

Use this approach and Kirchhoff's circuit laws to analyse the circuit shown in the figure and obtain three linear equations governing the currents I_1, I_2 and I_3. Show that the only possible frequencies of self-sustaining currents satisfy either



(a) $\omega^2 LC = 1$ or (b) $3\omega^2 LC = 1$. Find the corresponding current patterns and, in each case, by identifying parts of the circuit in which no current flows, draw an equivalent circuit that contains only one capacitor and one inductor.

9.6 *The simultaneous reduction to diagonal form of two real symmetric quadratic forms.*

Consider the two real symmetric quadratic forms $\mathsf{u}^{\mathrm{T}}\mathsf{A}\mathsf{u}$ and $\mathsf{u}^{\mathrm{T}}\mathsf{B}\mathsf{u}$, where u^{T} stands for the row matrix $(x \quad y \quad z)$, and denote by u^n those column matrices that satisfy

$$\mathsf{B}\mathsf{u}^n = \lambda_n \mathsf{A}\mathsf{u}^n, \tag{E9.1}$$

in which n is a label and the λ_n are real, non-zero and all different.

(a) By multiplying (E9.1) on the left by $(\mathsf{u}^m)^{\mathrm{T}}$, and the transpose of the corresponding equation for u^m on the right by u^n, show that $(\mathsf{u}^m)^{\mathrm{T}}\mathsf{A}\mathsf{u}^n = 0$ for $n \neq m$.
(b) By noting that $\mathsf{A}\mathsf{u}^n = (\lambda_n)^{-1}\mathsf{B}\mathsf{u}^n$, deduce that $(\mathsf{u}^m)^{\mathrm{T}}\mathsf{B}\mathsf{u}^n = 0$ for $m \neq n$.
(c) It can be shown that the u^n are linearly independent; the next step is to construct a matrix P whose columns are the vectors u^n.
(d) Make a change of variables $\mathsf{u} = \mathsf{P}\mathsf{v}$ such that $\mathsf{u}^{\mathrm{T}}\mathsf{A}\mathsf{u}$ becomes $\mathsf{v}^{\mathrm{T}}\mathsf{C}\mathsf{v}$, and $\mathsf{u}^{\mathrm{T}}\mathsf{B}\mathsf{u}$ becomes $\mathsf{v}^{\mathrm{T}}\mathsf{D}\mathsf{v}$. Show that C and D are diagonal by showing that $c_{ij} = 0$ if $i \neq j$, and similarly for d_{ij}.

Thus $\mathsf{u} = \mathsf{P}\mathsf{v}$ or $\mathsf{v} = \mathsf{P}^{-1}\mathsf{u}$ reduces both quadratics to diagonal form.

To summarise, the method is as follows:

(a) find the λ_n that allow (E9.1) a non-zero solution, by solving $|\mathsf{B} - \lambda\mathsf{A}| = 0$;
(b) for each λ_n construct u^n;
(c) construct the non-singular matrix P whose columns are the vectors u^n;
(d) make the change of variable $\mathsf{u} = \mathsf{P}\mathsf{v}$.

9.7 (*It is recommended that the reader does not attempt this question until exercise 9.6 has been studied.*)

If, in the pendulum system studied in section 9.1, the string is replaced by a second rod identical to the first then the expressions for the kinetic energy T and the potential energy V become (to second order in the θ_i)

$$T \approx Ml^2\left(\tfrac{8}{3}\dot{\theta}_1^2 + 2\dot{\theta}_1\dot{\theta}_2 + \tfrac{2}{3}\dot{\theta}_2^2\right),$$
$$V \approx Mgl\left(\tfrac{3}{2}\theta_1^2 + \tfrac{1}{2}\theta_2^2\right).$$

Determine the normal frequencies of the system and find new variables ξ and η that will reduce these two expressions to diagonal form, i.e. to

$$a_1\dot{\xi}^2 + a_2\dot{\eta}^2 \qquad \text{and} \qquad b_1\xi^2 + b_2\eta^2.$$

9.8 (*It is recommended that the reader does not attempt this question until exercise 9.6 has been studied.*)

Find a real linear transformation that simultaneously reduces the quadratic forms

$$3x^2 + 5y^2 + 5z^2 + 2yz + 6zx - 2xy,$$

$$5x^2 + 12y^2 + 8yz + 4zx$$

to diagonal form.

