An Introduction to Numerical Analysis

Solutions to Exercises

Endre Süli and David F. Mayers

University of Oxford

CAMBRIDGE UNIVERSITY PRESS

Cambridge

New York *Port Chester* *Melbourne* *Sydney*

Solution to Exercise 1.1

The fixed points are the solutions of

$$q(x) \equiv x^2 - 2x + c = 0.$$

Evidently $q(0) = c > 0$, $q(1) = c - 1 < 0$, $q(x) \to +\infty$ as $x \to \infty$, showing that $0 < \xi_1 < 1 < \xi_2$.

Now

$$\begin{aligned} x_{n+1} &= \tfrac{1}{2}(x_n^2 + c) \\ \xi_1 &= \tfrac{1}{2}(\xi_1^2 + c), \end{aligned}$$

so by subtraction

$$x_{n+1} - \xi_1 = \tfrac{1}{2}(x_n^2 - \xi_1^2) = \tfrac{1}{2}(x_n + \xi_1)(x_n - \xi_1).$$

It follows that if $|x_n + \xi_1| < 2$, then $|x_{n+1} - \xi_1| < |x_n - \xi_1|$. Now $\xi_1 + \xi_2 = 2$, so if $0 \le x_0 < \xi_2$ then $x_0 + \xi_1 < 2$, and evidently $x_0 + \xi_1 > 0$. Hence x_1 is closer to ξ_1 then was x_0, so also $0 \le x_1 < \xi_2$. An induction argument then shows that each x_n satisfies $0 \le x_n < \xi_2$, and

$$|x_n - \xi_1| < \left(\frac{x_0 + \xi_1}{2}\right)^n |x_0 - \xi_1|,$$

and $x_n \to \xi_1$.

Now x_{n+1} is independent of the sign of x_n, and is therefore also independent of the sign of x_0, and it follows that $x_n \to \xi_1$ for all x_0 such that $-\xi_2 < x_0 < \xi_2$.

The same argument shows that if $x_0 > \xi_2$ then $x_1 > x_0 > \xi_2$, and so $x_n \to \infty$. As before this means also that $x_n \to \infty$ if $x_0 < -\xi_2$.

If $x_0 = \xi_2$ then of course $x_n = \xi_2$ for all $n > 0$. If $x_0 = -\xi_2$, then $x_1 = \xi_2$, and again $x_n = \xi_2$ for all $n > 0$.

Solution to Exercise 1.2

Since $f'(x) = \mathrm{e}^x - 1$ and $f''(x) = \mathrm{e}^x$, $f'(x) > 0$ and $f''(x) > 0$ for all $x > 0$. It therefore follows from Theorem 1.9 that if $x_0 > 0$ then Newton's method converges to the positive root.

Similarly $f'(x) < 0$ and $f''(x) > 0$ in $(-\infty, 0)$ and the same argument shows that the method converges to the negative root if $x_0 < 0$.

If $x_0 = 0$ the method fails, as $f'(0) = 0$, and x_1 does not exist.

For this function f, Newton's method gives

$$\begin{aligned} x_{n+1} &= x_n - \frac{\exp(x_n) - x_n - 2}{\exp(x_n) - 1} \\ &= x_n - \frac{1 - (x_n + 2)\exp(-x_n)}{1 - \exp(x_n)} \\ &\approx x_n - 1 \quad n > 1. \end{aligned}$$

In fact, e^{-100} is very small indeed.

In the same way, when x_0 is large and negative, say $x_0 = -100$,

$$x_{n+1} \approx x_n - \frac{-x_n - 2}{-1} = -2.$$

Hence when $x_0 = 100$, the first few members of the sequence are $100, 99, 98, \ldots$; after 98 iterations x_n will get close to the positive root, and convergence becomes quadratic and rapid. About 100 iterations are required to give an accurate value for the root.

However, when $x_0 = -100$, x_1 is very close to -2, and is therefore very close to the negative root. Three, or possibly four, iterations should give the value of the root to six decimal places.

Solution to Exercise 1.3

Newton's method is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

To avoid calculating the derivative we might consider approximating the derivative by

$$f'(x_n) \approx \frac{f(x_n + \delta) - f(x_n)}{\delta},$$

where δ is small. The given iteration uses this approximation, with $\delta = f(x_n)$; if x_n is close to a root then we might expect that $f(x_n)$ is small.

If $x_n - \xi$ is small we can write

$$\begin{aligned} f(x_n) &= f(\xi) + (x_n - \xi)f'(\xi) + \tfrac{1}{2}(x_n - \xi)^2 f''(\xi) + \mathcal{O}(x_n - \xi)^3 \\ &= \eta f' + \tfrac{1}{2}\eta^2 f'' + \mathcal{O}(\eta)^3 \end{aligned}$$

where $\eta = x_n - \xi$, and f' and f'' are evaluated at $x = \xi$. Then

$$\begin{aligned} f(x_n + f(x_n)) - f(x_n) &= f(\xi + \eta + \eta f' + \tfrac{1}{2}\eta^2 f'') - f(\xi + \eta) \\ &= \eta (f')^2 + \tfrac{1}{2}\eta^2 [3 f' f'' + (f')^2 f''] + \mathcal{O}(\eta)^3. \end{aligned}$$

Hence

$$\begin{aligned} x_{n+1} - \xi &= x_n - \xi - \frac{[f(x_n)]^2}{f(x + n + f(x_n)) - f(x_n)} \\ &= \eta - \frac{\eta^2 [(f')^2 + \eta f' f'']}{\eta [(f')^2 + \tfrac{1}{2}\eta f'(3 + f') f'']} + \mathcal{O}(\eta)^3 \\ &= \eta^2 \frac{f''(1 + f')}{f'} + \mathcal{O}(\eta)^3. \end{aligned}$$

This shows that if $x_0 - \eta$ is sufficiently small the iteration converges quadratically. The analysis here requires that f''' is continuous in a neighbourhood of ξ, to justify the terms $\mathcal{O}(\eta^3)$. A more careful analysis might relax this to require only the continuity of f''.

The leading term in $x_{n+1} - \xi$ is very similar to that in Newton's method, but with an additional term.

The convergence of this method, starting from a point close to a root, is very similar to Newton's method. But if x_0 is some way from the root $f(x_n)$ will not be small, the approximation to the derivative $f'(x_n)$ is very poor, and the behaviour may be very different. For the example

$$f(x) = \mathrm{e}^x - x - 2$$

starting from $x_0 = 1,\ 10$ and -10 we find

0	1.000000
1	1.205792
2	1.153859
3	1.146328
4	1.146193
5	1.146193

0	10.000000
1	10.000000
2	10.000000
3	10.000000
4	10.000000
5	10.000000

0	-10.000000
1	-1.862331
2	-1.841412
3	-1.841406
4	-1.841406

The convergence from $x_0 = 1$ is satisfactory. Starting from $x_0 = -10$ we get similar behaviour to Newton's method, an immediate step to x_1 quite close to -2, and then rapid convergence to the negative root.

However, starting from $x_0 = 10$ gives a quite different result. This time $f(x_0)$ is roughly 20000 (which is not small), and $f(x_0 + f(x_0))$ is about 10^{9500}; the difference between x_0 and x_1 is excessively small. Although the iteration converges, the rate of convergence is so slow that for any practical purpose it is virtually stationary. Even starting from $x_0 = 3$ many thousands of iterations are required for convergence.

Solution to Exercise 1.4

The number of correct figures in the approximation x_n to the root ξ is

$$D_n = \text{integer part of } \{-\log_{10}|\xi - x_n|\}.$$

From (1.24) for Newton's method we have

$$\frac{|\xi - x_{n+1}|}{|\xi - x_n|^2} \to \frac{|f''(\xi)|}{2|f'(\xi)|}.$$

Hence

$$D_{n+1} \approx 2\,D_n - B,$$

where

$$B = \log_{10}\frac{|f''(\xi)|}{2|f'(\xi)|}.$$

If B is small then D_{n+1} is close to $2\,D_n$, but if B is significantly larger than 1 then D_{n+1} may be smaller than this.

In the example,

$$\begin{aligned} f(x) &= \mathrm{e}^x - x - 1.0000000005 \\ f'(x) &= \mathrm{e}^x - 1 \\ f''(x) &= \mathrm{e}^x, \end{aligned}$$

and $\xi = 0.0001$. Hence

$$B = \log_{10}\frac{\mathrm{e}^{0.0001}}{2(\mathrm{e}^{0.0001} - 1)} = 3.7$$

and the number of significant figures in the next iteration is about $2k-4$, not $2k$.

Starting from $x_0 = 0.0005$ the results of Newton's method are

0	0.000500000000000	3
1	0.000260018333542	3
2	0.000149241714302	4
3	0.000108122910746	5
4	0.000100303597745	6
5	0.000099998797906	9
6	0.000099998333362	14

where the last column shows the number of correct decimal places.

The root is $\xi = 0.000099998333361$ to 15 decimal places.

The number of correct figures increases by a factor quite close to 4.

Solution to Exercise 1.5

From (1.23) we have

$$\xi - x_{n+1} = -\frac{(\xi - x_n)^2 f''(\eta_n)}{2f'(x_n)}.$$

Now $f'(\xi) = 0$, so by the Mean Value Theorem

$$f'(x_n) - f'(\xi) = (x_n - \xi) f''(\chi_n)$$

for some value of χ_n between ξ and x_n. Hence

$$\xi - x_{n+1} = \frac{(\xi - x_n) f''(\eta_n)}{2f''(\chi_n)}.$$

Now $|f''(\eta_n)| < M$ and $|f''(\chi_n)| > m$, and so

$$|\xi - x_{n+1}| < K|\xi - x_n|,$$

where

$$K = \frac{M}{2m} < 1.$$

Hence if x_0 lies in the given interval, all x_n lie in the interval, and $x_n \to \xi$.

Then $\eta_n \to \xi$, $f''(\eta_n) \to f''(\xi)$ and $f''(\chi_n) \to f''(\xi)$. This shows that

$$\frac{\xi - x_{n+1}}{\xi - x_n} \to \tfrac{1}{2}$$

and convergence is linear, with asymptotic rate of convergence $\ln 2$.

For the example $f(x) = \mathrm{e}^x - 1 - x$, $f(0) = 0$, $f'(0) = 0$. Starting from $x_0 = 1$, Newton's method gives

0	1.000
1	0.582
2	0.319
3	0.168
4	0.086
5	0.044
6	0.022
7	0.011
8	0.006
9	0.003
10	0.001

showing $\xi - x_0$ reducing by a factor close to $\frac{1}{2}$ at each step.

Solution to Exercise 1.6

When $f(\xi) = f'(\xi) = f''(\xi) = 0$ we get from the definition of Newton's method, provided that f''' is continuous in some neighbourhood of ξ,

$$\begin{aligned}\xi - x_{n+1} &= \xi - x_n + \frac{f(x_n)}{f'(x_n)} \\ &= \xi - x_n + \frac{\frac{1}{6}(x_n - \xi)^3 f'''(\eta_n)}{\frac{1}{2}(x_n - \xi)^2 f'''(\chi_n)} \\ &= (\xi - x_n)\left\{1 - \frac{f'''(\eta_n)}{3f'''(\chi_n)}\right\}.\end{aligned}$$

If we now assume that in the neighbourhood $[\xi - k, \xi + k]$ of the root

$$0 < m < |f'''(x)| < M, \quad \text{where } M < 3m,$$

then

$$|\xi - x_{n+1}| < K|\xi - x_n|,$$

where

$$K = 1 - \frac{M}{3m} < 1.$$

Hence if x_0 is in this neighbourhood, all the x_n lie in the neighbourhood, and Newton's method converges to ξ. Also,

$$\frac{|\xi - x_{n+1}|}{|\xi - x_n|} \to \frac{2}{3},$$

so that convergence is linear, with asymptotic rate of convergence $\ln(3/2)$.

Solution to Exercise 1.7

The proof follows closely the proof of Theorem 1.9.

From (1.23) it follows that $x_{n+1} < \xi$, provided that x_n lies in the interval $I = [X, \xi]$. Since f is monotonic increasing and $f(\xi) = 0$, $f(x) < 0$ in I. Hence if $x_0 \in I$ the sequence (x_n) lies in I, and is monotonic increasing. As it is bounded above by ξ, it converges; since ξ is the only root of $f(x) = 0$ in I, the sequence converges to ξ. Since f'' is continuous it follows that

$$\frac{\xi - x_{n+1}}{(\xi - x_n)^2} = -\frac{f''(\eta_n)}{2f'(x_n)} \to -\frac{f''(\xi)}{2f'(\xi)},$$

so that convergence is quadratic.

Solution to Exercise 1.8

Neglecting terms of second order in ε we get

$$\begin{aligned} x_0 &= 1 + \varepsilon \\ x_1 &= -1 + \varepsilon \\ x_2 &= \tfrac{1}{2}\varepsilon \\ x_3 &= -1 - \varepsilon \\ x_4 &= -1 + \varepsilon \\ x_5 &= -1. \end{aligned}$$

Although this value of x_5 is not exact, it is clear that for sufficiently small ε the sequence converges to -1.

With x_0 and x_1 interchanged, the value of x_2 is of course the same, but x_3 and subsequent values are different:

$$\begin{aligned} x_0 &= -1 + \varepsilon \\ x_1 &= 1 + \varepsilon \\ x_2 &= \tfrac{1}{2}\varepsilon \\ x_3 &= 1 - \varepsilon \\ x_4 &= 1 + \varepsilon \\ x_5 &= 1. \end{aligned}$$

Solution to Exercise 1.9

The function φ has the form

$$\varphi(x_n, x_{n-1}) = \frac{x_n f(x_{n-1}) - x_{n-1} f(x_n) - \xi(f(x_{n-1}) - f(x_n))}{(x_n - \xi)(x_{n-1} - \xi)(f(x_{n-1}) - f(x_n))}$$

In the limit as $x_n \to \xi$ both numerator and denominator tend to zero, so we apply l'Hopital's rule to give

$$\begin{aligned}\lim_{x_n \to \xi} \varphi(x_n, x_{n-1}) &= \lim \frac{f(x_{n-1}) - x_{n-1} f'(x_n) + \xi f'(x_n))}{-f'(x_n)(x_n - \xi)(x_{n-1} - \xi) + (f(x_{n-1}) - f(x_n))(x_{n-1} - \xi)} \\ &= \frac{f(x_{n-1}) - x_{n-1} f'(\xi) + \xi}{(f(x_{n-1}) - f(\xi))(x_{n-1} - \xi)}\end{aligned}$$

so that

$$\psi(x_{n-1}) = \frac{f(x_{n-1}) - x_{n-1} f'(\xi) + \xi f'(\xi)}{(f(x_{n-1}) - f(\xi))(x_{n-1} - \xi)}.$$

In the limit as $x_{n-1} \to \xi$ the numerator and denominator of $\psi(x_{n-1})$ both tend to zero, so again we use l'Hopital's rule to give

$$\lim_{x_{n-1} \to \xi} \psi(x_{n-1}) = \lim_{x_{n-1} \to \xi} \frac{f'(x_{n-1}) - f'(\xi)}{f'(x_{n-1})(x_{n-1} - \xi) + (f(x_{n-1}) - f(\xi))}.$$

We must now use l'Hopital's rule again, to give finally

$$\begin{aligned}\lim_{x_{n-1} \to \xi} \psi(x_{n-1}) &= \lim \frac{f''(x_{n-1})}{f''(x_{n-1})(x_{n-1} - \xi) + f'(x_{n-1}) + f'(x_{n-1})} \\ &= \frac{f''(\xi)}{2f'(\xi)}.\end{aligned}$$

Now the limit of φ does not depend on the way in which x_n and x_{n-1} tend to ξ, so finally we have

$$\frac{x_{n+1} - \xi}{(x_n - \xi)(x_{n-1} - \xi)} \to \frac{f''(\xi)}{2f'(\xi)}.$$

Now assume that

$$\frac{x_{n+1} - \xi}{(x_n - \xi)^q} \to A;$$

then

$$\frac{x_n - \xi}{(x_{n-1} - \xi)^q} \to A;$$

or

$$\frac{(x_n - \xi)^{1/q}}{x_{n-1} - \xi} \to A^{1/q},$$

and so

$$\frac{x_{n+1}-\xi}{(x_n-\xi)^{q-1/q}(x_{n-1}-\xi)} \to A^{1+1/q}.$$

Comparing with the previous limit, we require

$$q-1/q=1, \quad \text{and} \quad A^{1+1/q}=\frac{f''(\xi)}{2f'(\xi)}.$$

This gives a quadratic equation for q, and since we clearly require that θ is positive we obtain $q=\frac{1}{2}(1+\sqrt{5})$, giving the required result.

Solution to Exercise 1.10

Fig. 1.6 shows a typical situation with $f''(x) > 0$, so the graph of f lies below the line PQ. Here P and Q are the points corresponding to u_n and v_n. Also R is the point corresponding to θ, so that $f(\theta) < 0$. Hence in the next iteration $u_{n+1} = \theta$ and $v_{n+1} = v_n$.

The same picture applies to the next step, and again $v_{n+2} = v_{n+1}$, and so on. Thus if $f'' > 0$ in $[u_N, v_N]$, and $f(u_N) < 0 < f(v_N)$ then $v_n = v_N$ for all $n \geq N$.

If on the other hand $f(u_N) > 0$ and $f(v_N) < 0$ we see in the same way that $u_n = u_N$ for all $n \geq N$.

Similar results are easily deduced if $f'' < 0$ in $[u_N, v_N]$; it is only necessary to replace f by the function $-f$.

Now returning to the situation in Fig. 1.6, the point v_n remains fixed, and the points u_n are monotonically increasing. Hence the sequence (u_n) is monotonically increasing for $n \geq N$, and is bounded above by v_N, and is therefore convergent to the unique solution of $f(x) = 0$ in the interval $[u_N, v_N]$. In the general situation, we see that one end of the interval $[u_n, v_n]$ eventually remains fixed, and the other end converges to the root.

Write $u_n = \xi + \delta$, and

$$\frac{u_{n+1} - \xi}{\delta} = \frac{(\xi + \delta) f(v_N) - v_N f(\xi + \delta) - \xi(f(v_N) - f(\xi + \delta))}{\delta(f(v_N) - f(\xi + \delta))}.$$

In the limit as $\delta \to 0$ the numerator and denominator both tend to zero, so we apply l'Hopital's rule to give

$$\begin{aligned} \lim_{\delta \to 0} \frac{u_{n+1} - \xi}{\delta} &= \lim_{\delta \to 0} \frac{f(v_N) - v_N f'(\xi + \delta) + \xi f'(\xi + \delta)}{f(v_n) - f(\xi + \delta) - \delta f'(\xi + \delta)} \\ &= \frac{f(v_N) - v_N f'(\xi) + \xi f'(\xi)}{f(v_N)} \end{aligned}$$

Hence the sequence (u_n) converges linearly to ξ, and the asymptotic rate of convergence is

$$-\ln\left\{1 - \frac{(v_N - \xi) f'(\xi)}{f(v_N)}\right\}$$

This may also be written

$$-\ln\left\{1 - \frac{f'(\xi)}{f'(\eta_N)}\right\}$$

for some η_N lying between ξ and v_N. Since $f(\xi) = 0$, it follows that

$\eta_N > \xi$. Evidently the closer v_N is to the root ξ, the closer $f'(\eta_N)$ is to $f'(\xi)$, and the more rapidly the iteration converges.

Asymptotically this method converges more slowly than the standard secant method. Its advantage is that if $f(u_0)$ and $f(v_0)$ have opposite signs the iteration is guaranteed to converge to a root lying in $[u_0, v_0]$; the method is therefore robust. However, it is easy to draw a situation where v_0 is far from ξ, and where the bisection method is likely to be more efficient.

Solution to Exercise 1.11

The sequence (x_n) converges to the two-cycle a, b if $x_{2n} \to a$ and $x_{2n+1} \to b$, or equivalently with a and b interchanged. So a and b are fixed points of the composite iteration $x_{n+1} = h(x_n)$, where $h(x) = g(g(x))$, and we define a stable two-cycle to be one which corresponds to a stable fixed point of h. Now

$$h'(x) = g'(g(x))\, g'(x);$$

if $h'(a) < 1$ the fixed point a of h is stable; since $g(a) = a$ it follows that if $|g'(a)g'(b)| < 1$ then the two-cycle a, b is stable. In the same way, if $|g'(a)g'(b)| > 1$ then the two-cycle is not stable.

For Newton's method

$$x_{n+1} = x_n - f(x_m)/f'(x_n),$$

and the corresponding function g is defined by

$$g(x) = x - f(x)/f'(x).$$

In this case

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}.$$

Hence, if

$$\left|\frac{f(a)f''(a)}{[f'(a)]^2}\right| \left|\frac{f(b)f''(b)}{[f'(b)]^2}\right| < 1$$

the two-cycle is stable.

Newton's method for the solution of $x^3 - x = 0$ has the two-cycle $a, -a$ if

$$-a = a - \frac{a^3 - a}{3a^2 - 1}$$
$$a = -a - \frac{-a^3 + a}{3a^2 - 1}.$$

These equations have the solution

$$a = \frac{1}{\sqrt{5}}.$$

Here $f'(a) = 3a^2 - 1 = -2/5$ and $f''(a) = 6a = 6/\sqrt{5}$. So

$$\left|\frac{f(a)f''(a)}{[f'(a)]^2}\right| \left|\frac{f(b)f''(b)}{[f'(b)]^2}\right| = 36,$$

and the two-cycle is not stable.