9.9 Three particles of mass m are attached to a light horizontal string having fixed ends, the string being thus divided into four equal portions each of length a and under a tension T. Show that for small transverse vibrations the amplitudes x^i of the normal modes satisfy $\mathsf{Bx} = (ma\omega^2/T)\mathsf{x}$, where B is the matrix

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

Estimate the lowest and highest eigenfrequencies using trial vectors $(3 \quad 4 \quad 3)^{\mathrm{T}}$ and $(3 \quad -4 \quad 3)^{\mathrm{T}}$. Use also the exact vectors $\left(1 \quad \sqrt{2} \quad 1\right)^{\mathrm{T}}$ and $\left(1 \quad -\sqrt{2} \quad 1\right)^{\mathrm{T}}$ and compare the results.

9.10 Use the Rayleigh–Ritz method to estimate the lowest oscillation frequency of a heavy chain of N links, each of length a $(= L/N)$, which hangs freely from one end. (Try simple calculable configurations such as all links but one vertical, or all links collinear, etc.)

9.5 Hints and answers

9.1 See figure 9.6.

9.3 (b) $x_1 = \epsilon(\cos\omega t + \cos\sqrt{2}\omega t)$, $x_2 = -\epsilon\cos\sqrt{2}\omega t$, $x_3 = \epsilon(-\cos\omega t + \cos\sqrt{2}\omega t)$. At various times the three displacements will reach $2\epsilon, \epsilon, 2\epsilon$ respectively. For example, x_1 can be written as $2\epsilon\cos[(\sqrt{2}-1)\omega t/2]\cos[(\sqrt{2}+1)\omega t/2]$, i.e. an oscillation of angular frequency $(\sqrt{2}+1)\omega/2$ and modulated amplitude $2\epsilon\cos[(\sqrt{2}-1)\omega t/2]$; the amplitude will reach 2ϵ after a time $\approx 4\pi/[\omega(\sqrt{2}-1)]$.

9.5 As the circuit loops contain no voltage sources, the equations are homogeneous, and so for a non-trivial solution the determinant of coefficients must vanish.
(a) $I_1 = 0$, $I_2 = -I_3$; no current in PQ; equivalent to two separate circuits of capacitance C and inductance L.
(b) $I_1 = -2I_2 = -2I_3$; no current in TU; capacitance $3C/2$ and inductance $2L$.

9.7 $\omega = (2.634g/l)^{1/2}$ or $(0.3661g/l)^{1/2}$; $\theta_1 = \xi + \eta$, $\theta_2 = 1.431\xi - 2.097\eta$.

9.9 Estimated, $10/17 < Ma\omega^2/T < 58/17$; exact, $2 - \sqrt{2} \le Ma\omega^2/T \le 2 + \sqrt{2}$.

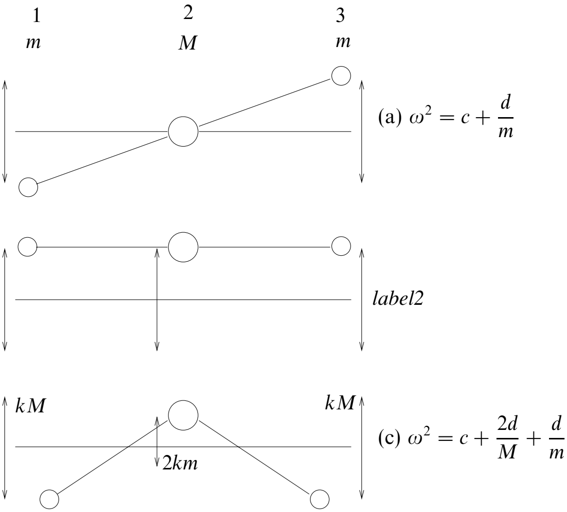

1
2
3
m
M
m
(a) $\omega^2 = c + \frac{d}{m}$
label2
kM
kM
$2km$
(c) $\omega^2 = c + \frac{2d}{M} + \frac{d}{m}$

Figure 9.6 The normal modes, as viewed from above, of the coupled pendulums in example 9.1.