Solution to Exercise 2.1

Multiplication by Q on the right reverses the order of the *columns* of A. Hence, writing $B = QAQ$,

$$B_{i,j} = A_{n+1-i,n+1-j}.$$

If L is lower triangular, then $L_{ij} = 0$ whenever $i < j$. Hence $(QLQ)_{ij} = 0$ whenever $n+1-i < n+1-j$, or when $i > j$, thus showing that QLQ is upper triangular.

Now suppose that A is a general $n \times n$ matrix, that $B = QAQ$, and that B can be written as $B = L_1U_1$, where L is unit lower triangular and U is upper triangular. Then

$$\begin{aligned} A &= QBQ \\ &= QL_1U_1Q \\ &= (QL_1Q)(QU_1Q) \end{aligned}$$

since $Q^2 = I$, the unit matrix. Now QL_1Q is unit upper triangular and QL_1Q is lower triangular, so we have the required form $A = UL$ with $L = QU_1Q$ and $U = QL_1Q$. This factorisation is possible if we can write $B = L_1U_1$, and this can be done if all the leading principal submatrices of B are nonsingular. The required condition is therefore that all the "trailing" principal submatrices of A are nonsingular.

The factorisation does not exist, for example, if

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 0 \end{pmatrix},$$

since the corresponding matrix B has $B_{11} = 0$.

Solution to Exercise 2.2

We can write

$$A = LU^*,$$

where L is unit lower triangular, and U^* is upper triangular. Now let D be the diagonal matrix whose diagonal elements are the diagonal elements of U_1, so that $d_{ii} = u^*_{ii}$, and define $U = D^{-1}U^*$; then $A = LDU$ as required, since U is unit upper triangular.

The given condition on A ensures that $u^*_{ii} \neq 0$ for $i = 1, \ldots, n$. The procedure needs to be modified slightly when $u^*_{nn} = 0$, since then D is singular. All that is necessary is to define $U = (D^*)^{-1}U^*$, where D^* is the matrix D but with the last element replaced by 1 and then to replace U_{nn} by 1.

If the factorisation $A = LU$ is known, we can use this procedure to find D and U such that $A = LDU$. Then $A^T = U^T DL^T$, which can be written

$$A^T = (U^T)(DL^T),$$

where U^T is unit lower triangular and DL^T is upper triangular. This is therefore the required factorisation of A^T.

Solution to Exercise 2.3

Suppose that the required result is true for $n = k$, so that any nonsingular $k \times k$ matrix A can be written as $PA = LU$. This is obviously true for $k = 1$.

Now consider any nonsingular $(n+1) \times (n+1)$ matrix A partitioned according to the first row and column. We locate the element in the first column with largest magnitude, or any one of them if there is more than one, and interchange rows if required. If the largest element is in row r we interchange rows 1 and r. We then write

$$P^{(1r)}A = \begin{pmatrix} \alpha & \boldsymbol{w}^T \\ \boldsymbol{p} & B \end{pmatrix} = \begin{pmatrix} \alpha & \mathbf{0}^T \\ \boldsymbol{p} & C \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{m}^T \\ \mathbf{0} & I \end{pmatrix}$$

where α is the largest element in the first column of A. Writing out the product we find that

$$\begin{aligned} \alpha \boldsymbol{m}^T &= \boldsymbol{w}^T \\ \boldsymbol{p}\boldsymbol{m}^T + C &= B. \end{aligned}$$

This gives

$$\boldsymbol{m} = \frac{1}{\alpha}\boldsymbol{w},$$

and

$$C = B - \frac{1}{\alpha}\boldsymbol{p}\boldsymbol{m}^T.$$

Note that if $\alpha = 0$ this implies that all the elements of the first column of A were zero, contradicting our assumption that A is nonsingular.

Now $\det(A) = \pm\alpha \det(C)$, and so the matrix C is also nonsingular, and as it is an $n \times n$ matrix we can use the inductive hypothesis to write

$$P^*C = L^*U^*.$$

It is then easy to see that

$$P^{(1r)}A = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & P^* \end{pmatrix} \begin{pmatrix} \alpha & \mathbf{0}^T \\ P^*\boldsymbol{p} & L^* \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{m}^T \\ \mathbf{0} & U^* \end{pmatrix}$$

since $P^*P^* = I$. Now defining the permutation matrix P by

$$P = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0}^T & P^* \end{pmatrix} P^{(1r)}$$

we obtain

$$PA = \begin{pmatrix} \alpha & \mathbf{0}^T \\ P^*\boldsymbol{p} & L^* \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{m}^T \\ \mathbf{0} & U^* \end{pmatrix},$$

which is the required factorisation of A.

The theorem therefore holds for every matrix of order $n+1$, and the induction is complete.

Consider the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

and attempt to write it in the form

$$A = \begin{pmatrix} p & 0 \\ q & r \end{pmatrix} \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}.$$

This would require

$$\begin{aligned} p &= 0 \\ ps &= 1 \\ q &= 0 \\ qs + r &= b. \end{aligned}$$

where the first two equations are clearly incompatible. The only possible permutation matrix P interchanges the rows, and the factorisation is obviously still impossible.

Solution to Exercise 2.4

Partitioning the matrices by the first k rows and columns, the equation $L\boldsymbol{y} = \boldsymbol{b}$ becomes

$$\begin{pmatrix} L_1 & O \\ C & L_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\beta} \end{pmatrix},$$

where we have used the fact that the first k rows of $\boldsymbol{b}$ are zero. Multiplying out this equation gives

$$\begin{aligned} L_1\boldsymbol{y}_1 &= \boldsymbol{0} \\ C\boldsymbol{y}_1 + L_2\boldsymbol{y}_2 &= \boldsymbol{\beta}, \end{aligned}$$

from which it is clear that $\boldsymbol{y}_1 = \boldsymbol{0}$, since L_1 and L_2 are nonsingular. Hence the first k rows of $\boldsymbol{y}$ are zero.

Column j of the inverse of L is $\boldsymbol{y}^{(j)}$, the solution of

$$L\boldsymbol{y}^{(j)} = \boldsymbol{e}^{(j)},$$

where $\boldsymbol{e}^{(j)}$ is column j of the unit matrix and has its only nonzero element in row j. Hence the first $j-1$ elements of $\boldsymbol{e}^{(j)}$ are zero, and by what we have just proved the first $j-1$ elements of $\boldsymbol{y}^{(j)}$ are zero. Thus in each column of the inverse all the elements above the diagonal element are zero, and the inverse is lower triangular.

Solution to Exercise 2.5

The operations on the matrix B are:

(i) Multiplying all the elements in a row by a scalar;
(ii) Adding a multiple of one row to another.

Each of these is equivalent to multiplying on the left by a nonsingular matrix. Evidently the effect of successive operations (i) is that the diagonal elements in the first n columns are each equal to 1, and the effect of successive operations (ii) is that all the offdiagonal elements are equal to zero. Hence the final result in the first n columns is the unit matrix.

Hence the combined effect of all these operations is equivalent to multiplying B on the left by A^{-1}; and the final result in the last columns is the inverse A^{-1}. The diagonal element b_{rr} at any stage is the determinant of the leading principal $r \times r$ submatrix of A, multiplied by each of the preceding scaling factors $1/b_{jj}$. Hence if all the leading principal submatrices of A are nonsingular, none of the diagonal elements b_{jj} used at any stage are zero.

At each stage, the scaling of the elements in row j requires $2n$ multiplications. The calculation of each term $b_{ik} - b_{ij}b_{jk}$ involves one multiplication, and there are $2n(n-1)$ such elements, as row j is not involved. Thus each stage requires $2n^2$ multiplications, and there are n stages, giving a total of $2n^3$ multiplications.

However, in stage j the first $j-1$ columns and the last $n-j+1$ columns are columns of the unit matrix. Hence the scaling of row j only involves n non zero elements, and in the calculating of the new elements b_{ik}, half of the factors b_{jk} are zero. This reduces the total number of multiplications from $2n^3$ to n^3.

Solution to Exercise 2.6

The initial matrix B is

$$B = \begin{pmatrix} 2\,4\,2\,1\,0\,0 \\ 1\,0\,3\,0\,1\,0 \\ 3\,1\,2\,0\,0\,1 \end{pmatrix}.$$

After the successive stages of the calculation the matrix B becomes

$$B = \begin{pmatrix} 1 & 2 & 1 & 1/2\,0\,0 \\ 0 & -2 & 2 & -1/2\,1\,0 \\ 0 & -5 & -1 & -3/2\,0\,1 \end{pmatrix},$$

$$B = \begin{pmatrix} 1\,0 & 3 & 0 & 1\,0 \\ 0\,1 & -1 & 1/4 & -1/2\,0 \\ 0\,0 & -6 & -1/4 & -5/2\,1 \end{pmatrix},$$

$$B = \begin{pmatrix} 1\,0\,0 & -1/8 & -1/4 & 1/2 \\ 0\,1\,0 & 7/24 & -1/12 & -1/6 \\ 0\,0\,1 & 1/24 & 5/12 & -1/6 \end{pmatrix}.$$

The inverse matrix A^{-1} consists of the last three columns of this.

Solution to Exercise 2.7

$$\begin{aligned}\sum_{i=1}^{n} |(A\boldsymbol{x})_i| &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \\ &\leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|\, |x_j| \\ &= \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \\ &\leq C \sum_{j=1}^{n} |x_j| \\ &= C \|\boldsymbol{x}\|_1.\end{aligned}$$

Now choose

$$C = \max_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}|;$$

then evidently

$$\sum_{i=1}^{n} |a_{ij}| \leq C.$$

Let k be the value of j for which this maximum is attained; define $\boldsymbol{x}$ to be the vector whose only nonzero element is a 1 in position k. Then

$$\begin{aligned}\sum_{i=1}^{n} |(A\boldsymbol{x})_i| &= \sum_{i=1}^{n} |a_{ik}| \\ &= C \\ &= C \|\boldsymbol{x}\|_1,\end{aligned}$$

so that $\|A\boldsymbol{x}\|_1 = C\|\boldsymbol{x}\|_1$, giving the result required.

Solution to Exercise 2.8

(i) Write $\alpha = \|\boldsymbol{v}\|_2$, so that

$$v_1^2 + \ldots + v_n^2 = \alpha^2.$$

It is then clear that $v_j^2 \leq \alpha^2$, $j = 1, \ldots, n$, and so $\|\boldsymbol{v}\|_\infty \leq \|\boldsymbol{v}\|_2$. Equality is attained by any vector $\boldsymbol{v}$ which has a single nonzero element.

Now write $\beta = \|\boldsymbol{v}\|_\infty$, so that $|v_j| \leq \beta$ for all j. Then

$$\|\boldsymbol{v}\|_2^2 = v_1^2 + \ldots + v_n^2 \leq \beta^2 + \ldots + \beta^2 = n\,\beta^2.$$

This is the required result; in this case equality is attained by any vector $\boldsymbol{v}$ in which all the elements have the same magnitude.

(ii) From the definition of $\|A\|_\infty$,

$$\|A\|_\infty = \max_{\boldsymbol{x}} \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

Choose $\boldsymbol{v}$ to be a vector $\boldsymbol{x}$ for which this maximum is attained. Then

$$\begin{aligned}
\|A\|_\infty \|\boldsymbol{v}\|_\infty &= \|A\boldsymbol{v}\|_\infty \\
&\leq \|A\boldsymbol{v}\|_2 \qquad \text{(see above)} \\
&\leq \|A\|_2 \|\boldsymbol{v}\|_2 \\
&\leq \|A\|_2 \sqrt{n}\, \|\boldsymbol{v}\|_\infty. \qquad \text{(see above)}
\end{aligned}$$

Division by $\|\boldsymbol{v}\|_\infty$ gives the required result.

To attain equality, we require equality throughout the argument. This means that $A\boldsymbol{v}$ must have a single nonzero element and that all the elements of $\boldsymbol{v}$ must have the same magnitude. Moreover $A\boldsymbol{v}$ must have its maximum possible size in both the 2-norm and the ∞-norm. Thus $\boldsymbol{v}$ must be an eigenvector of $A^T A$. An example is

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad \boldsymbol{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Evidently the rowsums are 2, so that $\|A\|_\infty = 2$. It is easy to see that $A^T A = 2I$, and has both eigenvalues equal to 2. Hence $\|A\|_2 = \sqrt{2}$, as required.

For the second inequality, choose $\boldsymbol{u}$ to be a vector which attains the maximum possible $\|A\boldsymbol{u}\|_2$. Then in the same way

$$\|A\|_2 \|\boldsymbol{u}\|_2 = \|A\boldsymbol{u}\|_2$$

$$\begin{aligned} &\leq \sqrt{m}\,\|A\boldsymbol{u}\|_\infty \\ &\leq \sqrt{m}\,\|A\|_\infty \|\boldsymbol{u}\|_\infty \\ &\leq \sqrt{m}\,\|A\|_\infty \|\boldsymbol{u}\|_2, \end{aligned}$$

and the result follows by division by $\|\boldsymbol{u}\|_2$. The argument follows as above, but we have to note that the vector $A\boldsymbol{u}$ has m elements.

To attain equality throughout we now require that $\boldsymbol{u}$ has a single nonzero element, that all the elements of $A\boldsymbol{u}$ have the same magnitude, and again that $A\boldsymbol{u}$ must have maximum size in both the 2-norm and the ∞-norm. Thus $\boldsymbol{u}$ must again be an eigenvector of A^TA. A rather trivial example has

$$A = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

which is a 2×1 matrix. Clearly $\|A\|_\infty = 1$, $A^TA = 2I$, so that $\|A\|_2 = \sqrt{2}$.

Solution to Exercise 2.9

We know that $\|A\|_2^2$ is the largest eigenvalue of $A^T A$, which is λ_n. In the same way $\|A^{-1}\|_2^2$ is the largest eigenvalue of

$$A^{-1^T} A^{-1} = (A\,A^T)^{-1},$$

and the eigenvalues of this matrix are the reciprocals of the eigenvalues of $A\,A^T$. Now the eigenvalues of $A\,A^T$ are the same as the eigenvalues of $A^T A$, so that $\|A^{-1}\|_2^2 = 1/\lambda_1$.

Now if Q is orthogonal, then $Q^T Q = I$, and all the eigenvalues of I are equal to 1. Hence by the result just proved, $\|Q\|_2 = 1$.

Conversely, if $\|A\|_2 = 1$, then the largest and smallest eigenvalues of $A^T A$ must be equal, so all the eigenvalues are equal. Hence $A^T A = \lambda I$, where λ is the eigenvalue. The matrix $A^T A$ is positive definite, so $\lambda > 0$, and writing

$$Q = \frac{1}{\sqrt{\lambda}} A$$

we have shown that $Q^T Q = I$, so that Q is orthogonal, as required.

Solution to Exercise 2.10

Let λ be an eigenvalue of A^TA and let $\boldsymbol{x} \neq \boldsymbol{0}$ be the associated eigenvector. Then, $A^TA\boldsymbol{x} = \lambda\boldsymbol{x}$, and therefore

$$\|A\boldsymbol{x}\|_2^2 = \boldsymbol{x}^TA^TA\boldsymbol{x} = \lambda\boldsymbol{x}^T\boldsymbol{x} = \lambda\|\boldsymbol{x}\|_2^2.$$

Hence, λ is a nonnegative real number.

Now let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$ and let $\|\cdot\|$ denote the associated subordinate matrix norm on $\mathbb{R}^{n\times n}$. Then,

$$\begin{aligned}|\lambda|\,\|\boldsymbol{x}\| = \|\lambda\boldsymbol{x}\| &= \|A^TA\boldsymbol{x}\| \\ &\leq \|A^TA\|\,\|\boldsymbol{x}\| \\ &\leq \|A^T\|\,\|A\|\,\|\boldsymbol{x}\|.\end{aligned}$$

Since $\boldsymbol{x} \neq \boldsymbol{0}$, it follows that

$$0 \leq \lambda \leq \|A^T\|\,\|A\|$$

for each eigenvalue λ of A^TA. By Theorem 2.9, we then have that

$$\|A\|_2 \leq \|A^T\|\,\|A\|$$

for any subordinate matrix norm $\|\cdot\|$. For example, with $\|\cdot\| = \|\cdot\|_\infty$, on noting that $\|A^T\|_\infty = \|A\|_1$, we conclude that

$$\|A\|_2^2 \leq \|A\|_1\,\|A\|_\infty,$$

as required.

Solution to Exercise 2.11

On multiplying A by A^T from the left, we deduce that

$$A^TA = \begin{bmatrix} n & 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Writing the eigenvalue problem $A^TA\boldsymbol{x} = \lambda\boldsymbol{x}$ in expanded form then gives

$$\begin{aligned} nx_1 + x_2 + x_3 + x_4 + \ldots + x_n &= \lambda x_1 \\ x_1 + x_2 &= \lambda x_2 \\ x_1 + x_3 &= \lambda x_3 \\ x_1 + x_4 &= \lambda x_4 \\ &\dots \\ x_1 + x_n &= \lambda x_n \end{aligned}$$

We observe that $\lambda = 1$ is an eigenvalue corresponding to the $(n-2)$ eigenvectors of A^TA of the form $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, with $x_1 = 0$, $x_2 + \ldots + x_n = 0$. The two remaining eigenvectors of A^TA are of the form $\boldsymbol{x} = (x_1, x_2, x_2, \ldots, x_2)$ where x_1 and x_2 are nonzero real numbers, and are found by solving the linear system

$$\begin{aligned} nx_1 + (n-1)x_2 &= \lambda x_1 \\ x_1 + x_2 &= \lambda x_2 \end{aligned}$$

which has a nontrivial solution when

$$\det \begin{pmatrix} n-\lambda & n-1 \\ 1 & 1-\lambda \end{pmatrix} = \lambda^2 - (n+1)\lambda + 1 = 0,$$

i.e., when

$$\lambda = \frac{1}{2}(n+1)\left[1 \pm \sqrt{1 - \tfrac{4}{(n+1)^2}}\right].$$

By Theorem 2.9, we then have that

$$\|A\|_2 = \frac{1}{2}(n+1)\left[1 + \sqrt{1 - \tfrac{4}{(n+1)^2}}\right].$$

Solution to Exercise 2.12

If $(I - B)$ is singular, there is a nonzero vector $\boldsymbol{x}$ such that

$$(I - B)\boldsymbol{x} = 0,$$

so that

$$\boldsymbol{x} = B\boldsymbol{x}.$$

Hence

$$\|\boldsymbol{x}\| = \|B\boldsymbol{x}\| \leq \|B\| \, \|\boldsymbol{x}\|,$$

and so $\|B\| \geq 1$. It then follows that if $\|A\| < 1$, then $I - A$ is not singular.

From the relation

$$(I - A)(I - A)^{-1} = I$$

it follows that

$$(I - A)^{-1} = I + A(I - A)^{-1}$$

and so

$$\begin{aligned} \|(I - A)^{-1}\| &\leq \|I\| + \|A(I - A)^{-1}\| \\ &\leq 1 + \|A\| \, \|(I - A^{-1}\|. \end{aligned}$$

Thus

$$(1 - \|A\|)\|(I - A)^{-1}\| \leq 1,$$

giving the required result.

Solution to Exercise 2.13

From $A\boldsymbol{x} = \boldsymbol{b}$ and $(A + \delta A)(\boldsymbol{x} + \boldsymbol{\delta x}) = \boldsymbol{b}$ it follows that

$$A\,\boldsymbol{\delta x} + \delta A\ \boldsymbol{x} + \delta A\ \boldsymbol{\delta x} = \boldsymbol{0}.$$

Then

$$(I + A^{-1}\delta A)\boldsymbol{\delta x} = -A^{-1}\delta A\ \boldsymbol{\delta x},$$

or

$$\boldsymbol{\delta x} = (I + A^{-1}\delta A)^{-1}A^{-1}\ \delta A\ \boldsymbol{\delta x}.$$

and so

$$\|\boldsymbol{\delta x}\| \leq \|(I + A^{-1}\delta A)^{-1}\|\ \|A^{-1}\delta A\|\,\|\boldsymbol{\delta x}\|.$$

Applying the result of Exercise 2.12 we get, provided that $\|A^{-1}\delta A\| < 1$,

$$\|\boldsymbol{\delta x}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|}\|A^{-1}\delta A\|\ \|blx\|,$$

which is the required result.

Solution to Exercise 2.14

Choose $\boldsymbol{x}$ to be an eigenvector of A^TA with eigenvalue λ, and $\boldsymbol{\delta x}$ to be an eigenvector with eigenvalue μ. Then

$$A\boldsymbol{x} = \boldsymbol{b},$$

so that

$$\boldsymbol{b}^T\boldsymbol{b} = \boldsymbol{x}^TA^TA\boldsymbol{x} = \boldsymbol{x}^T\lambda\boldsymbol{x} = \lambda\boldsymbol{x}^T\boldsymbol{x},$$

and

$$\|\boldsymbol{b}\|_2^2 = \lambda\|\boldsymbol{x}\|_2^2.$$

In the same way

$$\|\boldsymbol{\delta b}\|_2^2 = \mu\|\boldsymbol{\delta x}\|_2^2,$$

and so

$$\frac{\|\boldsymbol{\delta x}\|_2}{\|\boldsymbol{x}\|_2} = \left(\frac{\lambda}{\mu}\right)^{1/2}\frac{\|\boldsymbol{\delta b}\|_2}{\|\boldsymbol{b}\|_2}.$$

Now choose $\boldsymbol{x}$ so that $\lambda = \lambda_n$, the largest eigenvalue of A^TA, and $\boldsymbol{\delta x}$ so that $\mu = \lambda_1$, the smallest eigenvalue. Then

$$\frac{\lambda_n}{\lambda_1} = \kappa_2(A),$$

and equality is achieved as required, with $\boldsymbol{b} = A\boldsymbol{x}$ and $\boldsymbol{\delta b} = A\boldsymbol{\delta x}$.

Solution to Exercise 2.15

Following the notation of Theorem 2.12,

$$\begin{pmatrix} \boldsymbol{a} & A_n \end{pmatrix} = (\boldsymbol{q} \quad Q_n) \begin{pmatrix} \alpha & \boldsymbol{r}^T \\ \boldsymbol{0} & R_n \end{pmatrix},$$

we have

$$\boldsymbol{a} = \begin{pmatrix} 9 \\ 12 \\ 0 \end{pmatrix} \quad \text{and} \quad A_n = \begin{pmatrix} -6 \\ -8 \\ 20 \end{pmatrix}$$

and so

$$\alpha = \sqrt{(\boldsymbol{a}^T \boldsymbol{a})} = 15$$

and

$$\boldsymbol{q} = \begin{pmatrix} 3/5 \\ 4/5 \\ 0 \end{pmatrix}.$$

Then

$$\boldsymbol{r}^T = \boldsymbol{q}^T A_n = (3/5 \;\; 4/5 \;\; 0) \begin{pmatrix} -6 \\ -8 \\ 20 \end{pmatrix} = -10,$$

and

$$Q_n R_n = A_n - \boldsymbol{q}\, \boldsymbol{r}^T = \begin{pmatrix} 0 \\ 0 \\ 20 \end{pmatrix}.$$

Finally

$$Q_n = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

and

$$R_n = 20.$$

The required QR decomposition is therefore

$$\begin{pmatrix} 9 & -6 \\ 12 & -8 \\ 0 & 20 \end{pmatrix} = \begin{pmatrix} 3/5 & 0 \\ 4/5 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 15 & -10 \\ 0 & 20 \end{pmatrix}.$$

The required least squares solution is then given by solving

$$R_n \boldsymbol{x} = Q_n^T \boldsymbol{b},$$

or

$$\begin{pmatrix} 15 & -10 \\ 0 & 20 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3/5 & 4/5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 300 \\ 600 \\ 900 \end{pmatrix} = \begin{pmatrix} 660 \\ 900 \end{pmatrix}.$$

which gives

$$x_1 = 74, \quad x_2 = 45.$$

Solution to Exercise 3.1

Using the relations (3.2) and (3.3) we find in succession

$$\begin{aligned} l_{11} &= 2 \\ l_{21} &= 3 \\ l_{31} &= 1 \\ l_{22} &= 1 \\ l_{32} &= 0 \\ l_{33} &= 2. \end{aligned}$$

Hence $A = L\,L^T$ where

$$L = \begin{pmatrix} 2 & & \\ 3 & 1 & \\ 1 & 0 & 2 \end{pmatrix}.$$

Solution to Exercise 3.2

As in the solution to Exercise 1 we find that $A = L\,L^T$, where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix}.$$

Writing the system of equations $A\boldsymbol{x} = \boldsymbol{b}$, or $L\,L^T\boldsymbol{x} = \boldsymbol{b}$, in the form

$$\begin{aligned} L\,\boldsymbol{y} &= \boldsymbol{b}, \\ L^T\,\boldsymbol{x} &= \boldsymbol{y}, \end{aligned}$$

we find in succession

$$\begin{aligned} y_1 &= 4 \\ y_2 &= 1 \\ y_3 &= 1, \end{aligned}$$

and then

$$\begin{aligned} x_3 &= 2 \\ x_2 &= 0 \\ x_1 &= 1, \end{aligned}$$

which is the required solution.

Solution to Exercise 3.3

In the notation of equations (3.7) this example has

$$\begin{aligned} a_i &= -1, \qquad i = 2, \ldots, n \\ b_i &= 2, \qquad i = 1, \ldots, n \\ c_i &= -1, \qquad i = 1, \ldots, n-1. \end{aligned}$$

The elements l_j and u_j are determined uniquely by (3.7), with $u_1 = b_1 = 2$. Hence if we find expressions which satisfy all these equations then they are the values of l_j and u_j. Suppose now that $l_j = -(j-1)/j$. Then the first equation, $l_j = -1/u_{j-1}$, requires that

$$u_j = -1/l_{j+1} = (j+1)/j.$$

This also satisfies $u_1 = 2$, as required. The second of equations (3.7) is also satisfied, as

$$\begin{aligned} b_j - l_j c_{j-1} &= 2 + l_j \\ &= 2 - (j-1)/j \\ &= (j+1)/j \\ &= u_j. \end{aligned}$$

Hence all the equations are satisfied, and

$$\begin{aligned} l_j &= -(j-1)/j, \qquad j = 2, \ldots, n \\ u_j &= (j+1)/j, \qquad j = 1, \ldots, n-1. \end{aligned}$$

To find the determinant, note that $\det(T) = \det(L)\det(U)$, and $\det(L) = 1$ since L is unit lower triangular.

$$\begin{aligned} \det(U) &= u_1 u_2 \ldots u_n \\ &= \frac{2}{1}\frac{3}{2} \cdots \frac{n+1}{n} \\ &= n+1. \end{aligned}$$

Hence $\det(T) = n + 1$.

Solution to Exercise 3.4

Column j of the matrix T is the vector $c^{(j)}$, with the nonzero elements

$$c^{(j)}_{j-1} = -1, \quad c^{(j)}_j = 2, \quad c^{(j)}_{j+1} = -1,$$

with obvious modifications when $j = 1$ or $j = n$. Hence the scalar product of $c^{(j)}$ and $v^{(k)}$ is

$$M_{kj} = -v^{(k)}_{j-1} + 2v^{(k)}_j - v^{(k)}_{j+1}.$$

Since $v^{(k)}_i$ is a linear function of i, for $i = 1, \ldots, k$ and for $i = k, \ldots, n$, its is clear that $M_{kj} = 0$ when $j \neq k$. This is also true in the special cases $j = 1$ and $j = n$ since we can extend the definition of $v^{(k)}_j$ by the same linear functions, which gives $v^{(k)}_0 = 0$ and $v^{(k)}_n = 0$. The two linear functions give the same result when $j = k$. To find M_{kk} a simple calculation gives

$$\begin{aligned} M_{kk} &= -(k-1)(n+1-k) + 2k(n+1-k) - k(n+1-k-1) \\ &= n+1. \end{aligned}$$

Hence the scalar product of the vector $v^{(k)}/(n+1)$ with each column of T is column k of the unit matrix, so that $v^{(k)}/(n+1)$ is column k of the matrix T^{-1}. This shows that the elements of the inverse are

$$T^{-1}_{ik} = \begin{cases} \frac{i(n+1-k)}{n+1}, & i \leq k \\ \frac{k(n+1-i)}{n+1}, & i \geq k. \end{cases}$$

This matrix is clearly symmetric.

All the elements of the matrix are positive, and the sum of the elements in row i is

$$\begin{aligned} &\frac{n+1-i}{n+1} \sum_{k=1}^{i} k + \frac{i}{n+1} \sum_{k=i+1}^{n} (n+1-k) \\ &\quad = \frac{n+1-i}{n+1} \frac{i(i+1)}{2} + \frac{i}{n+1} \frac{(n-i)(n-i+1)}{2} \\ &\quad = \frac{i(n+1-i)}{2}. \end{aligned}$$

The $\infty-$norm of T^{-1} is the maximum rowsum; its value depends on whether n is odd or even. If n is odd, the maximum is attained when $i = (n+1)/2$, and is $(n+1)^2/8$. If n is even, the maximum is attained when $i = n/2$ and also when $i = n/2 + 1$; the maximum is $n(n+2)/8$.

Evidently the $\infty-$norm of T is 4, so finally the condition number is

$$\kappa_\infty(T) = \begin{cases} \frac{1}{2}(n+1)^2, & n \text{ odd} \\ \frac{1}{2}n(n+2), & n \text{ even.} \end{cases}$$

Solution to Exercise 3.5

With the notation of Theorem 3.4 we can now write

$$\begin{aligned} |u_j| &\geq \left| |b_j| - \|a_j| \left| \frac{c_{j-1}}{u_{j-1}} \right| \right| \\ &\geq \; | \, |b_j| - |a_j| \, | \\ &\geq \; |c_j| \\ &> \; 0. \end{aligned}$$

Note that the inequality '$> |c_j|$ ' now becomes '$\geq |c_j|$' but we can still deduce that $u_j \neq 0$ since c_j is not zero.

The matrix

$$T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

satisfies the given conditions, except that $c_2 = 0$. It is obviously singular, since the first two rows are identical. This also means that the leading 2×2 principal submatrix is singular, and so the LU factorisation does not exist.

Solution to Exercise 3.6

The proof is by induction on the order of the matrix; we suppose that the given result is true for every $n \times n$ tridiagonal matrix, and consider an $(n+1) \times (n+1)$ matrix T in the standard notation

$$T = \begin{pmatrix} b_1 & c_1 & & \\ a_2 & b_2 & c_2 & \\ & a_3 & b_3 & c_3 \\ & & \dots & \dots \end{pmatrix}.$$

We interchange the first two rows, if $|a_2| > |b_1|$, to get the largest element in the first column on the diagonal. The result will be

$$P^*T = \begin{pmatrix} b_1^* & c_1^* & d_1^* & \\ a_2^* & b_2^* & c_2^* & \\ & a_3 & b_3 & c_3 \\ & & \dots & \dots \end{pmatrix}$$

where the permutation matrix P^* may be the unit matrix, and then $d_1^* = 0$.

We can now write $P^*T = L_1U_1$, where

$$L_1 = \begin{pmatrix} 1 & & & \\ l_2 & 1 & & \\ & & 1 & \\ & & \dots & \dots \end{pmatrix},$$

$$U_1 = \begin{pmatrix} b_1^* & c_1^* & d_1^* & \\ & u_2 & v_2 & \\ & & a_3 & b_3 \\ & & \dots & \dots \end{pmatrix},$$

and

$$\begin{aligned} l_2 &= a_2/b_1^* \\ u_2 &= b_2 - l_2c_1^* \\ v_2 &= c_2 - l_2d_1^*. \end{aligned}$$

In the special case where $b_1 = a_2 = 0$, so that the first column of T is entirely zero, and the matrix is singular, we simply take $P^* = I$, $L_1 = I$ and $U_1 = T$.

In the matrix U_1 the $n \times n$ submatrix consisting of the last n rows

and columns is triple diagonal, and so by the induction hypothesis we can write

$$P_n \begin{pmatrix} u_2 & v_2 & \\ & a_3 & b_3 \\ & \dots & \dots \end{pmatrix} = L_n U_n.$$

This shows that

$$P^* T = \begin{pmatrix} 1 & \mathbf{0}^T \\ \boldsymbol{l}_2 & I_n \end{pmatrix} \begin{pmatrix} b_1^* & (c_1^*,\, d_1^*, 0, \dots) \\ \mathbf{0} & P_n L_n U_n \end{pmatrix}$$

where $\boldsymbol{l}_2 = (l_2, 0, 0, \dots, 0)^T$, and so

$$\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & P_n \end{pmatrix} P^* T = \begin{pmatrix} 1 & \mathbf{0}^T \\ P_n \boldsymbol{l}_2 & L_n \end{pmatrix} \begin{pmatrix} b_1^* & (c_1^*,\, d_1^*,\, 0, \dots) \\ \mathbf{0} & U_n \end{pmatrix}.$$

Thus the required result is true for all $(n+1) \times (n+1)$ matrices. Since it is obviously true for any 1×1 matrix, the induction is complete.

Solution to Exercise 3.7

This is an informal induction proof. The elements of L and U are determined by (2.19). Suppose that we are calculating the elements l_{ij} for a fixed i, with $j = 1, \ldots, i-1$, and we are about to calculate l_{ir}, where $r < i - p$. Then $b_{ir} = 0$, since B is Band(p, q). Now

$$l_{ir} = \frac{1}{u_{rr}} \left\{ b_{ir} - \sum_{k=1}^{r-1} l_{ik} u_{kr} \right\}.$$

Thus if $l_{i1} = l_{i2} = \ldots = l_{i,r-1} = 0$ it follows that $l_{ir} = 0$ also. Evidently $l_{i1} = 0$ so an induction argument shows that $l_{ik} = 0$ for all $k \leq r < i - p$. Hence L is Band$(p, 0)$.

The argument for the matrix U is similar. Calculating the elements u_{ij} in order, we are about to calculate u_{ir}, where $r > i + p$. Then $b_{ir} = 0$, and if

$$u_{1r} = u_{2r} = \ldots = u_{i-1,r} = 0$$

it follows that $u_{ir} = 0$, since

$$u_{ir} = b_{ir} - \sum_{k=1}^{i-1} l_{ik} u_{kr}.$$

Moreover it is clear that $u_{1r} = 0$, so the same induction argument shows that $u_{kr} = 0$ for all $k \leq i < r - p$. Hence U is Band$(0, q)$.

Solution to Exercise 3.8

With the usual notation, the equations determining the elements of L and U include

$$l_{21} = \frac{a_{21}}{a_{11}}$$

and

$$u_{24} = a_{24} - l_{21}a_{14}.$$

Now we are given that $a_{24} = 0$, but in general l_{21} and a_{14} are not zero. Hence in general a_{24} is not zero.

In the same way

$$l_{41} = \frac{a_{41}}{a_{11}}$$

and

$$l_{42} = \frac{1}{u_{22}}(a_{42} - l_{41}u_{12}).$$

Here $a_{42} = 0$, but there is no reason why l_{41} or u_{12} should be zero.

Although a subdiagonal of A is entirely zero, there is no reason why any of the elements of the same subdiagonal of L should be zero.

Solution to Exercise 4.1

Suppose that $\boldsymbol{g}$ is a contraction in the ∞-norm, as in (4.5). Observe that

$$\|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y})\|_p \leq n^{1/p}\|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y})\|_\infty \leq Ln^{1/p}\|\boldsymbol{x}-\boldsymbol{y}\|_\infty \leq Ln^{1/p}\|\boldsymbol{x}-\boldsymbol{y}\|_p\,.$$

Therefore, if $L < n^{-1/p}$, then $\boldsymbol{g}$ is a contraction in the p-norm.

Solution to Exercise 4.2

Substituting for x_1 we find

$$(7x_2 + 25)^2 + x_2^2 - 25 = 0,$$

or

$$50x_2^2 + 350x_2 + 600 = 0,$$

with solutions $x_2 = -3$ and -4. The corresponding values of x_1 are 4 and -3, giving the two solutions $(4, -3)^T$ and $(-3, -4)^T$.

The Jacobian matrix of $\mathbf{f}$ is

$$\begin{pmatrix} 2x_1 & 2x_2 \\ 1 & -7 \end{pmatrix}.$$

At the first solution $(4, -3)^T$ this is

$$\begin{pmatrix} 8 & -6 \\ 1 & -7 \end{pmatrix},$$

and at $(-3, -4)^T$ it is

$$\begin{pmatrix} -6 & -8 \\ 1 & -7 \end{pmatrix}.$$

Clearly the condition is not satisfied in either case.

If we change the sign of f_2 the solutions remain the same, but the signs of the elements in the second row of the Jacobian matrix are changed. The condition is now satisfied at the first solution $(4, -3)^T$, as the matrix becomes

$$\begin{pmatrix} 8 & -6 \\ -1 & 7 \end{pmatrix}.$$

If we replace $\boldsymbol{f}$ by $\boldsymbol{f}^*$ the solutions are still the same, and the Jacobian matrix becomes

$$\begin{pmatrix} 1 - 2x_1 & -7 - 2x_2 \\ -1 & 7 \end{pmatrix}.$$

At the second solution $(-3, -4)^T$ this is

$$\begin{pmatrix} 7 & 1 \\ 1 & -7 \end{pmatrix}.$$

The relaxation parameter $\lambda = 1/7$ will give convergence in both cases.

Solution to Exercise 4.3

Clearly $\varphi(0) = g(u)$ and $\varphi(1) = g(v)$. The function φ is differentiable, and

$$\varphi'(t) = g'((1-t)u + tv)\,(-u + v).$$

Hence by applying the Mean Value Theorem to φ, there is a real number θ in $(0, 1)$ such that

$$\varphi(1) - \varphi(0) = \varphi'(\theta),$$

which gives

$$g(v) - g(u) = (v - u)g'((1-\theta)u + \theta v).$$

The result now follows by defining $\eta = (1-\theta)u + \theta v$, which lies on the line between u and v, and is therefore in Ω, since Ω is convex.

Since $|g'(\zeta)| < 1$ there is a value of δ such that

$$|g'(z)| \leq k = \tfrac{1}{2}[1 + |g'(\zeta)|] < 1$$

for all z such that $|z - \zeta| \leq \delta$. Convergence of the iteration follows in the usual way, since

$$\begin{aligned} |z_{n+1} - \zeta| &= |g(z^n) - g(\zeta)| \\ &= |z_n - \zeta|\,|g'(\eta)| \\ &\leq k|z_n - \zeta|. \end{aligned}$$

Hence $|z_n - \zeta| < k^n|z_0 - \zeta|$ provided that $|z_0 - \zeta| \leq \delta$, and the sequence (z_n) converges to ζ.

Solution to Exercise 4.4

The iteration

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{g}^*(\boldsymbol{x}^{(n)})$$

is in component form

$$\begin{aligned} x_1^{(n+1)} &= u(x_1^{(n)}, x_2^{(n)}) \\ x_2^{(n+1)} &= v(x_1^{(n)}, x_2^{(n)}) \end{aligned}$$

and the complex iteration $z_{n+1} = g(z_n)$ gives

$$x_1^{(n+1)} + \imath x_2^{(n+1)} = u(x_1^{(n)}, x_2^{(n)}) + \imath v(x_1^{(n)}, x_2^{(n)}),$$

with $\imath = \sqrt{-1}$, which are obviously identical.

The condition $|g(\zeta)| < 1$ gives

$$u_x^2 + v_x^2 < 1,$$

evaluated at the fixed point $\zeta = \zeta_1 + \imath\zeta_2$.

In its real form a sufficient condition for convergence is $\|J(\zeta_1, \zeta_2)\|_\infty < 1$, and the Jacobian matrix is

$$J = \begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix}.$$

Using the Cauchy-Riemann relations this gives the sufficient condition

$$|u_x| + |v_x| < 1.$$

This is a more restrictive condition than that obtained from the complex form, as it leads to

$$|u_x|^2 + |v_x|^2 < 1 - 2|u_x v_x|.$$

Solution to Exercise 4.5

Clearly $g_1(1,1) = 1$ and $g_2(1,1) = 1$, so that (1,1) is a fixed point. The Jacobian matrix is

$$J = \begin{pmatrix} \frac{2}{3}x_1 & -\frac{2}{3}x_2 \\ \frac{2}{3}x_2 & \frac{2}{3}x_1 \end{pmatrix}.$$

Evidently at the fixed point $\|J\|_\infty = \frac{4}{3} > 1$, so the sufficient condition for convergence is not satisfied.

However, as in Exercise 4 this iteration corresponds to complex iteration with the function $g(z) = \frac{1}{3}(z^2 + 3 + i)$, as can be seen by writing down the real and imaginary parts. Then $g'(z) = \frac{2}{3}z$, and at the fixed point

$$|g'(1+i)| = |\tfrac{2}{3}\sqrt{2}| < 1,$$

so the iteration converges.

Solution to Exercise 4.6

In component form the function $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x} - K(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})$ is

$$g_i(\boldsymbol{x}) = x_i - \sum_{r=1}^{k} K_{ir}(\boldsymbol{x}) f_r(\boldsymbol{x}).$$

Differentiating with respect to x_j gives the (i, j) element of the Jacobian matrix of $\boldsymbol{g}$ as

$$\begin{aligned}
\frac{\partial g_i}{\partial x_j} &= \delta_{ij} - \frac{\partial}{\partial x_j} \sum_{r=1}^{k} K_{ir}(\boldsymbol{x}) f_r(\boldsymbol{x}) \\
&= \delta_{ij} - \sum_{r=1}^{k} \frac{\partial K_{ir}}{\partial x_j} f_r - \sum_{r=1}^{k} K_{ir} \frac{\partial f_r}{\partial x_j} \\
&= \delta_{ij} - \sum_{r=1}^{k} \frac{\partial K_{ir}}{\partial x_j} f_r - \sum_{r=1}^{k} K_{ir} J_{rj},
\end{aligned}$$

all evaluated at the point $\boldsymbol{x}$.

When we evaluate this at the point $\boldsymbol{\xi}$, we know that $\boldsymbol{f}(\boldsymbol{\xi}) = \boldsymbol{0}$, so that $f_r = 0$ for each value of r. Moreover, K is the inverse of the Jacobian matrix, J, of $\boldsymbol{f}$, so that

$$\sum_{r=1}^{k} K_{ir} J_{rj} = \delta_{ij}.$$

Hence all the elements of the Jacobian matrix of $\boldsymbol{g}$ vanish at the point $\boldsymbol{\xi}$.