10

Vector calculus

In chapter 7 we discussed the algebra of vectors, and in chapter 8 we considered how to transform one vector into another using a linear operator. In this chapter and the next we discuss the calculus of vectors, i.e. the differentiation and integration both of vectors describing particular bodies, such as the velocity of a particle, and of vector fields, in which a vector is defined as a function of the coordinates throughout some volume (one-, two- or three-dimensional). Since the aim of this chapter is to develop methods for handling multi-dimensional physical situations, we will assume throughout that the functions with which we have to deal have sufficiently amenable mathematical properties, in particular that they are continuous and differentiable.

10.1 Differentiation of vectors

Let us consider a vector **a** that is a function of a scalar variable u. By this we mean that with each value of u we associate a vector $\mathbf{a}(u)$. For example, in Cartesian coordinates $\mathbf{a}(u) = a_x(u)\mathbf{i} + a_y(u)\mathbf{j} + a_z(u)\mathbf{k}$, where $a_x(u)$, $a_y(u)$ and $a_z(u)$ are scalar functions of u and are the components of the vector $\mathbf{a}(u)$ in the x-, y- and z- directions respectively. We note that if $\mathbf{a}(u)$ is continuous at some point $u = u_0$ then this implies that each of the Cartesian components $a_x(u)$, $a_y(u)$ and $a_z(u)$ is also continuous there.

Let us consider the derivative of the vector function $\mathbf{a}(u)$ with respect to u. The derivative of a vector function is defined in a similar manner to the ordinary derivative of a scalar function $f(x)$ given in chapter 2. The small change in the vector $\mathbf{a}(u)$ resulting from a small change Δu in the value of u is given by $\Delta\mathbf{a} = \mathbf{a}(u + \Delta u) - \mathbf{a}(u)$ (see figure 10.1). The derivative of $\mathbf{a}(u)$ with respect to u is defined to be

$$\frac{d\mathbf{a}}{du} = \lim_{\Delta u \to 0} \frac{\mathbf{a}(u + \Delta u) - \mathbf{a}(u)}{\Delta u}, \tag{10.1}$$

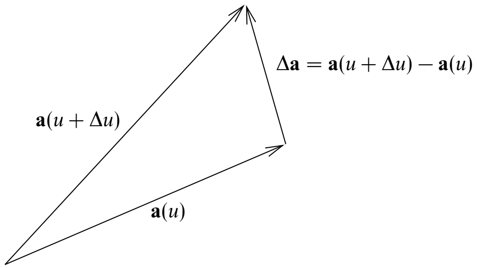

$\Delta\mathbf{a} = \mathbf{a}(u + \Delta u) - \mathbf{a}(u)$
$\mathbf{a}(u + \Delta u)$
$\mathbf{a}(u)$

Figure 10.1 A small change in a vector $\mathbf{a}(u)$ resulting from a small change in u.

assuming that the limit exists, in which case $\mathbf{a}(u)$ is said to be differentiable at that point. Note that $d\mathbf{a}/du$ is also a vector, which is not, in general, parallel to $\mathbf{a}(u)$. In Cartesian coordinates, the derivative of the vector $\mathbf{a}(u) = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}$ is given by

$$\frac{d\mathbf{a}}{du} = \frac{da_x}{du}\mathbf{i} + \frac{da_y}{du}\mathbf{j} + \frac{da_z}{du}\mathbf{k}.$$

Perhaps the simplest application of the above is to finding the velocity and acceleration of a particle in classical mechanics. If the time-dependent position vector of the particle with respect to the origin in Cartesian coordinates is given by $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$ then the velocity of the particle is given by the vector

$$\mathbf{v}(t) = \frac{d\mathbf{r}}{dt} = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k}.$$