Solution to Exercise 4.7

Evidently at a solution $x_1 = x_2$, and $2x_1^2 = 2$. Hence there are two solutions, $(1,1)^T$ and $(-1,-1)^T$. The Jacobian matrix of $\boldsymbol{f}$ is

$$J = \begin{pmatrix} 2x_1 & 2x_2 \\ 1 & -1 \end{pmatrix},$$

and its inverse is

$$J^{-1} = \frac{1}{2(x_1+x_2)} \begin{pmatrix} 1 & 2x_2 \\ 1 & 2x_1 \end{pmatrix}.$$

Hence Newton's method gives

$$\begin{aligned} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} &= \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} - \frac{1}{2(x_1^{(0)}+x_2^{(0)})} \begin{pmatrix} 1 & 2x_2^{(0)} \\ 1 & 2x_1^{(0)} \end{pmatrix} \begin{pmatrix} (x_1^{(0)})^2 + (x_2^{(0)})^2 - 2 \\ x_1^{(0)} - x_2^{(0)} \end{pmatrix} \\ &= \frac{1}{2(x_1^{(0)}+x_2^{(0)})} \begin{pmatrix} \left(x_1^{(0)}\right)^2 + \left(x_2^{(0)}\right)^2 + 2 \\ \left(x_1^{(0)}\right)^2 + \left(x_2^{(0)}\right)^2 + 2 \end{pmatrix} \end{aligned}$$

Thus $x_1^{(n)} = x_2^{(n)}$ for all positive values of n, and if we write $x_1^{(n)} = x_2^{(n)} = x^{(n)}$ the iteration becomes

$$x^{(n+1)} = \frac{\left(x^{(n)}\right)^2 + 1}{2x^{(n)}}.$$

Evidently if $x^{(0)} > 0$ then $x^{(n)} > 1$ for all positive n. Moreover

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

and $0 < x - 1 < 2x$ when $x > 1$. Hence $x^{(n)} \to 1$ as $n \to \infty$. It then follows from

$$x^{(n+1)} - 1 = \frac{1}{2x^{(n)}}(x^{(n)} - 1)^2$$

that

$$\frac{x^{(n+1)} - 1}{(x^{(n)} - 1)^2} \to \frac{1}{2} \qquad \text{as } n \to \infty,$$

so that convergence is quadratic.

A trivial modification of the argument shows that if $x_1^{(0)} + x_2^{(0)} < 0$ then the iteration converges quadratically to the solution $(-1,-1)^T$.

Solution to Exercise 4.8

We say that $(\boldsymbol{x}^{(k)})$ converges to $\boldsymbol{\xi}$ linearly if (4.19) holds with $q = 1$, $0 < \mu < 1$ and $\varepsilon_k = \|\boldsymbol{x}^{(k)} - \boldsymbol{\xi}\|_\infty$. The rate of convergence is defined as $\log_{10}(1/\mu) = -\log_{10}\mu$.

The Jacobian matrix of $\boldsymbol{f}$ is

$$J = \begin{pmatrix} 2x_1 & 2x_2 \\ 1 & 1 \end{pmatrix},$$

whose inverse is

$$J^{-1} = \frac{1}{2(x_1 - x_2)} \begin{pmatrix} 1 & -2x_2 \\ -1 & 2x_1 \end{pmatrix},$$

provided that $x_1 \neq x_2$. Newton's method then gives, provided that $x_1^{(n)} \neq x_2^{(n)}$,

$$\begin{pmatrix} x_1^{(n+1)} \\ x_2^{(n+1)} \end{pmatrix} = \begin{pmatrix} x_1^{(n)} \\ x_2^{(n)} \end{pmatrix} - \frac{1}{2(x_1^{(n)} - x_2^{(n)})} \begin{pmatrix} 1 & -2x_2^{(n)} \\ -1 & 2x_1^{(n)} \end{pmatrix} \begin{pmatrix} \left(x_1^{(n)}\right)^2 + \left(x_2^{(n)}\right)^2 \\ x_1^{(n)} + x_2^{(n)} - 2 \end{pmatrix},$$

and a simple calculation then shows that $x_1^{(n+1)} + x_2^{(n+1)} = 2$.

When $x_1^{(0)} = 1 + \alpha,\ x_2^{(0)} = 1 - \alpha$ we find that

$$x_1^{(1)} = 1 + \tfrac{1}{2}\alpha, \qquad x_2^{(1)} = 1 - \tfrac{1}{2}\alpha.$$

[These are exact values, not approximations for small α.]

We have shown that for any $\boldsymbol{x}^{(0)}$, provided $x_1^{(0)} \neq x_2^{(0)}$, the result of the first iteration satisfies $x_1^{(1)} + x_2^{(1)} = 2$. Hence we can write $x_1^{(1)} = 1 + \alpha,\ x_2^{(1)} = 1 - \alpha$. Then

$$x_1^{(n)} = 1 + \alpha/2^{n-1}, \qquad x_2^{(n)} = 1 - \alpha/2^{n-1}$$

for $n \geq 1$; this shows that the $(\boldsymbol{x}^{(n)})$ converges linearly to $(1,1)^T$, with rate of convergence $\ln 2$.

Convergence is not quadratic, because the Jacobian matrix of $\boldsymbol{f}$ is singular at the limit point $(1,1)^T$.

Solution to Exercise 4.9

With the given value of z we get

$$\begin{aligned} z+2 &= (2k+\tfrac{1}{2})\imath\pi + \ln[(2k+\tfrac{1}{2})\pi] + \eta + 2 \\ &= \mathrm{e}^z \\ &= (2k+\tfrac{1}{2})\imath\pi \mathrm{e}^\eta. \end{aligned}$$

Hence

$$\eta = \ln\left(1 - \imath\frac{\ln[(2k+\frac{1}{2}\pi] + \eta + 2}{2k+\frac{1}{2}\pi}\right).$$

Now

$$|\ln(1+\imath t)| < |t|$$

for all nonzero real values of t, so that

$$|\eta| < \frac{\ln[(2k+\frac{1}{2})\pi] + |\eta| + 2}{(2k+\frac{1}{2})\pi},$$

which gives

$$|\eta| < \frac{\ln[(2k+\frac{1}{2}\pi)] + 2}{(2k+\frac{1}{2})\pi}.$$

This is enough to give the required result, since for sufficiently large k the other terms in the numerator and denominator become negligible, and $|\eta|$ is bounded in the limit by $\ln k/2k\pi$.

To give a formal analysis, we obviously find in the denominator that

$$(2k+\tfrac{1}{2})\pi > 2\pi k.$$

For the numerator, we note that $1 < \ln k$ for $k > 1$, and so

$$\begin{aligned} \ln[(2k+\tfrac{1}{2})\pi] + 2 &< \ln[(2k+\tfrac{1}{2})\pi] + 2\ln k \\ &< 3\ln k + \ln[(2+\tfrac{1}{2k})\pi] \\ &< 3\ln k + \ln(3\pi) \\ &< 3\ln k + \ln(3\pi)\ln k \\ &< [3+\ln(3\pi)]\ln k. \end{aligned}$$

Finally

$$|\eta| < C\frac{\ln k}{k},$$

where

$$C = \frac{3+\ln(3\pi)}{2\pi}.$$

Solution to Exercise 5.1

(i) Lemma 5.2
For the Householder matrix

$$H = I - \frac{2}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^T,$$

which is clearly symmetric, we have

$$\begin{aligned} H\,H^T &= H^2 \\ &= (I - \alpha\boldsymbol{v}\boldsymbol{v}^T)(I - \alpha\boldsymbol{v}\boldsymbol{v}^T) \\ &= I - 2\alpha\boldsymbol{v}\,\boldsymbol{v}^T + \alpha^2\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{v}\boldsymbol{v}^T \\ &= I + (\alpha^2\boldsymbol{v}^T\boldsymbol{v} - 2\alpha)\boldsymbol{v}\,\boldsymbol{v}^T \\ &= I, \end{aligned}$$

where we have written $\alpha = 2/(\boldsymbol{v}^T\boldsymbol{v})$. Hence H is orthogonal.

(ii)Lemma 5.3
Since H_k is a Householder matrix,

$$H_k = I - \frac{2}{\boldsymbol{v}_k^T\boldsymbol{v}_k}\boldsymbol{v}_k\boldsymbol{v}_k^T,$$

for some vector $\boldsymbol{v}_k$. Now define the vector $\boldsymbol{v}$, with n elements, partitioned as

$$\begin{pmatrix} 0 \\ \boldsymbol{v}_k \end{pmatrix},$$

so that the first $n-k$ elements of $\boldsymbol{v}$ are zero. Then $\boldsymbol{v}^T\boldsymbol{v} = \boldsymbol{v}_k^T\boldsymbol{v}_k$, and

$$\begin{aligned} I - \frac{2}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v}\,\boldsymbol{v}^T &= \begin{pmatrix} I_{n-k} & 0 \\ 0 & I_k \end{pmatrix} - \frac{2}{\boldsymbol{v}^T\boldsymbol{v}}\begin{pmatrix} 0 \\ \boldsymbol{v}_k \end{pmatrix}(0 \quad \boldsymbol{v}_k^T) \\ &= \begin{pmatrix} I_{n-k} & 0 \\ 0 & H_k \end{pmatrix} \\ &= A, \end{aligned}$$

so that A is a Householder matrix.

Solution to Exercise 5.2

Since this is a 4×4 matrix, two Householder transformations are required.

In the first transformation the Householder vector is $\boldsymbol{v} = (0, 4, 2, 2)^T$. The result of the first transformation is the matrix

$$\begin{pmatrix} 2 & -3 & 0 & 0 \\ -3 & 1 & 3 & 4 \\ 0 & 3 & -3 & -9 \\ 0 & 4 & -9 & -2 \end{pmatrix}.$$

In the second transformation the Householder vector is $\boldsymbol{v} = (0, 0, 8, 4)^T$, and the result is

$$\begin{pmatrix} 2 & -3 & 0 & 0 \\ -3 & 1 & -5 & 0 \\ 0 & -5 & -11 & -3 \\ 0 & 0 & -3 & 6 \end{pmatrix}.$$

This is the required tridiagonal form. Note that the first row and column, and the leading 2×2 principal minor, are unaffected by the second transformation.

Solution to Exercise 5.3

Taking $\theta = 0$, the corresponding Sturm sequence is

$$\begin{aligned} p_0 &= 1 \\ p_1 &= 3 - \theta = 3 \\ p_2 &= (2-\theta)(3) - 1 = 5 \\ p_3 &= (4-\theta)(5) - (4)(3) = 8 \\ p_4 &= (1-\theta)(8) - \alpha^2(5) = 8 - 5\alpha^2. \end{aligned}$$

If $5\alpha^2 < 8$ there are 4 agreements in sign, while if $5\alpha^2 > 8$ there are 3 agreements in sign.

Taking $\theta = 1$, the corresponding Sturm sequence is

$$\begin{aligned} p_0 &= 1 \\ p_1 &= 3 - \theta = 2 \\ p_2 &= (2-\theta)(2) - 1 = 1 \\ p_3 &= (4-\theta)(1) - (4)(2) = -5 \\ p_4 &= (1-\theta)(-5) - \alpha^2(1) = -\alpha^2. \end{aligned}$$

In this case there are always three agreements in sign.

Hence if $5\alpha^2 > 8$ the number of agreements in sign is the same, so no eigenvalues lie in $(0,1)$, while if $5\alpha^2 < 8$ there is one more agreement in sign for $\theta = 0$ than for $\theta = 1$, so there is exactly one eigenvalue in $(0,1)$. If $5\alpha^2 = 8$ then there is an eigenvalue at 0.

Solution to Exercise 5.4

If $H\boldsymbol{x} = c\,\boldsymbol{y}$, where c is a scalar, then

$$\boldsymbol{x}^T H^T H \boldsymbol{x} = c^2 \boldsymbol{y}^T \boldsymbol{y};$$

but since H is a Householder matrix, $H^T H = I$, and so

$$c^2 = \frac{\boldsymbol{x}^T \boldsymbol{x}}{\boldsymbol{y}^T \boldsymbol{y}}.$$

Writing

$$H = I - \frac{2}{\boldsymbol{v}^T \boldsymbol{v}} \boldsymbol{v}\, \boldsymbol{v}^T$$

this shows that

$$\boldsymbol{x} - \frac{2}{\boldsymbol{v}^T \boldsymbol{v}} (\boldsymbol{v}^T \boldsymbol{x})\, \boldsymbol{v} = c\boldsymbol{y}.$$

Hence $\boldsymbol{v} = \alpha(\boldsymbol{x} - c\boldsymbol{y})$, where c is known, and α is some scalar. Now H is unaffected by multiplying $\boldsymbol{v}$ by an arbitrary nonzero scalar, and the required Householder matrix is $H = H(\boldsymbol{v})$, where $\boldsymbol{v} = \boldsymbol{x} \pm c\boldsymbol{y}$, and

$$c = \sqrt{\frac{\boldsymbol{x}^T \boldsymbol{x}}{\boldsymbol{y}^T \boldsymbol{y}}}.$$

There are clearly two possible such Householder matrices.

Solution to Exercise 5.5

From

$$(D + \varepsilon A)(\boldsymbol{e} + \varepsilon \boldsymbol{u}) = (\lambda + \varepsilon\mu)(\boldsymbol{e} + \varepsilon \boldsymbol{u})$$

we find by equating powers of ε^0 and ε^1 that

$$D\boldsymbol{e} = \lambda \boldsymbol{e}$$

and

$$D\boldsymbol{u} + A\boldsymbol{e} = \lambda \boldsymbol{u} + \mu \boldsymbol{e}.$$

Since D is diagonal it follows from the first of these that $\lambda = d_{jj}$ for some j, and that $\boldsymbol{e}$ is the unit vector whose only nonzero element is $e_j = 1$. From the second equation,

$$d_{kk}u_k + a_{kj} = d_{jj}u_k + \mu\delta_{kj}, \qquad k = 1, \ldots, n.$$

Taking $k = j$ we see at once that $\mu = a_{jj}$. For $k \neq j$ it also follows that

$$u_k = \frac{a_{kj}}{d_{jj} - d_{kk}}.$$

Since the eigenvector $\boldsymbol{e} + \varepsilon \boldsymbol{u}$ has to be normalised, we have

$$(\boldsymbol{e}^T + \varepsilon \boldsymbol{u}^T)(\boldsymbol{e} + \varepsilon \boldsymbol{u}) = 1.$$

Comparing coefficients of powers of ε, we first see that $\boldsymbol{e}^T\boldsymbol{e} = 1$, which already holds, and then $\boldsymbol{e}^T\boldsymbol{u} = 0$. This means that $u_j = 0$.

Solution to Exercise 5.6

Multiplying out the partitioned equation and equating powers of ε gives, for the leading term,

$$\begin{aligned} d_{11}\boldsymbol{e} &= \lambda\boldsymbol{e} \\ D_{n-k}\boldsymbol{f} &= \lambda\boldsymbol{f}. \end{aligned}$$

From the first of these we see that $\lambda = d_{11}$; since none of the elements of the diagonal matrix D_{n-k} are equal to d_{11} the second equation shows that $\boldsymbol{f} = \boldsymbol{0}$.

Now equating coefficients of ε we get

$$\begin{aligned} A_1\boldsymbol{e} + d_{11}\boldsymbol{u} + A_2\boldsymbol{f} &= \lambda\boldsymbol{u} + \mu\boldsymbol{e} \\ A_2^T\boldsymbol{e} + A_3\boldsymbol{f} + D_{n-k}\boldsymbol{v} &= \lambda\boldsymbol{v}. \end{aligned}$$

Making use of the facts that $\lambda = d_{11}$ and $\boldsymbol{f} = \boldsymbol{0}$ the first of these equations shows that μ is an eigenvalue of A_1, and the second shows that

$$A_2^T\boldsymbol{e} + D_{n-k}\boldsymbol{v} = \lambda\boldsymbol{v}.$$

This equation determines $\boldsymbol{v}$, since the matrix $D_{n-k} - \lambda I$ is not singular.

Now equating coefficients of ε^2 we get

$$\begin{aligned} d_{11}\boldsymbol{x} + A_1\boldsymbol{u} + A_2\boldsymbol{v} &= \lambda\boldsymbol{x} + \mu\boldsymbol{u} + \nu\boldsymbol{e} \\ A_2^T\boldsymbol{u} + D_{n-k}\boldsymbol{y} + A_3\boldsymbol{v} &= \lambda\boldsymbol{y} + \mu\boldsymbol{v} + \nu\boldsymbol{f}. \end{aligned}$$

The first of these reduces to

$$(A_1 - \mu I)\boldsymbol{u} = \nu\boldsymbol{e} - A_2\boldsymbol{v}.$$

Determination of $\boldsymbol{u}$ from this equation is not quite straightforward, since the matrix $A_1 - \mu I$ is singular. However, if we know the eigenvalues θ_j and corresponding eigenvectors $\boldsymbol{w}_j$ of A_1 we can write

$$\boldsymbol{u} = \sum_j \beta_j \boldsymbol{w}_j.$$

If we number these eigenvalues so that $\mu = \theta_1$ and $\boldsymbol{e} = \boldsymbol{w}_1$, we see that

$$\sum_j (\theta_j - \theta_1)\beta_j \boldsymbol{w}_j = \nu\boldsymbol{w}_1 - A_2\boldsymbol{v}.$$

Multiplying by $\boldsymbol{w}_i^T$ we obtain

$$(\theta_i - \theta_1)\beta_i = -\boldsymbol{w}_i^T A_2 \boldsymbol{v};$$

This determines β_i for $i \neq 1$, since we have assumed that the eigenvalues θ_j are distinct. The equation does not determine the coefficient β_1; as in Exercise 5 this is given by the requirement that the eigenvector of A is normalised. Writing down this condition and comparing coefficients of ε we see that $\boldsymbol{e}^T\boldsymbol{u} = 0$, thus showing that $\beta_1 = 0$.

Solution to Exercise 5.7

Since $A - \mu I = QR$ and Q is orthogonal, $Q^T(A - \mu\, I) = R$. Hence

$$\begin{aligned} B &= RQ + \mu\, I \\ &= Q^T(A - \mu\, I)Q + \mu\, I \\ &= Q^T A\, Q - \mu Q^T\, I\, Q + \mu\, I \\ &= Q^T A\, Q. \end{aligned}$$

Hence B is an orthogonal transformation of A.

Also

$$\begin{aligned} B^T &= (Q^T AQ)^T \\ &= Q^T A^T Q \\ &= Q^T AQ \end{aligned}$$

since A is symmetric. Hence $B^T = B$, so that B is symmetric.

We build up the matrix Q as a product of plane rotation matrices $R^{p,p+1}(\varphi_p)$ as in (5.34), with $p = 1, \ldots, n-1$. The first of these rotations, with $p = 1$, replaces rows 1 and 2 of the matrix $A - \mu I$ with linear combinations of these two rows, such that the new $(2,1)$ element is zero. since the matrix is tridiagonal the new element $(1,3)$ will in general be nonzero. The second rotation, with $p = 2$, carries out a similar operation on rows 2 and 3; in the result the element $(3,2)$ becomes zero, and $(2,4)$ may be nonzero.

We now form the matrix RQ; this involves taking the matrix R and applying the same sequence of plane rotations on the right, but with each rotation transposed. Since R is upper triangular, and the first rotation operates on columns 1 and 2, a single nonzero element appears below the diagonal at position $(2,1)$. In the same way the second rotation operates on columns 2 and 3, and introduces a nonzero element at $(3,2)$. Hence finally the only nonzero elements in the matrix B below the diagonal are in positions $(i+1, i)$, for $i = 1, \ldots, n-1$. But we have just shown that B is symmetric; hence B is also tridiagonal.

Solution to Exercise 5.8

For this matrix the shift $\mu = a_{nn} = 0$; we see that

$$A = QR = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Evidently Q is orthogonal and R is upper triangular. It is then easy to verify that $B = RQ = A$. Hence in successive iterations of the QR algorithm all the matrices $A^{(k)}$ are the same as A, so they do not converge to a diagonal matrix.

Solution to Exercise 5.9

We use the shift $a_{n,n} = 10$, so that

$$A - \mu I = \begin{pmatrix} 3 & 4 \\ 4 & 0 \end{pmatrix}.$$

The QR decomposition of this matrix is

$$A - \mu I = QR = \begin{pmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix} \begin{pmatrix} 5 & \frac{12}{5} \\ 0 & -\frac{16}{5} \end{pmatrix}.$$

Then the new matrix $A^{(1)}$ is

$$RQ + \mu I = \begin{pmatrix} 5 & \frac{12}{5} \\ 0 & -\frac{16}{5} \end{pmatrix} \begin{pmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix} + 10\, I$$

which is

$$\begin{pmatrix} \frac{373}{25} & -\frac{64}{25} \\ -\frac{64}{25} & \frac{202}{25} \end{pmatrix}.$$

Solution to Exercise 6.1

The inequalities follow immediately from Theorem 6.2, with $n = 1$, $x_0 = -1$ and $x_1 = 1$.

As an example we need a function whose second derivative is constant, so that $f''(\xi) = M$ for any value of ξ; thus we choose $f(x) = x^2$, and $M_2 = 2$. Since $f(1) = f(-1) = 1$ the interpolation polynomial of degree 1 is $p_1(x) = 1$, and at $x = 0$

$$|f(0) - p_1(0)| = |0 - 1| = 1 = M_2/2$$

as required.

Solution to Exercise 6.2

(i) The interpolation polynomial is

$$p_1(x) = \frac{x-a}{-a}0^3 + \frac{x}{a}a^3 = xa^2.$$

The difference is

$$f(x) - p_1(x) = x^3 - xa^2 = x(x-a)(x+a).$$

Theorem 6.2 gives

$$f(x) - p_1(x) = \frac{x(x-a)}{2!}f''(\xi) = 3x(x-a)\xi.$$

Comparing these we see that

$$\xi = (x+a)/3.$$

(ii) The same calculation for $f(x) = (2x-a)^4$ gives

$$p_1(x) = \frac{x-a}{-a}(-a)^4 + \frac{x}{a}(2a-a)^4,$$

and the difference is

$$\begin{aligned} f(x) - p_1(x) &= (2x-a)^4 + (x-a)a^3 - xa^3 \\ &= 8x(x-a)(2x^2 - 2ax + a^2). \end{aligned}$$

Comparing with

$$\frac{x(x-a)}{2!}48(2\xi - a)^2$$

we see that there are two values of ξ, given by

$$\xi = \tfrac{1}{2}a \pm \left(\frac{2x^2 - 2ax + a^2}{12}\right)^{1/2}.$$

Solution to Exercise 6.3

We know that

$$q(x_i) = y_i, \qquad i = 0, \ldots, n,$$

and that

$$r(x_i) = y_i, \qquad i = 1, \ldots, n+1.$$

Now suppose that $1 \leq j \leq n$; then $q(x_j) = r(x_j) = y_j$. Hence

$$\begin{aligned} p(x_j) &= \frac{(x_j - x_0)r(x_j) - (x_j - x_{n+1})q(x_j)}{x_{n+1} - x_0} \\ &= \frac{(x_j - x_0)y_j - (x_j - x_{n+1})y_j}{x_{n+1} - x_0} \\ &= y_j. \end{aligned}$$

Clearly

$$\begin{aligned} p(x_0) &= \frac{(x_0 - x_0)r(x_0) - (x_0 - x_{n+1})q(x_0)}{x_{n+1} - x_0} \\ &= q(x_0) \\ &= y_0, \end{aligned}$$

and in the same way $p(x_{n+1}) = y_{n+1}$.