The direction of the velocity vector is along the tangent to the path $\mathbf{r}(t)$ at the instantaneous position of the particle, and its magnitude $|\mathbf{v}(t)|$ is equal to the speed of the particle. The acceleration of the particle is given in a similar manner by

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt} = \frac{d^2x}{dt^2}\mathbf{i} + \frac{d^2y}{dt^2}\mathbf{j} + \frac{d^2z}{dt^2}\mathbf{k}.$$

►*The position vector of a particle at time t in Cartesian coordinates is given by $\mathbf{r}(t) = 2t^2\mathbf{i} + (3t - 2)\mathbf{j} + (3t^2 - 1)\mathbf{k}$. Find the speed of the particle at $t = 1$ and the component of its acceleration in the direction $\mathbf{s} = \mathbf{i} + 2\mathbf{j} + \mathbf{k}$.*

The velocity and acceleration of the particle are given by

$$\mathbf{v}(t) = \frac{d\mathbf{r}}{dt} = 4t\mathbf{i} + 3\mathbf{j} + 6t\mathbf{k},$$
$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt} = 4\mathbf{i} + 6\mathbf{k}.$$

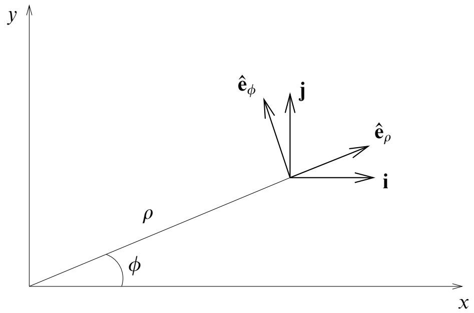


Figure 10.2 Unit basis vectors for two-dimensional Cartesian and plane polar coordinates.

The speed of the particle at $t = 1$ is simply

$$|\mathbf{v}(1)| = \sqrt{4^2 + 3^2 + 6^2} = \sqrt{61}.$$

The acceleration of the particle is constant (i.e. independent of t), and its component in the direction $\mathbf{s}$ is given by

$$\mathbf{a} \cdot \hat{\mathbf{s}} = \frac{(4\mathbf{i} + 6\mathbf{k}) \cdot (\mathbf{i} + 2\mathbf{j} + \mathbf{k})}{\sqrt{1^2 + 2^2 + 1^2}} = \frac{5\sqrt{6}}{3}. \blacktriangleleft$$

Note that in the case discussed above **i**, **j** and **k** are fixed, time-independent basis vectors. This may not be true of basis vectors in general; when we are not using Cartesian coordinates the basis vectors themselves must also be differentiated. We discuss basis vectors for non-Cartesian coordinate systems in detail in section 10.10. Nevertheless, as a simple example, let us now consider two-dimensional plane polar coordinates ρ, ϕ.

Referring to figure 10.2, imagine holding ϕ fixed and moving radially outwards, i.e. in the direction of increasing ρ. Let us denote the unit vector in this direction by $\hat{\mathbf{e}}_\rho$. Similarly, imagine keeping ρ fixed and moving around a circle of fixed radius in the direction of increasing ϕ. Let us denote the unit vector tangent to the circle by $\hat{\mathbf{e}}_\phi$. The two vectors $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_\phi$ are the basis vectors for this two-dimensional coordinate system, just as **i** and **j** are basis vectors for two-dimensional Cartesian coordinates. All these basis vectors are shown in figure 10.2.

An important difference between the two sets of basis vectors is that, while **i** and **j** are constant in magnitude *and direction*, the vectors $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_\phi$ have constant magnitudes but their directions change as ρ and ϕ vary. Therefore, when calculating the derivative of a vector written in polar coordinates we must also differentiate the basis vectors. One way of doing this is to express $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_\phi$

in terms of **i** and **j**. From figure 10.2, we see that

$$\hat{\mathbf{e}}_\rho = \cos\phi\,\mathbf{i} + \sin\phi\,\mathbf{j},$$
$$\hat{\mathbf{e}}_\phi = -\sin\phi\,\mathbf{i} + \cos\phi\,\mathbf{j}.$$