Hence $p(x_j) = y_j,\ j = 0, \ldots, n+1$, showing that $p(x)$ is the Lagrange interpolation polynomial for the points $\{(x_i, y_i) \ : \ i = 0, \ldots, n+1\}$.

Solution to Exercise 6.4

We find that

$$\begin{aligned}
\pi_{n+1}(1-1/n) &= (2-1/n)(2-3/n)\ldots(1/n)(-1/n) \\
&= \frac{(2n-1)(2n-3)\ldots(1)(-1)}{n^{n+1}} \\
&= -\frac{1}{n^{n+1}}\,\frac{(2n)!}{2.4.\ldots.(2n)} \\
&= -\frac{(2n)!}{n^{n+1}\,2^n\,n!}.
\end{aligned}$$

Substitution of Stirling's formula (cf. also Chapter 2, Section 2.1) gives

$$\begin{aligned}
\pi_{n+1}(1-1/n) &\sim -\frac{1}{n^{n+1}\,2^n}\,\frac{(2n)^{2n+1/2}\mathrm{e}^{-2n}}{n^{n+1/2}\,\mathrm{e}^{-n}} \\
&= -\frac{2^{n+1/2}\,\mathrm{e}^{-n}}{n}, \qquad n\to\infty
\end{aligned}$$

as required.

Solution to Exercise 6.5

Suppose that there exists $q_{2n+1} \in \mathcal{P}_{2n+1}$, different from p_{2n+1} and also having these properties. Define $r = p_{2n+1} - q_{2n+1}$; then $r(x_i) = 0,\ i = 0, \ldots, n$. By Rolle's Theorem there exist n points $\xi_j,\ j = 1, \ldots, n$, one between each consecutive pair of x_i, at which $r'(\xi_j) = 0$. Applying Rolle's Theorem again, there exist $n-1$ points $\eta_k,\ k = 1, \ldots, n-1$, at which $r''(\eta_k) = 0$. But we also know that $r''(x_i) = 0,\ i = 0, \ldots, n$, making a total of $2n$ points at which r'' vanishes. However, the proof breaks down at this point, since we cannot be sure that each η_k is distinct from all the points x_i; all we know is that η_k lies between x_{k-1} and x_{k+1} and it is possible that $\eta_k = x_k$.

Suppose that $p_5(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + c_5 x^5$. The given conditions lead to the equations

$$\begin{aligned}
c_0 - c_1 + c_2 - c_3 + c_4 - c_5 &= 1 \\
c_0 &= 0 \\
c_0 + c_1 + c_2 + c_3 + c_4 + c_5 &= 1 \\
2c_2 - 6c_3 + 12c_4 - 20c_5 &= 0 \\
2c_2 &= 0 \\
2c_2 + 6c_3 + 12c_4 + 20c_5 &= 0.
\end{aligned}$$

Adding equations 1 and 3, and using equation 2, gives

$$c_2 + c_4 = 2.$$

Now $c_0 = c_2 = 0$, so that $c_4 = 2$. Adding equations 4 and 6 gives

$$4c_2 + 24c_4 = 0,$$

which is clearly inconsistent. Hence this system of equations have no solution.

With the change $p_5(-1) = -1$, adding equations 1 and 3 now gives

$$c_2 + c_4 = 0,$$

so the same argument now leads to $c_0 = c_2 = c_4 = 0$. The equations now only require that

$$\begin{aligned}
c_1 + c_3 + c_5 &= 1 \\
6c_3 + 20c_5 &= 0.
\end{aligned}$$

Hence $c_5 = \alpha$ can be chosen arbitrarily, and then $c_3 = -\frac{10}{3}\alpha,\ c_1 =$

$1 + \frac{7}{3}\alpha$. A general form of the polynomial satisfying the given conditions is

$$p_5(x) = x + \beta(7x - 10x^3 + 3x^5),$$

where β is arbitrary.

Solution to Exercise 6.6

The polynomial $l_0(x)$ must satisfy the conditions

$$l_0(x_0) = 1, \;\; l_0(x_i) = l'(x_i) = 0, \;\; i = 1, \ldots, n.$$

This shows that

$$l_0(x) = \prod_{i=1}^{n} \frac{(x - x_i)^2}{(x_0 - x_i)^2}.$$

The polynomial $h_i(x)$ must satisfy

$$h_i(x_j) = \delta_{ij}, \;\; j = 0, \ldots, n,$$

$$h_i'(x_j) = 0, \quad j = 1, \ldots, n.$$

It must therefore contain the factors $(x - x_j)^2$, $j = 1, \ldots, n$, $j \neq i$, and $(x - x_0)$. There must be one other linear factor, and so

$$h_i(x) = (1 + \alpha(x - x_i)) \frac{x - x_0}{x_i - x_0} [L_i(x)]^2,$$

where

$$L_i(x) = \prod_{j=1, j\neq i}^{n} \frac{x - x_j}{x_i - x_j},$$

and the value of α is determined by the condition $h_i'(x_i) = 0$. It is easy to see that

$$h_i'(x_i) = \alpha + \frac{1}{x_i - x_0} + 2L_i'(x_i)$$

and so

$$\alpha = -\frac{1}{x_i - x_0} - 2L_i'(x_i).$$

In the same way the polynomial $k_i(x)$ must satisfy

$$k_i(x_j) = 0, \;\; j = 0, \ldots, n,$$

and

$$k_i'(x_j) = \delta_{ij}, \quad j = 1, \ldots, n.$$

It is easy to see that

$$k_i(x) = (x - x_i)[L_i(x)]^2 \frac{x - x_0}{x_i - x_0}.$$

Each of these polynomials is of degree $2n$.

As in the proof of Theorem 6.4, we consider the function

$$\psi(t) = f(t) - p_{2n}(t) - \frac{f(x) - p_{2n}(x)}{\pi(x)^2}\pi(t)^2,$$

where now

$$\pi(x) = (x - x_0)\prod_{i=1}^{n}(x - x_i)^2.$$

If x is distinct from all of the x_j, then $\psi(t)$ vanishes at all the points x_j, and at x. Hence by Rolle's Theorem $\psi'(t)$ vanishes at $n+1$ points lying between them; $\psi'(t)$ also vanishes at x_j, $j = 1, \ldots, n$. This means that $\psi'(t)$ vanishes at $2n + 1$ distinct points, so by repeated applications of Rolle's Theorem we see that $\psi^{(2n+1)}(\eta) = 0$ for some η in (a, b). This gives the required result.

Solution to Exercise 6.7

The expressions are a natural extension of those in Exercise 6.

We now define

$$L_i(x) = \prod_{j=1, j\neq i}^{n-1} \frac{x - x_j}{x_i - x_j},$$

and we see that

$$\begin{aligned} l_0(x) &= L_0(x)^2 \frac{x - x_n}{x_0 - x_n}, \\ l_n(x) &= L_n(x)^2 \frac{x - x_0}{x_n - x_0}. \end{aligned}$$

These are both polynomials of degree $2n-1$, and satisfy the conditions we require.

The polynomials h_i and k_i are given by

$$h_i(x) = [1 + \alpha(x - x_i)] \frac{(x - x_0)(x - x_n)}{(x_i - x_0)(x_i - x_n)} L_i(x)^2,$$

where

$$\alpha = -\frac{1}{x_i - x_0} - \frac{1}{x_i - x_n} + 2L_i'(x_i)$$

and

$$k_i(x) = (x - x_i)[L_i(x)]^2 \frac{(x - x_0)(x - x_n)}{(x_i - x_0)(x_i - x_n)}.$$

The error bound is obtained as in Exercise 6 by considering the function

$$\psi(t) = f(t) - p_{2n-1}(t) - \frac{f(x) - p_{2n-1}(x)}{\pi(x)^2} \pi(t)^2,$$

where now

$$\pi(x) = (x - x_0)(x - x_n) \prod_{i=1}^{n-1} (x - x_i)^2.$$

Solution to Exercise 6.8

The zeros of the polynomial are at $-2, -1, 0, 1, 2$ and 3. These are symmetric about the centre point $\frac{1}{2}$, and the polynomial has even degree. Hence it is symmetric about the point $x = \frac{1}{2}$. Since it does not vanish in the interval $(0, 1)$ the maximum is attained at the midpoint, $\frac{1}{2}$. The maximum value is

$$q(\tfrac{1}{2}) = \tfrac{225}{64}.$$

Write $x = x_k + \theta\pi/8$, where k is an integer and $0 \leq \theta < 1$. Then writing $u(x) = p(\theta)$, where $p(\theta)$ is the Lagrange interpolation polynomial, we see that p is defined by the interpolation points $(j, f(x_j)), j = -2, \ldots, 3$. Then the difference between $f(x)$ and $u(x)$ is just the error in the Lagrange polynomial, which is

$$\sin x - u(x) \quad = \quad \frac{\prod_{j=-2}^{3}(\theta - j)}{6!} g^{VI}(\eta),$$

where $-2 < \eta < 3$, and g^{VI} denotes the 6th derivative of the function $g(\theta) = f(x_k + \theta\pi/8)$ with respect to θ. Hence

$$g^{VI}(\eta) = (\pi/8)^6 f^{VI}(\xi),$$

where $x_{k-2} < \xi < x_{k+3}$. Now $f(x) = \sin x$, so

$$|g^{VI}(\eta)| \leq (\pi/8)^6.$$

Hence using the above bound on the polynomial.

$$\begin{aligned} |\sin x - u(x)| &= \left| \frac{p(\theta)}{6!} g^{VI}(\eta) \right| \\ &\leq \frac{225}{64} \frac{1}{6!} \left(\frac{\pi}{8}\right)^6 \\ &\approx 0.000018. \end{aligned}$$

Solution to Exercise 6.9

In the interpolation polynomial the coefficient of $f(x_j)$, where $0 \leq j \leq n-1$, is

$$\begin{aligned}\prod_{r=0,r\neq j}^{2n-1} \frac{x-x_r}{x_j-x_r} &= \prod_{r=0,r\neq j}^{n-1} \frac{x-x_r}{x_j-x_r} \prod_{r=0}^{n-1} \frac{x-x_r-\varepsilon}{x_j-x_r-\varepsilon} \\ &= \frac{\varphi_j(x)}{\varphi_j(x_j)} \frac{\varphi_j(x-\varepsilon)}{\varphi_j(x_j-\varepsilon)} \left(\frac{x-x_j-\varepsilon}{-\varepsilon}\right).\end{aligned}$$

In the same way the coefficient of $f(x_{n+j})$ is

$$\prod_{r=0}^{n-1} \frac{x-x_r}{x_j+\varepsilon-x_r} \prod_{r=0,r\neq j}^{n-1} \frac{x-x_r-\varepsilon}{x_j+\varepsilon-x_r-\varepsilon} = \frac{x-x_j}{\varepsilon} \frac{\varphi_j(x)}{\varphi_j(x_j)} \frac{\varphi_j(x-\varepsilon)}{\varphi_j(x_j+\varepsilon)}.$$

Adding these two terms gives the required expression.

In the limit as $\varepsilon \to 0$ it is clear that

$$\varphi_j(x-\varepsilon) \to \varphi_j(x).$$

The required limit can therefore be written

$$\frac{[\varphi_j(x)]^2}{\varphi_j(x_j)} \lim_{\varepsilon\to 0} \frac{G(\varepsilon)}{\varepsilon} = \frac{[\varphi_j(x)]^2}{\varphi_j(x_j)} G'(0),$$

where

$$G(\varepsilon) = \frac{x-x_j}{\varphi_j(x_j+\varepsilon)} f(x_j+\varepsilon) - \frac{x-x_j-\varepsilon}{\varphi_j(x_j-\varepsilon)} f(x_j),$$

since it is clear that $G(0) = 0$.

Now

$$\begin{aligned}G'(0) &= \frac{x-x_j}{\varphi_j(x_j)} f'(x_j) - \frac{x-x_j}{[\varphi_j(x_j)]^2} \varphi_j'(x_j) f(x_j) \\ &\quad + \frac{1}{\varphi_j(x_j)} f(x_j) - \frac{x-x_j}{[\varphi_j(x_j)]^2} \varphi_j'(x_j) f(x_j).\end{aligned}$$

In the notation of (6.14) we have

$$\begin{aligned}L_j(x) &= \frac{\varphi_j(x)}{\varphi_j(x_j)}, \\ L_j'(x) &= \frac{\varphi_j'(x)}{\varphi_j(x_j)}.\end{aligned}$$

A little algebraic simplification then shows that as $\varepsilon \to 0$ the terms involving $f(x_j)$ and $f(x_{j+n})$ tend to $H_j(x)f(x_j) + K_j(x)f'(x_j)$, which are the corresponding terms in the Hermite interpolation polynomial.

Solution to Exercise 6.10

For $f(x) = x^5$, since $f(0) = 0,\ f'(0) = 0$, there is no contribution to the interpolation polynomial from these two terms. In the notation of (6.14) we find that

$$\begin{aligned} L_1(x) &= \frac{x}{a}, \\ K_1(x) &= \frac{x^2}{a^2}(x-a), \\ H_1(x) &= \frac{x^2}{a^2}\left[1 - \frac{2}{a}(x-a)\right]. \end{aligned}$$

Hence

$$\begin{aligned} p_3(x) &= \frac{x^2}{a^2}(x-a)\,5a^4 + \frac{x^2}{a^2}\left[1 - \frac{2}{a}(x-a)\right]\,a^5 \\ &= 3a^2x^3 - 2a^3x^2. \end{aligned}$$

The error is therefore

$$\begin{aligned} x^5 - p_3(x) &= x^2(x^3 - 3a^2x + 2a^3) \\ &= x^2(x-a)^2(x+2a) \\ &= \frac{x^2(x-a)^2}{4!}\,5!\xi, \end{aligned}$$

where $\xi = (x+2a)/5$. This is the required result, since $f^{IV}(\xi) = 5!\xi$.

Solution to Exercise 6.11

Since $f(z)$ is holomorphic in D, the only poles of $g(z)$ are the points where the denominator vanishes, namely the point $z = x$ and the points $z = x_j,\ j = 0, \ldots, n$. Since they are distinct, they are all simple poles. Hence the residue at $z = x$ is

$$\begin{aligned} R_x &= \lim_{z \to x} (z - x) g(z) \\ &= \lim_{z \to x} f(z) \prod_{k=0}^{n} \frac{x - x_k}{z - x_k} \\ &= f(x). \end{aligned}$$

In the same way the residue at $z = x_j$ is

$$\begin{aligned} R_j &= \lim_{z \to x_j} (z - x_j) g(z) \\ &= \lim_{z \to x_j} (z - x_j) \frac{f(z)}{z - x} \prod_{k=0}^{n} \frac{x - x_k}{z - x_k} \\ &= -f(x_j) \prod_{k=0, k \neq j}^{n} \frac{x - x_k}{x_j - x_k} \\ &= -L_j(x) f(x_j). \end{aligned}$$

Hence the sum of the residues is just $f(x) - p_n(x)$. The Cauchy Residue Theorem then gives the required result.

The contour C consists of two semicircles of radius K, joined by two straight lines of length $b - a$. Hence the length of the contour is $2\pi K + 2(b - a)$. Hence

$$\begin{aligned} \left| \int_C g(z) \, \mathrm{d}z \right| &\leq \left| \int_C |g(z)| \, \mathrm{d}z \right| \\ &\leq [2\pi K + 2(b - a)] G, \end{aligned}$$

provided that $|g(z)| \leq G$ for all z on C. Now for all z on C we know that $|z - x| \leq K$ and $|z - x_j| \leq K$, and since x and all the x_j lie in $[a, b]$ we also know that $|x - x_j| \leq (b - a)$. Hence

$$|g(z)| \leq M \frac{(b - a)^{n+1}}{K^{n+1}}, \qquad z \in C.$$

Using the result of the Residue Theorem, this shows that

$$|f(x) - p_n(x)| < \frac{(b - a + \pi K) M}{\pi} \left(\frac{b - a}{K} \right)^{n+1},$$

as required. Since $K > b-a$, the right-hand side tends to zero as $n \to \infty$, which is the condition for p_n to converge uniformly to f, for $x \in [a, b]$.

The function $f(z)$ has poles at $z = \pm\imath$. For the interval $[a, b]$ we require that the corresponding contour C does not contain these poles; this requires that the distance from $\imath$ to any point on $[a, b]$ must be greater than $b - a$. For the symmetric interval $[-a, a]$ the closest point to $\imath$ is $x = 0$, with distance 1. Hence the length of the interval must be less than 1, and we require that $a < 1/2$.

This is the condition required by the above proof, so it is a sufficient condition, but may not be necessary. Evidently it is not satisfied by $[-5, 5]$.

Solution to Exercise 6.12

Expanding $f(h)$ and $f(-h)$ about the point $x = 0$, we get

$$\begin{aligned} f(h) &= f(0) + hf'(0) + \tfrac{1}{2}h^2 f''(0) + \tfrac{1}{6}h^3 f'''(\xi_1), \\ f(-h) &= f(0) - hf'(0) + \tfrac{1}{2}h^2 f''(0) - \tfrac{1}{6}h^3 f'''(\xi_2), \end{aligned}$$

where $\xi_1 \in (0, h)$ and $\xi_2 \in (-h, 0)$. Hence,

$$f(h) - f(-h) = 2hf'(0) + \tfrac{1}{6}h^3 \left(f'''(\xi_1) + f'''(\xi_2)\right) .$$

The continuity of f''' on $[-h, h]$ implies the existence of ξ in $(-h, h)$ such that $\frac{1}{2}(f'''(\xi_1) + f'''(\xi_2)) = f'''(\xi)$. Therefore,

$$\frac{f(h) - f(-h)}{2h} - f'(0) = \tfrac{1}{6}h^2 f'''(\xi),$$

and hence

$$E(h) = \tfrac{1}{6}h^2 f'''(\xi) + \frac{\varepsilon_+ - \varepsilon_-}{2h}.$$

Taking the absolute value of both sides, and bounding $|f'''(\xi)|$ by M_3 and $|\varepsilon_+|$ and $|\varepsilon_-|$ by ε, we have that

$$|E(h)| \leq \tfrac{1}{6}h^2 M_3 + \frac{\varepsilon}{h}.$$

Let as consider the right-hand side of this inequality as a function of $h > 0$; clearly, it is a positive and convex function which tends to $+\infty$ if $h \to +0$ or $h \to +\infty$.

Setting the derivative of $h \mapsto \frac{1}{6}h^2 M_3 + \frac{\varepsilon}{h}$ to zero yields,

$$\tfrac{1}{3}hM_3 - \frac{\varepsilon}{h^2} = 0,$$

and therefore,

$$h = \left(\frac{3\varepsilon}{M_3}\right)^{1/3}$$

gives the value of h for which the bound on $|E(h)|$ is minimized.

Solution to Exercise 7.1

The weight w_k is defined by

$$w_k = \int_a^b L_k(x)\mathrm{d}x = \int_a^b \prod_{j=1, j\neq k}^{n} \left(\frac{x - x_j}{x_k - x_j}\right) \mathrm{d}x.$$

Now if $x_k = a + kh$ then $x_{n-k} = a + (n-k)h = b - kh = a + b - x_k$. Making the change of variable $y = a + b - x$ we get

$$\begin{aligned} w_k &= -\int_b^a \prod_{j=1, j\neq k}^{n} \left(\frac{a + b - y - x_j}{x_k - x_j}\right) \mathrm{d}y \\ &= \int_a^b \prod_{j=1, j\neq k}^{n} \left(\frac{x_{n-j} - y}{x_{n-j} - x_{n-k}}\right) \mathrm{d}y \\ &= \int_a^b \prod_{i=1, i\neq n-k}^{n} \left(\frac{y - x_i}{x_{n-k} - x_i}\right) \mathrm{d}y \\ &= w_{n-k}, \end{aligned}$$

where we have replaced j by $n - i$ in the last product.

Solution to Exercise 7.2

The Newton–Cotes formula using $n+1$ points is exact for every polynomial of degree n. Now suppose that n is even and consider the polynomial

$$q_{n+1}(x) = [x - (a+b)/2]^{n+1}.$$

This is a polynomial of odd degree, and is antisymmetric about the midpoint of the interval $[a, b]$. Hence

$$\int_a^b q_{n+1}(x)\mathrm{d}x = 0.$$

The Newton–Cotes approximation to this integral is

$$\sum_{k=0}^{n} w_k \, q_{n+1}(x_k).$$

Now Exercise 1 has shown that $w_{n-k} = w_k$, and the antisymmetry of q_{n+1} shows that $q_{n+1}(x_k) = -q_{n+1}(x_{n-k})$. Hence the Newton–Cotes formula also gives zero, so it is exact for the polynomial q_{n+1}.

Any polynomial of degree $n+1$ may be written

$$p_{n+1} = c\, q_{n+1} + p_n,$$

where c is a constant and p_n is a polynomial of degree n. Hence the Newton–Cotes formula is also exact for p_{n+1}.

Solution to Exercise 7.3

It is clear that a quadrature formula is exact for all polynomials of degree n if, and only if, it is exact for the particular functions $f(x) = x^r,\ r = 0, \ldots, n$. The given formula is exact for the function x^r provided that

$$\int_{-1}^{1} x^r \, \mathrm{d}x = w_0(-\alpha)^r + w_1\alpha^r,$$

that is, if

$$\frac{1-(-1)^{r+1}}{r+1} = \alpha^r[(-1)^r w_0 + w_1].$$

To be exact for polynomials of degree 1 we must therefore satisfy this equation for $r = 0$ and $r = 1$. This gives

$$\begin{aligned} 2 &= w_0 + w_1 \\ 0 &= \alpha[-w_0 + w_1]. \end{aligned}$$

Since α is not zero, these give at once $w_0 = w_1 = 1$.

For the formula to be exact for all polynomials of degree 2 we also require to satisfy the equation with $r = 2$, giving

$$\frac{2}{3} = \alpha^2[w_0 + w_1] = 2\alpha^2.$$

As $\alpha > 0$ this requires the unique value $\alpha = 1/\sqrt{3}$.

The same equation with $r = 3$ gives

$$0 = \alpha^3[-w_0 + w_1]$$

which is satisfied, so when $\alpha = 1/\sqrt{3}$ the formula is exact for $x^r,\ r = 0, 1, 2, 3$. It is therefore exact for every polynomial of degree 3.

[Note that it is also exact for the function x^r for every odd integer r.]

Solution to Exercise 7.4

As in Exercise 3 the formula is exact for all polynomials of degree 3 if, and only if, it is exact for the functions $x^r,\ r = 0, 1, 2, 3$. This requires that

$$\begin{aligned} 2 &= w_0 + w_1 + w_2 + w_3 \\ 0 &= -w_0 - \tfrac{1}{3}w_1 + \tfrac{1}{3}w_2 + w_3 \\ \tfrac{2}{3} &= w_0 + \tfrac{1}{9}w_1 + \tfrac{1}{9}w_2 + w_3 \\ 0 &= -w_0 - \tfrac{1}{27}w_1 + \tfrac{1}{27}w_2 + w_3. \end{aligned}$$

From the 2nd and 4th of these equations it is easy to see that $w_0 = w_3$ and $w_1 = w_2$; this also follows from Exercise 1. The 1st and 3rd equations then become

$$\begin{aligned} 2 &= 2w_0 + 2w_1 \\ \tfrac{2}{3} &= 2w_0 + \tfrac{2}{9}w_1. \end{aligned}$$

This leads to $w_0 = w_3 = \frac{1}{4},\ \ w_1 = w_2 = \frac{3}{4}$.

Solution to Exercise 7.5

By symmetry the integral is zero for every odd power of x, and both formulae also give the result zero. Hence the difference is zero for the functions x, x^3, x^5. Both formulae give the correct result for all polynomials of degree 3; hence we only have to consider the functions x^4 and x^6

For x^4 the integral has the value $2/5$ and a simple calculation shows that the two formulae give the results (i) $2/3$ and (ii) $14/27$. Hence the errors are

$$x^4 : \qquad (i) - 4/15, \qquad -16/135.$$

Similarly for x^6 the integral is $2/7$, and the two formulae give the results (i) $2/3$ and (ii) $122/243$, giving for the errors

$$x^6 : \qquad (i) - 8/21, \qquad -368/1701.$$

Hence for the polynomial

$$p_5 = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + c_5 x^5$$

the respective errors are (i) $-\frac{4}{15}c_4$ and (ii) $-\frac{16}{135}c_4$, so the second is more accurate, by a factor $\frac{4}{9}$.

For the polynomial

$$p_6 = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + c_5 x^5 + c_6 x^6$$

the respective errors are (i)$-\frac{4}{15}c_4 - \frac{8}{21}c_6$ and (ii) $-\frac{16}{135}c_4 - \frac{368}{1701}$. Hence if we choose a polynomial for which $c_6 = -\frac{7}{10}c_4$ the error in Simpson's rule is zero, but the error in formula (ii) is not.

Solution to Exercise 7.6

Simple algebra shows that the errors in the approximation by the trapezium rule of the integrals $\int_0^1 x^4\,\mathrm{d}x$ and $\int_0^1 x^5\,\mathrm{d}x$ are $-3/10$ and $-1/3$ respectively.

Similarly the errors in the approximation by Simpson's Rule of the same integrals are $-1/120$ and $-1/48$ respectively.

Hence the errors in the approximation of

$$\int_0^1 (x^5 - Cx^4)\,\mathrm{d}x$$

by the trapezium rule and Simpson's rule are

$$\begin{array}{ll} \frac{3}{10}C - \frac{1}{3} & \text{(trapezium rule)} \\ \frac{1}{120}C - \frac{1}{48} & \text{(Simpson's rule).} \end{array}$$

The trapezium rule gives the correct value of this integral when $C = 10/9$.

Moreover the trapezium rule gives a more accurate result than Simpson's rule for this integral when

$$|\tfrac{3}{10}C - \tfrac{1}{3}| < |\tfrac{1}{120}C - \tfrac{1}{48}|.$$

A sketch of the graphs of the two functions of C on the left and right of this inequality shows that it is satisfied when C lies between the two extreme values which are the solutions of

$$\begin{aligned} \tfrac{3}{10}C - \tfrac{1}{3} &= \tfrac{1}{120}C - \tfrac{1}{48} \\ \tfrac{3}{10}C - \tfrac{1}{3} &= -[\tfrac{1}{120}C - \tfrac{1}{48}]. \end{aligned}$$

These values are $\frac{15}{14}$ and $\frac{85}{74}$ as required.

Solution to Exercise 7.7

To determine c_{-1}, c_0, c_1 and c_2, we demand that the quadrature rule integrates 1, x, x^2 and x^3 exactly, for, then it will integrate any polynomial from $\mathcal{P}_3$ exactly. Hence,

$$\begin{aligned} c_{-1} + c_0 + c_1 + c_2 &= 1 \\ -c_{-1} + c_1 + 2c_2 &= \tfrac{1}{2} \\ c_{-1} + c_1 + 4c_2 &= \tfrac{1}{3} \\ -c_{-1} + c_1 + 8c_2 &= \tfrac{1}{4}\,. \end{aligned}$$

Solving this linear system yields

$$c_{-1} = c_2 = -\tfrac{1}{24}\,, \qquad c_0 = c_1 = \tfrac{13}{24}\,.$$

Suppose that f and its derivatives up to and including order 4 are defined and continuous on the closed interval $[-1, 2]$ which includes the interval of integration, $[0, 1]$, as well as all the quadrature points, $-1, 0, 1, 2$. Consider

$$E(f) = \int_0^1 f(x)\mathrm{d}x - Q(f) = \int_0^1 [f(x) - p_3(x)]\mathrm{d}x\,,$$

where p_3 is the Lagrange interpolation polynomial of f of degree 3 on the interval $[-1, 2]$ with interpolation points $-1, 0, 1, 2$. Hence, by the remainder theorem for Lagrange interpolation,

$$|E(f)| \leq \frac{M_4}{(3+1)!} \int_0^1 |\pi_4(x)|\mathrm{d}x\,,$$

where $\pi_4(x) = (x+1)x(x-1)(x-2)$. Now, $|\pi_4(x)| = (1-x^2)x(2-x)$ for all $x \in [0, 1]$, and therefore,

$$\int_0^1 |\pi_4(x)|\mathrm{d}x = \tfrac{11}{30}\,.$$

Thus,

$$|E(f)| \leq \tfrac{11}{720} M_4\,,$$

where $M_4 = \max_{x\in[-1,2]} |f(x)|$.

Solution to Exercise 7.8

From the definition we see that

$$\begin{aligned} T(m) &= \frac{b-a}{m}[\tfrac{1}{2}f_0 + f_2 + f_4 \ldots] \\ T(2m) &= \frac{b-a}{2m}[\tfrac{1}{2}f_0 + f_1 + f_2 + \ldots], \end{aligned}$$

where we use the notation

$$f_j = f(a + j(b-a)/2m).$$

Hence

$$\begin{aligned} \tfrac{4}{3}T(2m) - \tfrac{1}{3}T(m) &= \frac{b-a}{6m}[2f_0 + 4f_1 + 4f_2 + 4f_3 + 4f_4 + \ldots \\ &\qquad -f_0 - 2f_2 - 2f_4 - \ldots] \\ &= \frac{b-a}{6m}[f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_2 + \ldots], \end{aligned}$$

which agrees with $S(2m)$.

Solution to Exercise 7.9

From Theorem 7.4 it follows that

$$\int_a^b f(x)\mathrm{d}x - T(m) = A/m^2 + E(m)$$

where $m^2E(m) \to 0$. Hence

$$\begin{aligned}
\frac{T(m) - T(2m)}{T(2m) - T(4m)} &= \frac{-A/m^2 + A/4m^2 - E(m) + E(2m)}{-A/4m^2 + A/16m^2 - E(2m) + E(4m)} \\
&= \frac{-12A - 16m^2[E(m) - E(2m)]}{-3A - 16m^2[E(2m) - E(4m)]} \\
&\to 4.
\end{aligned}$$

Table 7.3 for the values $m = 4, 8, 16$ gives this ratio the values $3.72, 3.91, 3.97$ respectively; these indicate satisfactory convergence to limiting value 4.

Solution to Exercise 7.10

In the same way as in Exercise 9 we get

$$\begin{aligned}
\frac{T(m)-T(2m)}{T(2m)-T(4m)} &= \frac{-A/m^\alpha + A/(2m)^\alpha - E(m) + E(2m)}{-A/(2m)^\alpha + A/(4m)^\alpha - E(2m) + E(4m)} \\
&= \frac{A(-4^\alpha + 2^\alpha) - (4m)^\alpha [E(m) - E(2m)]}{A(-2^\alpha + 1) - (4m)^\alpha [E(2m) - E(4m)]} \\
&\to 2^\alpha.
\end{aligned}$$

Table 7.4 for the values $m = 4, 8, 16$ gives this ratio the values $2.47, 2.49, 2.50$ respectively; these are consistent with a value of α such that $2^\alpha = 2.5$, or $\alpha = \ln 2.5 / \ln 2$. The value $\alpha = 4/3$ fits this data quite well.

Solution to Exercise 7.11

With the notation of Section 6.3

$$\begin{aligned} L_0(x) &= \tfrac{1}{2}(1-x) \\ L_1(x) &= \tfrac{1}{2}(x+1) \\ H_0(x) &= \tfrac{1}{4}(1-x)^2[1+(x+1)] \\ H_1(x) &= \tfrac{1}{4}(x+1)^2[1-(x-1)] \\ K_0(x) &= \tfrac{1}{4}(1-x)^2(x+1) \\ K_1(x) &= \tfrac{1}{4}(x+1)^2(x-1). \end{aligned}$$

The required polynomial is then

$$p_3(x) = H_0(x)f(-1) + H_1(x)f(1) + K_0(x)f'(-1) + K_1(x)f'(1),$$

and

$$f(x) - p_3(x) = \frac{(x+1)^2(x-1)^2}{24} f^{IV}(\xi)$$

for some $\xi \in (-1,1)$.

Integration of this equation gives the required result, with

$$\begin{aligned} |E| &= \tfrac{1}{24}\left|\int_{-1}^{1} (x^2-1)^2 f^{IV}(\xi)\,\mathrm{d}x\right| \\ &\leq \tfrac{1}{24} M_4 \int_{-1}^{1} (x^2-1)^2\,\mathrm{d}x \\ &= \tfrac{2}{45} M_4, \end{aligned}$$

where

$$M_4 = \max_{x\in(-1,1)} |f^{IV}(x)|.$$

Solution to Exercise 7.12

We have seen that

$$\begin{aligned} q_1(t) &= -t \\ q_2(t) &= -\tfrac{1}{2}t^2 + \tfrac{1}{6} \\ q_3(t) &= -\tfrac{1}{6}t^3 + \tfrac{1}{6}t. \end{aligned}$$

By integration it follows that

$$\begin{aligned} q_4(t) &= -\tfrac{1}{24}t^4 + \tfrac{1}{12}t^2 + A_4 \\ q_5(t) &= -\tfrac{1}{120}t^5 + \tfrac{1}{36}t^3 + A_4 t + A_5. \end{aligned}$$

Now $q_5(t)$ must be an odd function, so $A_5 = 0$, and $q_5(1) = 0$, so that

$$A_4 = \tfrac{1}{120} - \tfrac{1}{36} = -\tfrac{7}{360}.$$

Hence

$$\begin{aligned} q_4(t) &= -\tfrac{1}{24}t^4 + \tfrac{1}{12}t^2 - \tfrac{7}{360} \\ q_5(t) &= -\tfrac{1}{120}t^5 + \tfrac{1}{36}t^3 - \tfrac{7}{360}t. \end{aligned}$$

In the same way

$$\begin{aligned} q_6(t) &= -\tfrac{1}{720}t^6 + \tfrac{1}{144}t^4 - \tfrac{7}{720}t^2 + A_6 \\ q_7(t) &= -\tfrac{1}{5040}t^7 + \tfrac{1}{720}t^5 - \tfrac{7}{2160}t^3 + A_6 t + A_7. \end{aligned}$$

Since q_7 is an odd function, $A_7 = 0$; since $q_7(1) = 0$

$$A_6 = \tfrac{1}{5040} - \tfrac{1}{720} + \tfrac{7}{2160} = \tfrac{31}{15120}.$$

Hence

$$\begin{aligned} q_6(t) &= -\tfrac{1}{720}t^6 + \tfrac{1}{144}t^4 - \tfrac{7}{720}t^2 + \tfrac{31}{15120} \\ q_7(t) &= -\tfrac{1}{5040}t^7 + \tfrac{1}{720}t^5 - \tfrac{7}{2160}t^3 + \tfrac{31}{15120}t. \end{aligned}$$

For the coefficients c_r we obtain from the definition

$$\begin{aligned} c_1 &= q_2(1)/2^2 = -\tfrac{1}{12} \\ c_2 &= q_4(1)/2^4 = \tfrac{1}{720} \\ c_3 &= q_6(1)/2^6 = -\tfrac{1}{30240}. \end{aligned}$$

Solution to Exercise 7.13

Since $\sin x$ is an odd function, $\int_{-\pi}^{\pi} \sin rx \mathrm{d}x = 0$ for all values of r. For the same reason the composite trapezium rule will also give the value zero, so the rule gives the correct value of the integral for every value of r, whether or not it is a multiple of m.

When $r = 0$, $\cos rx \equiv 1$, so the composite trapezium rule gives the exact result in this case also.

Now suppose that r is a positive integer; then the composite trapezium rule for $\int_{-\pi}^{\pi} \cos rx \mathrm{d}x$ becomes

$$T_m = h[\tfrac{1}{2}\cos(-r\pi) + \sum_{j=1}^{m-1} \cos(-r\pi + jr\theta) + \tfrac{1}{2}\cos r\pi],$$

where $\theta = 2\pi/m$. Since the first and last terms are equal, this can be written, using the given relations,

$$\begin{aligned} T_m &= h\sum_{j=1}^{m} \cos(-r\pi + jr\theta) \\ &= (-1)^r h \sum_{j=1}^{m} \cos jr\theta \\ &= (-1)^r h[\sin(m+\tfrac{1}{2})r\theta - \sin\tfrac{1}{2}r\theta]/\sin\tfrac{1}{2}r\theta \\ &= 0, \end{aligned}$$

since $mr\theta = 2r\pi$. Hence, in general, the composite trapezium rule gives the result zero, which is the correct value of the integral.

However, the argument breaks down when $\sin\frac{1}{2}r\theta = 0$, which occurs when r is a multiple of m. Now if $r = km$ then

$$\cos(-r\pi + jr\theta) = \cos(-km\pi + 2jr\pi/m) = \cos(-r\pi + 2jk\pi) = (-1)^r.$$

Thus all the terms in the sum are equal, and the result of the composite trapezium rule is

$$T_m = (-1)^r mh = (-1)^r 2\pi.$$

[Note that if $m = 1$ the result is $(-1)^r 2\pi$ for every positive value of r, since r is always a multiple of 1.]

Solution Exercise 8.1

(i) Write $M = \|f\|_\infty$, so that

$$|f(x)| \leq M, \quad \text{for } a \leq x \leq b.$$

Then

$$\begin{aligned} \{\|f\|_2\}^2 &= \int_a^b w(x)[f(x)]^2 \mathrm{d}x \\ &\leq \int_a^b w(x) M^2 \mathrm{d}x \\ &= M^2 C^2. \end{aligned}$$

Taking square roots gives the required result.

(ii) Many examples are possible; here are two, where for simplicity we have assumed that the weight function $w(x) \equiv 1$:

(iia) Define

$$f(x) = \frac{2M}{1 + k^2(x-c)^2},$$

where

$$c = \tfrac{1}{2}(a+b),$$

and k is a constant to be determined. Then clearly $\|f\|_\infty = 2M > M$, as required. For the $2-$norm we find that

$$\begin{aligned} \{\|f\|_\infty\}^2 &= \int_a^b \frac{4M^2}{1 + k^2(x-c)^2} \mathrm{d}x \\ &= \frac{4M^2}{k} \left\{ \tan^{-1} \tfrac{1}{2}k(b-a) - \tan^{-1} \tfrac{1}{2}k(a-b) \right\} \\ &\leq \frac{4M^2\pi}{k}, \end{aligned}$$

and we have the required property if

$$k = \frac{8M^2\pi}{\varepsilon}.$$

(iib) Another function with the required properties is

$$f(x) = \begin{cases} \frac{2M}{\delta}(a + \delta - x), & a \leq x \leq x + \delta \\ 0, & a + \delta \leq x \leq b\,. \end{cases}$$

Then, as before, $\|f\|_\infty = 2M > M$. Now

$$\begin{aligned} \{\|f\|_\infty\}^2 &= \frac{4M^2}{\delta^2}\int_a^{a+\delta}(a+\delta-x)^2\mathrm{d}x \\ &= \frac{4M^2\delta}{3} \end{aligned}$$

and we have the required property if $\delta = 3\varepsilon/8M^2$.

Solution of Exercise 8.2

Suppose that $\hat{p}_n$ is the minimax polynomial of degree n for the even function f on $[-a, a]$. Then there is a sequence of $n+2$ points x_j at which

$$f(x_j) - \hat{p}_n(x_j) = (-1)^j E, \quad j = 1, \ldots, n+2$$

where $|E| = \|f - \hat{p}_n\|_\infty$. Now define the polynomial q_n by

$$q_n(x) = \hat{p}_n(-x);$$

then

$$f(-x_j) - q_n(-x_j) = (-1)^j E, \quad j = 1, \ldots, n+2$$

since f is an even function. Also

$$\begin{aligned} \|f - q_n\|_\infty &= \max |f(x) - q_n(x)| \\ &= \max |f(-x) - \hat{p}_n(-x)| \\ &= \|f - \hat{p}_n\|_\infty, \end{aligned}$$

since the interval $[-a, a]$ is symmetric. Thus the points

$$(-x_j), \quad j = n+2, n+1, \ldots, 1$$

form a sequence of critical points for the approximation q_n, and q_n is therefore a minimax polynomial for f. But the minimax approximation is unique, so that $q_n \equiv \hat{p}_n$, and $\hat{p}_n$ is an even polynomial.

This means in particular that the minimax polynomial of odd degree $2n+1$ is an even polynomial, and the coefficient of x^{2n+1} is zero. It is therefore identical to $\hat{p}_{2n}$, which is the minimax polynomial of degree $2n+1$, as well as being the minimax polynomial of degree $2n$. The minimax polynomial $\hat{p}_{2n}$ has a sequence of $2n+3$ critical points.

The proof has also shown that the critical points are symmetrical in the interval $[-a, a]$; if x_j is a critical point, so is $-x_j$.

Solution to Exercise 8.3

The minimax polynomial of degree n is an odd function; the minimax polynomial approximation of degree $2n+1$ is therefore also the minimax approximation of degree $2n+2$.

The proof follows the same lines as that of Exercise 2. With the same notation,

$$q_n(x) = \hat{p}_n(-x);$$

then

$$-f(-x_j) - q_n(-x_j) = (-1)^j E, \quad j = 1, \ldots, n+2$$

since f is an odd function. Thus the points

$$(-x_j), \quad j = n+2, n+1, \ldots, 1$$

form a sequence of critical points for the approximation $-q_n$ to the function $f(x)$, and $-q_n$ is therefore a minimax polynomial for f. But the minimax approximation is unique, so that $-q_n \equiv \hat{p}_n$, and $\hat{p}_n$ is an odd polynomial.

This means in particular that the minimax polynomial of even degree $2n+2$ is an odd polynomial, and the coefficient of x^{2n+2} is zero. It is therefore identical to $\hat{p}_{2n+1}$, which is the minimax polynomial of degree $2n+2$, as well as being the minimax polynomial of degree $2n+1$. The minimax polynomial $\hat{p}_{2n+1}$ has a sequence of $2n+4$ critical points.

The proof has also shown that the critical points are symmetrical in the interval $[-a, a]$; if x_j is a critical point, so is $-x_j$.