Since **i** and **j** are constant vectors, we find that the derivatives of the basis vectors $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_\phi$ with respect to t are given by

$$\frac{d\hat{\mathbf{e}}_\rho}{dt} = -\sin\phi\frac{d\phi}{dt}\,\mathbf{i} + \cos\phi\frac{d\phi}{dt}\,\mathbf{j} = \dot{\phi}\,\hat{\mathbf{e}}_\phi, \tag{10.2}$$
$$\frac{d\hat{\mathbf{e}}_\phi}{dt} = -\cos\phi\frac{d\phi}{dt}\,\mathbf{i} - \sin\phi\frac{d\phi}{dt}\,\mathbf{j} = -\dot{\phi}\,\hat{\mathbf{e}}_\rho, \tag{10.3}$$

where the overdot is the conventional notation for differentiation with respect to time.

►*The position vector of a particle in plane polar coordinates is* $\mathbf{r}(t) = \rho(t)\hat{\mathbf{e}}_\rho$. *Find expressions for the velocity and acceleration of the particle in these coordinates.*

Using result (10.4) below, the velocity of the particle is given by

$$\mathbf{v}(t) = \dot{\mathbf{r}}(t) = \dot{\rho}\,\hat{\mathbf{e}}_\rho + \rho\,\dot{\hat{\mathbf{e}}}_\rho = \dot{\rho}\,\hat{\mathbf{e}}_\rho + \rho\dot{\phi}\,\hat{\mathbf{e}}_\phi,$$

where we have used (10.2). In a similar way its acceleration is given by

$$\begin{aligned}\mathbf{a}(t) &= \frac{d}{dt}(\dot{\rho}\,\hat{\mathbf{e}}_\rho + \rho\dot{\phi}\,\hat{\mathbf{e}}_\phi)\\ &= \ddot{\rho}\,\hat{\mathbf{e}}_\rho + \dot{\rho}\,\dot{\hat{\mathbf{e}}}_\rho + \rho\dot{\phi}\,\dot{\hat{\mathbf{e}}}_\phi + \rho\ddot{\phi}\,\hat{\mathbf{e}}_\phi + \dot{\rho}\dot{\phi}\,\hat{\mathbf{e}}_\phi\\ &= \ddot{\rho}\,\hat{\mathbf{e}}_\rho + \dot{\rho}(\dot{\phi}\,\hat{\mathbf{e}}_\phi) + \rho\dot{\phi}(-\dot{\phi}\,\hat{\mathbf{e}}_\rho) + \rho\ddot{\phi}\,\hat{\mathbf{e}}_\phi + \dot{\rho}\dot{\phi}\,\hat{\mathbf{e}}_\phi\\ &= (\ddot{\rho} - \rho\dot{\phi}^2)\,\hat{\mathbf{e}}_\rho + (\rho\ddot{\phi} + 2\dot{\rho}\dot{\phi})\,\hat{\mathbf{e}}_\phi. \blacktriangleleft\end{aligned}$$

Here we have used (10.2) and (10.3).

10.1.1 Differentiation of composite vector expressions

In composite vector expressions each of the vectors or scalars involved may be a function of some scalar variable u, as we have seen. The derivatives of such expressions are easily found using the definition (10.1) and the rules of ordinary differential calculus. They may be summarised by the following, in which we assume that **a** and **b** are differentiable vector functions of a scalar u and that ϕ is a differentiable scalar function of u:

$$\frac{d}{du}(\phi\mathbf{a}) = \phi\frac{d\mathbf{a}}{du} + \frac{d\phi}{du}\mathbf{a}, \tag{10.4}$$
$$\frac{d}{du}(\mathbf{a}\cdot\mathbf{b}) = \mathbf{a}\cdot\frac{d\mathbf{b}}{du} + \frac{d\mathbf{a}}{du}\cdot\mathbf{b}, \tag{10.5}$$
$$\frac{d}{du}(\mathbf{a}\times\mathbf{b}) = \mathbf{a}\times\frac{d\mathbf{b}}{du} + \frac{d\mathbf{a}}{du}\times\mathbf{b}. \tag{10.6}$$