Solution of Exercise 8.4

(i) Since g is an odd function and the interval is symmetric, the approximation will be an odd polynomial, so take

$$p_2(x) = c_1 x.$$

There will be four alternate maxima and minima, symmetric in the interval, including the points ± 1. The other two will be internal extrema, which we take to be $\pm\theta$. The condition that these two points are extrema of the error gives

$$c_1 - \cos\theta = 0,$$

and the condition that the magnitudes of the extrema are equal gives

$$\begin{aligned} c_1 - \sin 1 &= E \\ c_1\theta - \sin\theta &= -E. \end{aligned}$$

From these equations it is easy to deduce that

$$(1+\theta)\cos\theta - \sin 1 - \sin\theta = 0.$$

This equation for θ has exactly one root in (0,1); it is easy to see that the left-hand side is monotonic decreasing in (0,1). Having determined this value, we see at once that $c_1 = \cos\theta$ and $E = \cos\theta - \sin 1$. The numerical values are $\theta = 0.4937$, $c_1 = 0.8806$ and $E = 0.0391$, giving

$$p_2(x) = 0.8806x, \quad \|g - p_2\|_\infty = 0.0391.$$

(ii) Since h is an even function and the interval is symmetric, the approximation will be an even polynomial, so take

$$p_3(x) = c_0 + c_2 x^2.$$

There will be five alternate maxima and minima, symmetrical in the interval, including the points -1, 0 and 1. The other two will be internal extrema, which we take at $\pm\alpha$. The condition that these points are extrema gives

$$2c_2\alpha + 2\alpha\sin(\alpha^2) = 0,$$

and the condition that the magnitudes of the maxima and minima are all equal gives

$$\begin{aligned} c_0 + c_2 - \cos 1 &= E, \\ c_0 + c_2\alpha^2 - \cos(\alpha^2) &= -E \\ c_0 - 1 &= E. \end{aligned}$$

From these we obtain in succession

$$\begin{aligned} c_2 &= \cos 1 - 1, \\ \sin(\alpha^2) &= 1 - \cos 1, \\ E &= \tfrac{1}{2}\{c_2(1-\alpha^2) + \cos(\alpha^2) - \cos 1\}, \\ c_0 &= 1 + E. \end{aligned}$$

The numerical values are:

$$p_3(x) = 1.0538 - 0.4597x^2, \quad \|\cos(x^2) - p_3\|_\infty = E = 0.0538,$$

with $\alpha = 0.6911$.

Solution to Exercise 8.5

Since p_n is a polynomial, it is a continuous function. Suppose that $p_n(0) = A$. In the interval $[-\eta, \eta]$, where η is an arbitrarily small number, there will be points x at which $H(x) = 1$, and points at which $H(x) = -1$. At such points $|H(x) - p_n(x)|$ will be arbitrarily close to $|A - 1|$ and $|A + 1|$ respectively. Hence $\|H - p_n\|_\infty$ cannot be smaller than $|A - 1|$ or $|A + 1|$. If A is nonzero, one of these quantities will be greater than 1, and if $A = 0$ we shall have $\|f - p_n\|_\infty \geq 1$.

By the previous argument the polynomial of degree zero of best approximation to $H(x)$ on $[-1, 1]$ must have $p_0(0) = 0$. It is therefore the zero polynomial, $p_0(x) = 0$, and is unique.

The polynomial of best approximation, of degree 1, must also have $p_1(0) = 0$, so must have the form $p_1(x) = c_1 x$. Then the difference $e_1(x) = H(x) - c_1 x$ is antisymmetric about $x = 0$, so that $e(-x) = -e(x)$. The difference attains its extreme values at 0 and at ± 1; $|e(1)| = |1 - c_1|$, and $|e(x)|$ takes values arbitrarily close to 1 close to $x = 0$. Hence $\|H - p_1\|_\infty = 1$ provided that $|1 - c_1| \leq 1$; the polynomial of best approximation of degree 1 is not unique. Any polynomial $p_1(x) = c_1 x$, with $0 \leq c_1 \leq 2$, is a polynomial of best approximation.

Solution to Exercise 8.6

It is evidently very easy to construct a function f which is zero at each of the points $t_i,\ i = 1, \ldots, k$, but is not zero everywhere; any polynomial which has these points as zeros, for example, is such a function. For such a function $Z_k(f) = 0$, but f is not identically zero. Hence $Z_k(.)$ does not satisfy the first of the axioms for a norm. However, if p_n is a polynomial of degree $n < k$, then if $Z_k(p_n) = 0$ the polynomial p_n must vanish at the k points t_j, so it must vanish identically. Thus $Z_k(.)$ is a norm on the space of polynomials of degree n, if $k > n$; it is easy to see that the other axioms hold.

Suppose that the polynomial q_1 satisfies the given conditions

$$f(0) - q_1(0) = -[f(\tfrac{1}{2}) - q_1](\tfrac{1}{2}) = f(1) - q_1(1).$$

If q^* is a polynomial of degree 1 which gives a smaller value for Z, then $Z_3(f - q^*) < Z_3(f - q_1)$; hence $q^*(x) - q_1(x)$ must be negative at t_1 and t_3, and positive at t_2. This is impossible if q^* is a polynomial of degree 1. Hence q_1 is the polynomial which minimises $Z(f - p_1)$ over all polynomials of degree 1.

Writing $q_1(x) = \alpha + \beta x$ the conditions give

$$\begin{aligned} 1 - \alpha &= E \\ \mathrm{e}^{1/2} - \alpha - \tfrac{1}{2}\beta &= -E \\ \mathrm{e} - \alpha - \beta &= E. \end{aligned}$$

These easily lead to $\beta = \mathrm{e} - 1$, $\alpha = \frac{3}{4} + \frac{1}{2}\sqrt{e} - \frac{1}{2}\mathrm{e}$. The error is

$$Z_3(f - q_1) = E = \tfrac{1}{4} - \tfrac{1}{2}\sqrt{e} + \tfrac{1}{2}\mathrm{e}.$$

A straightforward, but tedious, approach to the case $k = 4$ is to solve each of the four problems obtained by choosing three out of these four points. In each case, having constructed the polynomial approximation p_1, evaluate $Z_4(f - p_1)$, and the required approximation is the one which gives the least value to this quantity.

Alternatively, choose three of the four points and construct the polynomial p_1 which minimises $Z_3(f - p_1)$. Evaluate $|f(t^*) - p_1(t^*)|$ at the point t^* which was omitted. If this value does not exceed $Z_3(f - p_1)$, then p_1 is the required approximation. If it does exceed $Z_3(f - p_1)$ replace one of the points chosen, for which $f(t_j) - p_1(t_j)$ has the same sign as $f(t^*) - p_1(t^*)$, by t^*, and repeat the process. Continue this repeated choice of three of the four points until the required polynomial is reached.

Solution to Exercise 8.7

Since A is a fixed nonzero constant, the problem of finding the polynomial of best approximation to $f(x) \equiv 0$ by polynomials of the given form is equivalent to choosing the coefficients to minimise

$$\|0 - p_n\|_\infty,$$

which is the same as requiring to minimise

$$\|x^n - (1/A)q_{n-1}\|_\infty.$$

This is just the problem of approximating the function x^n by a polynomial of lower degree; hence we should choose the polynomial $p_n(x)$ so that

$$x^n - (1/A)\sum_{k=0}^{n-1} a_k x^k = 2^{-n-1}T_n(x),$$

where T_n is the Chebyshev polynomial of degree n. The formal result is

$$p_n(x) = \frac{A}{2^{n+1}}T_n(x),$$

Solution to Exercise 8.8

Clearly,

$$f(x) = a_{n+1}\left[x^{n+1} - \sum_{k=0}^{n} \frac{-a_k}{a_{n+1}} x^k\right] .$$

We seek the minimax polynomial $p_n \in \mathcal{P}_n$ for f on the interval $[-1, 1]$ in the form

$$p_n(x) = \sum_{k=0}^{n} b_k x^k .$$

Thus,

$$f(x) - p_n(x) = a_{n+1}\left[x^{n+1} - \sum_{k=0}^{n} \frac{b_k - a_k}{a_{n+1}} x^k\right] .$$

According to Theorem 8.6, the $\|\cdot\|_\infty$ norm of the right-hand side is smallest when

$$\sum_{k=0}^{n} \frac{b_k - a_k}{a_{n+1}} x^k = x^{n+1} - 2^{-n} T_{n+1}(x) .$$

Therefore, the required minimax polynomial for f is

$$p_n(x) = f(x) - a_{n+1} 2^{-n} T_{n+1}(x) .$$

Solution to Exercise 8.9

The minimax polynomial must be such that $f(x) - p_1(x)$ has three alternating extrema in $[-2, 1]$. Since f is convex, two of these extrema are at the ends -2 and 1, and the other must clearly be at 0. Graphically, the line p_1 must be parallel to the chord joining $(-2, f(-2))$ and $(1, f(1))$. Thus

$$p_1(x) = c_0 - \tfrac{1}{3}x.$$

The alternating extrema are than

$$\begin{aligned} f(-2) - p_1(-2) &= 2 - (c_0 + \tfrac{2}{3}) = \tfrac{4}{3} - c_0 \\ f(0) - p_1(0) &= 0 - (c_0) = -c_0 \\ f(1) - p_1(1) &= 1 - (c_0 - \tfrac{1}{3}) = \tfrac{4}{3} - c_0. \end{aligned}$$

These have the same magnitude if

$$\tfrac{4}{3} - c_0 = -(-c_0),$$

so that the minimax polynomial is

$$p_1(x) = \tfrac{2}{3} - \tfrac{1}{3}x,$$

and $\|f - p_1\|_\infty = \tfrac{2}{3}$.

Solution to Exercise 8.10

A standard trigonometric relation gives

$$\cos n\theta = \tfrac{1}{2}[\cos(n+1)\theta + \cos(n-1)\theta];$$

Writing $x = \cos\theta$, $T_n(x) = \cos n\theta$, gives (a).

Suppose that (b) is true for polynomials of degree up to and including n. Then it follows from (a) that T_{n+1} is a polynomial of degree $n+1$ with leading coefficient 2^n; hence (b) follows by induction, since $T_0 = 1$.

Evidently $T_0(x) = 1$ is an even function and $T_1(x) = x$ is an odd function. Then (c) follows by induction, using (a).

The zeros of $T_n(x)$ are $x_j = \cos\theta_j$, where θ_j is a zero of $\cos\theta$. Evidently

$$\cos n\frac{(j-\frac{1}{2})\pi}{n} = \cos(j-\tfrac{1}{2})\pi = 0,$$

giving (d); these values of x_j are distinct, and lie in $(-1,1)$.

Part(e) is obvious, since $|T_n(x)| = |\cos n\theta| \leq 1$, provided that $|x| \leq 1$ to ensure that θ is real.

Part (f) follows from the fact that $\cos n\theta = \pm 1$ when $\theta = k\pi/n$, $k = 0, \ldots, n$.

Solution to Exercise 8.11

There are many possible examples, most easily when f is an oscillatory function. For example, consider

$$f(x) = \sin \frac{3\pi(x-a)}{b-a}.$$

This function attains its maxima and minima, all of magnitude 1, at the points

$$x_j = a + \frac{j}{6}(b-a), \quad j = 1, 3, 5.$$

Hence the polynomial p_1 of degree 1 of best approximation to f on $[a, b]$ is $p_1(x) \equiv 0$, and none of the three critical points is equal to a or b.

Solution to Exercise 8.12

By expanding the binomial we see that

$$(1-x+tx)^n = \sum_{k=0}^{n} p_{nk}(x)t^k.$$

Substituting $t=1$ shows at once that

$$1 = \sum_{k=0}^{n} p_{nk}(x).$$

Differentiate with respect to t, giving

$$nx(1-x+tx)^{n-1} = \sum_{k=0}^{n} p_{nk}(x)k\,t^{k-1};$$

substituting $t=1$ then gives

$$nx = \sum_{k=0}^{n} kp_{nk}(x).$$

Differentiate again with respect to t, giving

$$n(n-1)x^2(1-x+tx)^{n-2} = \sum_{k=0}^{n} p_{nk}(x)k(k-1)\,t^{k-2};$$

substituting $t=1$ then gives

$$n(n-1)x^2 = \sum_{k=0}^{n} k(k-1)p_{nk}(x).$$

Now

$$\begin{aligned}\sum_{k=0}^{n}(x-k/n)^2 p_{nk}(x) &= x^2\sum_{k=0}^{n} p_{nk}(x) - (2x/n)\sum_{k=0}^{n} kp_{nk}(x)\\ &\quad +(1/n^2)\sum_{k=0}^{n} k^2 p_{nk}(x)\\ &= x^2 - (2x/n)nx + (1/n^2)[n(n-1)x^2+nx]\\ &= \frac{x(1-x)}{n}.\end{aligned}$$

From the definition of $p_n(x)$, and using the facts that each $p_{nk}(x)\geq 0$ when $0\leq x\leq 1$ and $\sum_{k=0}^{n} p_{nk}(x)=1$, we find

$$f(x)-p_n(x) = \sum_{k=0}^{n}[f(x)-f(k/n)]p_{nk}(x),$$

and

$$|f(x) - p_n(x)| \leq \sum_{k=0}^{n} |f(x) - f(k/n)| p_{nk}(x).$$

Then

$$\begin{aligned} \sum_{1} |f(x) - f(k/n)| p_{nk}(x) &< \sum_{1} (\varepsilon/2) p_{nk}(x) \\ &= (\varepsilon/2) \sum_{k=0}^{n} p_{nk}(x) \\ &= \varepsilon/2. \end{aligned}$$

For the other sum

$$\begin{aligned} \sum_{1} |f(x) - f(k/n)| p_{nk}(x) &\leq 2M \sum_{2} p_{nk}(x) \\ &\leq 2M/\delta^2 \sum_{2} (x - k/n)^2 p_{nk}(x) \\ &\leq 2M/\delta^2 \sum_{k=0}^{n} (x - k/n)^2 p_{nk}(x) \\ &= 2M/\delta^2 \frac{x(1-x)}{n} \\ &\leq 2M/\delta^2 \frac{1}{4n} \\ &= \frac{M}{2\delta^2 n}, \end{aligned}$$

since all the terms in the sums are non-negative, and $0 \leq x(1-x) \leq 1/4$. Thus if we choose $N_0 = M/\delta^2\varepsilon$ we obtain

$$\sum_{2} |f(x) - f(k/n)| p_{nk}(x) \leq \varepsilon/2.$$

Finally adding together these two sums we obtain

$$|f(x) - p_n(x)| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

when $n \geq N_0$. Since the value of N_0 does not depend on the particular value of x chosen, this inequality holds for all x in $[0, 1]$.

Solution to Exercise 9.1

We start with $\varphi_0(x) = 1$, and use the Gram-Schmidt process. It is useful to note that

$$\int_0^1 -\ln(x)\, x^k \mathrm{d}x = 1/(k+1)^2.$$

Writing

$$\varphi_1(x) = x - a_0\varphi_0(x)$$

we need

$$\begin{aligned} a_0 &= \frac{\int_0^1 -\ln(x)\, x\, \varphi_0(x)\mathrm{d}x}{\int_0^1 -\ln(x)[\varphi_0(x)]^2\mathrm{d}x} \\ &= -\tfrac{1}{4} \end{aligned}$$

and so

$$\varphi_1(x) = x - 1/4.$$

Now writing

$$\varphi_2(x) = x^2 - b_0\varphi_0(x) - b_1\varphi_1(x)$$

we find in the same way that

$$\begin{aligned} b_0 &= 1/9 \\ b_1 &= 5/7. \end{aligned}$$

Hence

$$\begin{aligned} \varphi_2(x) &= x^2 - (1/9) - (5/7)(x - 1/4) \\ &= x^2 - (5/7)x + (17/252). \end{aligned}$$

Solution to Exercise 9.2

Since the polynomials $\varphi_j(x)$ are orthogonal on the interval $[-1,1]$ we know that

$$\int_{-1}^{1} \varphi_i(t)\,\varphi_j(t)\,\mathrm{d}t = 0, \qquad i \neq j.$$

In this integral we make the change of variable

$$t = (2x-a-b)/(b-a), \qquad x = \tfrac{1}{2}[(b-a)t + a + b],$$

and it becomes

$$\int_a^b \varphi_i((2x-a-b)/(b-a))\,\varphi_j((2x-a-b)/(b-a))\,\frac{2}{b-a}\mathrm{d}x.$$

This shows that the new polynomials form an orthogonal system on the interval $[a,b]$.

From the Legendre polynomials

$$\begin{aligned} \varphi_0(t) &= 1 \\ \varphi_1(t) &= t \\ \varphi_2(t) &= t^2 - 1/3 \end{aligned}$$

we write $t = 2x-1$ and get the orthogonal polynomials on $[0,1]$ in the form

$$\begin{aligned} \varphi_0(x) &= 1 \\ \varphi_1(x) &= 2x-1 \\ \varphi_2(x) &= (2x-1)^2 - 1/3 \\ &= 4x^2 - 4x + 2/3 \end{aligned}$$

The different normalisation from the polynomials in Example 9.5 is unimportant.

Solution to Exercise 9.3

We are given that

$$\int_0^1 x^\alpha \, \varphi_i(x) \varphi_j(x) \, \mathrm{d}x = 0, \qquad i \neq j.$$

Making the change of variable $x = t/b$ this becomes

$$\int_0^b t^\alpha \, \varphi_i(t/b) \, \varphi_j(t/b) \, \mathrm{d}x / b^{\alpha+1} = 0, \qquad i \neq j.$$

This shows that the polynomials $\psi_j(x)$ defined by

$$\psi_j(x) = \varphi_j(x/b), \qquad j = 0, 1, \ldots$$

are orthogonal over the interval $[0, b]$ with the weight function $w(x) = x^\alpha$.

Solution to Exercise 9.4

Suppose that the required result is true for some value of k, with $0 \leq k < n-1$. Then

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^k (1-x^2)^n = (1-x^2)^{n-k} q_k(n).$$

Differentiation then gives

$$\begin{aligned}
\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^{k+1} (1-x^2)^n &= \frac{\mathrm{d}}{\mathrm{d}x}(1-x^2)^{n-k} q_k(x) \\
&= -2(n-k)x(1-x^2)^{n-k-1} q_k(x) \\
&\qquad +(1-x^2)^{n-k} q_k'(x) \\
&= (1-x^2)^{n-k-1}[-2(n-k)x q_k(x) + (1-x^2) q_k'(x)] \\
&= (1-x^2)^{n-k-1} q_{k+1}(x),
\end{aligned}$$

which is of the required form with k replaced by $k+1$. Since the result is trivially true for $k=0$ it is thus true for all k such that $0 \leq k < n$.

This means that every derivative of order less than n of the function $(1-x^2)^n$ has the term $(1-x^2)$ as a factor, and therefore vanishes at $x = \pm 1$.

Write D for $\mathrm{d}/\mathrm{d}x$, and suppose that $0 \leq i < j$. Then by integrating by parts

$$\begin{aligned}
\int_{-1}^{1} \varphi_i(x)\varphi_j(x)\mathrm{d}x &= \int_{-1}^{1} D^i(1-x^2)^i \, D^j(1-x^2)^j \, \mathrm{d}x \\
&= \left[D^i(1-x^2)^i \, D^{j-1}(1-x^2)^j\right]_{-1}^{1} \\
&\qquad - \int_{-1}^{1} D^{i+1}(1-x^2)^i \, D^{j-1}(1-x^2)^n \, \mathrm{d}x \\
&= -\int_{-1}^{1} D^{i+1}(1-x^2)^i \, D^{j-1}(1-x^2)^j \, \mathrm{d}x,
\end{aligned}$$

since we have proved that the derivatives vanish at ± 1.

The process of integration by parts can be repeated until we find that

$$\int_{-1}^{1} \varphi_i(x)\varphi_j(x)\mathrm{d}x = (-1)^i \int_{-1}^{1} D^{2i}(1-x^2)^i \, D^{j-i-1}(1-x^2)^j \, \mathrm{d}x.$$

This integral is zero, since $D^{2i}(1-x^2)^i$ is a constant, and the function $D^{j-i-1}(1-x^2)^j$ vanishes at ± 1, $0 \leq i < j$. The polynomials $\varphi_i(x)$ and $\varphi_j(x)$ are therefore orthogonal as required.

Taking $j = 0, 1, 2, 3$ we get

$$\begin{aligned}
\varphi_0(x) &= 1, \\
\varphi_1(x) &= D(1-x^2) \\
&= -2x, \\
\varphi_2(x) &= D^2(1-x^2)^2 \\
&= D(-4x+4x^3) \\
&= -4+12x^2, \\
\varphi_3(x) &= D^3(1-x^2)^3 \\
&= D^2(-6x+12x^3-6x^5) \\
&= D(-6+36x^2-30x^4) \\
&= 72x-120x^3.
\end{aligned}$$

These are the same, apart from constant scale factors as the polynomials given in Example 9.6.

Solution to Exercise 9.5

This is very similar to Exercise 4, and we shall only give an outline solution.

Suppose that

$$D^k\, x^j \mathrm{e}^{-x} = x^{j-k} q_k(x) \mathrm{e}^{-x},$$

where $q_k(x)$ is a polynomial of degree k. Then by differentiation

$$\begin{aligned} D^{k+1}\, x^j \mathrm{e}^{-x} &= \mathrm{e}^{-x}[(j-k)x^{j-k-1} q_k(x) + x^{j-k} q_k'(x) - x^{j-k} q_k(x)] \\ &= x^{j-k-1} q_{k+1} \mathrm{e}^{-x}, \end{aligned}$$

and the first result follows by induction.

To prove that the polynomials form an orthogonal system, write

$$\int_0^\infty \mathrm{e}^{-x} \varphi_i(x) \varphi_j(x) \mathrm{d}x = \int_0^\infty D^i(x^i \mathrm{e}^{-x}) \varphi_j(x) \mathrm{d}x,$$

and the orthogonality follows by repeated integration by parts, as in Exercise 4.

The first members of the sequence are

$$\begin{aligned} \varphi_0(x) &= 1, \\ \varphi_1(x) &= \mathrm{e}^x D(x\mathrm{e}^{-x}) = 1 - x, \\ \varphi_2(x) &= \mathrm{e}^x D[(2x - x^2)\mathrm{e}^{-x}] = 2 - 4x + x^2, \\ \varphi_3(x) &= \mathrm{e}^x D^2[(3x^2 - x^3)\mathrm{e}^{-x}] \\ &= \mathrm{e}^x D[(6x - 3x^2 - 3x^2 + x^3)\mathrm{e}^{-x}] \\ &= 6 - 12x + 9x^2 - x^3. \end{aligned}$$

Solution to Exercise 9.6

Evidently $\varphi_{j+1}(x) - C_j x \varphi_j(x)$ is in general a polynomial of degree $j+1$, but if C_j is chosen to be equal to the ratio of the leading coefficients of $\varphi_{j+1}(x)$ and $\varphi_j(x)$ then the coefficient of the leading term is zero, and the result is a polynomial of degree j only. Hence it can be expressed as a linear combination of the polynomials $\varphi_k(x)$ in the form

$$\varphi_{j+1}(x) - C_j x \varphi_j(x) = \sum_{k=0}^{j} \alpha_{j,k} \varphi_k(x),$$

where

$$\alpha_{j,k} = \frac{1}{A_k} \int_a^b w(x)[\varphi_{j+1}(x) - C_j x \varphi_j(x)]\varphi_k(x) \mathrm{d}x,$$

where

$$A_k = \int_a^b w(x)[\varphi_k(x)]^2 \mathrm{d}x.$$

Now $k \leq j$ in the sum, and so

$$\int_a^b w(x)\varphi_{j+1}(x)\varphi_k(x)\mathrm{d}x = 0;$$

Moreover $\varphi_j(x)$ is orthogonal to every polynomial of lower degree, and so

$$\int_a^b w(x)\varphi_j(x)\, x\, \varphi_k(x)\mathrm{d}x = 0, \qquad k+1 < j.$$

These two equations show that

$$\alpha_{j,k} = 0, \quad k = 0, \ldots, j-1.$$

Hence

$$\varphi_{j+1}(x) - (C_j x + D_j)\varphi_j(x) + E_j \varphi_{j-1}(x) = 0, \qquad j > 0,$$

where $D_j = \alpha_{j,j}$ and $E_j = -\alpha_{j,j-1}$.

Solution to Exercise 9.7

Since C_j is the ratio of the leading coefficients of the polynomials $\varphi_{j+1}(x)$ and $\varphi_j(x)$, which are both positive, C_j is also positive.

We saw in Exercise 6 that

$$\begin{aligned}\alpha_{j,j-1} &= \int_a^b w(x)[\varphi_{j+1}(x) - C_j x\varphi_j(x)]\varphi_{j-1}(x)\mathrm{d}x \\ &= -C_j \int_a^b w(x)x\varphi_j(x)\varphi_{j-1}(x)\mathrm{d}x,\end{aligned}$$

since $\varphi_{j-1}(x)$ and $\varphi_{j+1}(x)$ are orthogonal.

Now the same argument as in Exercise 6, but with j replaced by $j-1$, shows that $\varphi_j(x) - C_{j-1}x\varphi_{j-1}(x)$ is a polynomial of degree $j-1$; it is therefore orthogonal to $\varphi_j(x)$. This shows that

$$\int_a^b w(x)[\varphi_j(x) - C_{j-1}x\varphi_{j-1}(x)]\varphi_j(x)\mathrm{d}x.$$

Hence

$$\int_a^b w(x)x\varphi_j(x)\varphi_{j-1}(x)\mathrm{d}x = \frac{1}{C_{j-1}}\int_a^b w(x)[\varphi_j(x)]^2\mathrm{d}x,$$

which is positive. Hence E_j is positive.

The proof of the interlacing property follows closely the proof of Theorem 5.8; it proceeds by induction. Suppose that the zeros of $\varphi_j(x)$ and $\varphi_{j-1}(x)$ interlace, and that ξ and η are two consecutive zeros of $\varphi_j(x)$. Then

$$\varphi_{j+1}(\xi) = -E_j\varphi_{j-1}(\xi), \quad \varphi_{j+1}(\eta) = -E_j\varphi_{j-1}(\eta).$$

But there is exactly one zero of $\varphi_{j-1}(x)$ between ξ and η, so that $\varphi_{j-1}(\xi)$ and $\varphi_{j-1}(\eta)$ have opposite signs. Hence $\varphi_{j+1}(\xi)$ and $\varphi_{j+1}(\eta)$ also have opposite signs, and there is a zero of $\varphi_{j+1}(x)$ between ξ and η. This has located at least $j-1$ zeros of $\varphi_{j+1}(x)$. Now suppose that ζ is the largest zero of $\varphi_j(x)$; then ζ is greater than all the zeros of $\varphi_{j-1}(x)$, and $\varphi_{j-1}(\zeta) > 0$, since the leading coefficient of each of the polynomials is positive. Hence $\varphi_{j+1}(\zeta) < 0$, and there is a zero of $\varphi_{j+1}(x)$ greater than ζ. By a similar argument $\varphi_{j+1}(x)$ has a zero which is smaller than the smallest zero of $\varphi_j(x)$. This has now located all the zeros of $\varphi_{j+1}(x)$, and shows that the zeros of $\varphi_j(x)$ and $\varphi_{j+1}(x)$ interlace.

To start the induction, the same argument shows that the zero of $\varphi_1(x)$ lies between the zeros of $\varphi_2(x)$.

Solution to Exercise 9.8

We can write

$$\varphi_{n+1}(x) = c_{n+1}^{n+1} x^{n+1} - q_n(x),$$

where $q_n(x)$ is a polynomial of degree n.

Now the best polynomial approximation of degree n to x^{n+1} is determined by the condition that $x^{n+1} - p_n(x)$ is orthogonal to $\varphi_j(x)$, for $j = 0, \ldots, n$. But the above equation shows that

$$x^{n+1} - \frac{q_n(x)}{c_{n+1}^{n+1}} = \frac{\varphi_{n+1}(x)}{c_{n+1}^{n+1}},$$

which clearly satisfies this orthogonal condition. Hence the best polynomial approximation is

$$p_n(x) = \frac{c_{n+1}^{n+1} x^{n+1} - \varphi_{n+1}(x)}{c_{n+1}^{n+1}}.$$

and the expression for the $2-$norm of the difference follows immediately.

Using $w(x) = 1$ on $[-1, 1]$ we know that

$$\varphi_3(x) = x^3 - \tfrac{5}{3}x.$$

so the best approximation to x^3 is the polynomial $\frac{5}{3}x$. The norm of the error is given by

$$\begin{aligned} \|x^3 - p_2\|_2^2 &= \int_{-1}^{1} [x^3 - \tfrac{5}{3}x]^2 \mathrm{d}x \\ &= 152/189. \end{aligned}$$

Notice that in this example $c_j^j = 1$ for every j.

Solution to Exercise 9.9

The proof is by induction, but we shall not fill in all the details. The polynomial $\varphi_k(x)$ is constructed by writing

$$\varphi_k(x) = x^k - a_0\varphi_0(x) - \ldots - a_{k-1}\varphi_{k-1}(x),$$

where

$$a_j = \frac{\int_{-a}^{a} w(x)x^k\varphi_j(x)\mathrm{d}x}{\int_{-a}^{a} w(x)[\varphi_j(x)]^2\mathrm{d}x}.$$

Suppose that the required result is true for polynomials of degree less than k. Since $w(x)$ is an even function, a_j is zero whenever j and k have opposite parity. Thus in a polynomial of odd degree all the even coefficients are zero, and in a polynomial of even degree all the odd coefficients are zero. This gives the required result.

The coefficients γ_j are given by

$$\gamma_j = \frac{\int_{-a}^{a} w(x)f(x)\varphi_j(x)\mathrm{d}x}{\int_{-a}^{a} w(x)[\varphi_j(x)]^2\mathrm{d}x}.$$

Evidently γ_j is zero if $f(x)$ is an even function and j is odd, since we have just shown that φ_j is then an odd function. Similarly if $f(x)$ is an odd function.

Solution to Exercise 9.10

By definition H is an odd function, so by the results of Exercise 9 the polynomial of best approximation of degree 0 is just

$$p_0(x) = 0,$$

and the polynomials of best approximation of degrees 1 and 2 are the same. Moreover p_1 has the form

$$p_1(x) = \gamma_1 \varphi_1(x).$$

The orthogonal polynomials in this case are the Legendre polynomials, so $\varphi_1(x) = x$. Hence

$$\begin{aligned} \gamma_1 &= \frac{\int_{-1}^{1} H(x)\varphi_1(x)\mathrm{d}x}{\int_{-1}^{1} [\varphi_1(x)]^2 \mathrm{d}x} \\ &= \frac{2\int_{0}^{1} x\mathrm{d}x}{\int_{-1}^{1} x^2 \mathrm{d}x} \\ &= \frac{1}{2/3} \\ &= \frac{3}{2}. \end{aligned}$$

Hence the polynomials of best approximation are

$$p_1(x) = p_2(x) = \tfrac{3}{2}x.$$

Solution to Exercise 10.1

We saw in Exercise that the sequence of orthogonal polynomials for the weight function $-\ln x$ on $[0,1]$ begins with

$$\varphi_0(x) = 1, \quad \varphi_1(x) = x - \tfrac{1}{4}, \quad \varphi_2(x) = x^2 - \tfrac{5}{7}x + \tfrac{17}{252}.$$

For $n = 0$ there is just one quadrature point, the zero of $\varphi_1(x)$, which is at $\frac{1}{4}$. The corresponding quadrature weight is

$$W_0 = \int_0^1 -\ln x \, \mathrm{d}x = 1.$$

For $n = 1$ the two quadrature points are the zeros of $\varphi_1(x)$, which are

$$x_0 = \tfrac{5}{14} - \tfrac{1}{42}\sqrt{106}, \quad x_1 = \tfrac{5}{14} + \tfrac{1}{42}\sqrt{106}.$$

The corresponding weights are

$$W_0 = \int_0^1 -\ln x \, (x - x_1)^2/(x_0 - x_1)^2 \, \mathrm{d}x = \tfrac{1}{2} + \tfrac{9}{424}\sqrt{106}$$

and

$$W_1 = \tfrac{1}{2} - \tfrac{9}{424}\sqrt{106}.$$

Note that in this case a good deal of heavy algebra is saved by determining the weights by direct construction, requiring the quadrature formula to be exact for polynomials of degrees 0 and 1. This leads to the two equations

$$\begin{aligned} W_0 + W_1 &= \int -\ln x \, \mathrm{d}x = 1 \\ W_0 x_0 + W_1 x_1 &= \int -\ln x \, x \, \mathrm{d}x = \tfrac{1}{4}. \end{aligned}$$

Solution of these equations gives the same values as above for W_0 and W_1.

Solution to Exercise 10.2

The Gauss quadrature formula

$$\int_a^b w(x)f(x)\mathrm{d}x = \sum_{k=0}^{n} W_k f(x_k)$$

is exact when $f(x)$ is any polynomial of degree $2n+1$. It is therefore exact for the polynomial $L_k(x)$, which has degree n. Since $L_k(x_k) = 1$ and $L_k(x_j) = 0$ for $k \neq j$, this shows that

$$\int_a^b w(x)L_k(x)\mathrm{d}x = W_k.$$

Solution to Exercise 10.3

The first two of the sequence of orthogonal polynomials for the weight function $w(x) = x$ on the interval $[-1, 1]$ are easily found to be

$$\varphi_0(x) = 1$$

and

$$\varphi_1(x) = x - c,$$

where

$$c = \int_0^1 x\, x \mathrm{d}x \Big/ \int_0^1 x \,\mathrm{d}x = 2/3.$$

Hence the Gauss quadrature formula for $n = 0$ with this weight function on $[0, 1]$ has quadrature point $2/3$ and weight

$$W_0 = \int_0^1 x \,\mathrm{d}x = 1/2.$$

The error of this quadrature formula, for a function with a continuous second derivative, is given by Theorem 10.1 as

$$\frac{f''(\eta)}{2!} \int_0^1 x(x - 2/3)^2 \,\mathrm{d}x = \tfrac{1}{72} f''(\eta).$$

This gives the required result.

Solution to Exercise 10.4

We know that the Chebyshev polynomials $T_r(x)$ form an orthogonal system with weight function $w(x) = (1-x^2)^{-1/2}$ on the interval $[-1, 1]$. Hence the quadrature points x_j for $n = 0$ are the zeros of $T_{n+1}(x)$, which are

$$x_j = \cos[(2j+1)\pi/(2n+2)], \quad j = 0, \ldots, n.$$

Suppose that for some value of n

$$\sum_{j=0}^{n} \cos(2j+1)\theta = \frac{\sin(2n+2)\theta}{2\sin\theta}.$$

Then

$$\begin{aligned}
\sum_{j=0}^{n+1} \cos(2j+1)\theta &= \frac{\sin(2n+2)\theta}{2\sin\theta} + \cos(2n+3)\theta \\
&= \frac{\sin(2n+2)\theta + 2\sin\theta\cos(2n+3)\theta}{2\sin\theta} \\
&= \frac{\sin(2n+4)\theta}{2\sin\theta},
\end{aligned}$$

so the same result is true with n replaced by $n+1$. The result is trivially true for $n = 0$, so it holds for all positive integer n.

If $\theta = p\pi$, where p is an integer, then

$$\sum_{j=0}^{n} \cos(2j+1)p\pi = \sum_{j=0}^{n} (-1)^p = (-1)^p(n+1).$$

Hence

$$\sum_{j=0}^{n} \cos k(2j+1)\pi/(2n+2) = \begin{cases} 0 & k = 1, \ldots, n \\ n+1 & k = 0. \end{cases}$$

which means that

$$\sum_{j=0}^{n} T_k(x_j) = 0, \quad k = 1, \ldots, n.$$

But T_k is orthogonal to T_0 for $k > 0$, so

$$\int_{-1}^{1} (1-x^2)^{-1/2} T_k(x) \mathrm{d}x = 0, \quad k = 1, \ldots, n.$$

Moreover

$$\sum_{j=0}^{n} T_0(x_j) = n+1$$

and

$$\int_{-1}^{1} (1-x^2)^{-1/2} T_0(x) \mathrm{d}x = \pi.$$

Putting these together we find that

$$\frac{1}{n+1} \sum_{j=0}^{n} T_k(x_j) = \frac{1}{\pi} \int_{-1}^{1} (1-x^2)^{-1/2} T_k(x) \mathrm{d}x.$$

Thus the quadrature formula with weight function $w(x) = (1-x^2)^{-1/2}$ on the interval $[-1,1]$, with quadrature points and weights

$$x_j = \cos(2j+1)\pi/(2n+2), \quad W_j = \pi/(n+1), \quad k = 0, \ldots, n$$

is exact for the polynomials $T_k(x),\ k = 0, \ldots, n$. It is therefore exact for every polynomial of degree n, and because of the choice of the quadrature points x_j it is also exact for every polynomial of degree $2n+1$.

Solution to Exercise 10.5

Using the same notation as in Section 10.5, write

$$\pi(x) = \prod_{k=1}^{n} (x - x_k^*)^2.$$

This is evidently a polynomial of degree $2n$, and hence the quadrature formula

$$\int_a^b w(x)\pi(x)\mathrm{d}x = W_0\pi(a) + \sum_{k=1}^{n} W_k \pi(x_k)$$

is exact. But by the definition $\pi(x_k) = 0,\ k = 1, \ldots, n$, so that

$$\int_a^b w(x)\pi(x)\mathrm{d}x = W_0\pi(a),$$

Since $w(x)$ and $\pi(x)$ are positive on $[a, b]$, it follows that $W_0 > 0$.

Solution to Exercise 10.6

By integration by parts

$$\begin{aligned}\int_0^\infty \mathrm{e}^{-x} x L_j'(x) p_r(x) \mathrm{d}x &= \left[\mathrm{e}^{-x} x p_r(x) L_j(x)\right]_0^\infty \\ &\quad - \int_0^\infty \mathrm{e}^{-x} L_j(x)[-x p_r(x) + p_r(x) + x p_r'(x)]\mathrm{d}x \\ &= \int_0^\infty \mathrm{e}^{-x} L_j(x) x p_r(x) \mathrm{d}x \\ &\quad - \int_0^\infty \mathrm{e}^{-x} L_j(x)[p_r(x) + x p_r'(x)]\mathrm{d}x.\end{aligned}$$

Since $p_r(x) + x p_r'(x)$ is a polynomial of degree r, and $r < j$, the last term is zero by the orthogonality properties of the Laguerre polynomials $L_j(x)$. This gives the required result, and shows that the polynomials defined by

$$\varphi_j(x) = L_j(x) - L_j'(x)$$

form an orthogonal system for the weight function $w(x) = \mathrm{e}^{-x} x$ on the interval $[0, \infty]$.

The Radau quadrature formula (10.27) thus gives

$$\int_0^\infty \mathrm{e}^{-x} p_{2n}(x)\mathrm{d}x = W_0 p_{2n}(0) + \sum_{k=1}^n W_k p_{2n}(x_k)$$

where the quadrature points x_k, $k = 1, \ldots, n$ are the zeros of the polynomial $L_n(x) - L_n'(x)$. From Exercise 5 we have $L_1(x) = 1 - x$, so that $L_1(x) - L_1'(x) = 2 - x$, which gives the quadrature point $x_1 = 2$.

For the Gauss quadrature formula with weight function $w(x) = x\mathrm{e}^{-x}$ the corresponding weight is

$$W_1^* = \int_0^\infty x\mathrm{e}^{-x} = 1.$$

Hence using (10.28) we have

$$W_1 = W_1^*/x_1^* = 1/2,$$

and

$$W_0 = 1 - 1/2 = 1/2.$$

Therefore

$$\int_0^\infty \mathrm{e}^{-x} p_2(x)\mathrm{d}x = \tfrac{1}{2} p_2(0) + \tfrac{1}{2} p_2(2).$$

Solution to Exercise 10.7

If we divide the polynomial $p_{2n-1}(x)$ by the polynomial $(x-a)(b-x)$, which has degree 2, the quotient is a polynomial of degree $2n-3$, and the remainder $\rho(x)$ has degree 1. This remainder can then be written in the form $\rho(x) = r(x-a) + s(x-b)$.

Using a similar notation to that of Section 10.5, we now write

$$\int_a^b w(x)p_{2n-1}(x)\mathrm{d}x = \int_a^b w(x)(x-a)(b-x)q_{2n-3}(x)\mathrm{d}x + \int_a^b w(x)\rho(x)\mathrm{d}x.$$

Now let $x_k^*, W_k^*,\ k = 1, \ldots, n-1$ be the quadrature points and weights respectively for a Gauss quadrature formula using the modified weight function $w(x)(x-a)(b-x)$ over the interval $[a,b]$. This modified weight function is non-negative on $[a,b]$, and the quadrature formula is exact for every polynomial of degree $2n-3$.

Hence

$$\int_a^b w(x)(x-a)(b-x)q_{2n-3}(x) = \sum_{k=1}^{n-1} W_k^* q_{2n-3}(x_k^*),$$

so that

$$\begin{aligned}\int_a^b w(x)p_{2n-1}(x)\mathrm{d}x &= \sum_{k=1}^{n-1} \frac{W_k^*}{x_k^*-a)(b-x_k^*)} p_{2n-1}(x_k^*) \\ &\quad + \int_a^b w(x)\rho(x)\mathrm{d}x - \sum_{k=1}^{n-1} W_k^* \frac{\rho(x_k^*)}{(x_k^*-a)(b-x_k^*)}\end{aligned}$$

Now

$$\int_a^b w(x)\rho(x)\mathrm{d}x = r\int_a^b w(x)(x-a)\mathrm{d}x + s\int_a^b w(x)(b-x)\mathrm{d}x,$$

and

$$\frac{\rho(x_k^*)}{(x_k^*-a)(b-x_k^*)} = \frac{r}{b-x_k^*} + \frac{s}{x_k^*-a}.$$

It follows from the definitions of $\rho(x), r$ and s that

$$r = p_{2n-1}(b)/(b-a), \quad s = p_{2n-1}(a)/(b-a).$$

We have therefore constructed the Lobatto quadrature formula

$$\int_a^b w(x)f(x)\mathrm{d}x = W_0 f(a) + \sum_{k=1}^{n-1} W_k f(x_k) + W_n f(b),$$

which is exact when f is any polynomial of degree $2n-1$. The quadrature points are $x_k = x_k^*$, and the weights are

$$W_k = \frac{W_k^*}{(x_k^* - a)(b - x_k^*)}, \quad k = 1, \ldots, n-1,$$

and

$$W_0 = \int_a^b w(x)(b-x)\mathrm{d}x - \sum_{k=1}^{n-1} W_k(b - x_k),$$

$$W_n = \int_a^b w(x)(x-a)\mathrm{d}x - \sum_{k=1}^{n-1} W_k(x_k - a).$$

The weights w_k, $k = 1, \ldots, n-1$ are clearly positive, since the weights W_k^* are positive.

To show that W_0 is positive, we apply the quadrature formula to the polynomial

$$P(x) = (b-x)\prod_{k=1}^{n-1}(x - x_k)^2.$$

This is a polynomial of degree $2n-1$, so the quadrature formula is exact. The polynomial $P(x)$ vanishes at the points x_k, $k = 1, \ldots, n-1$, and at b, so we find that

$$\int_a^b w(x)P(x)\mathrm{d}x = W_0 P(a).$$

This shows that $W_0 \geq 0$, since $P(x)$ and $w(x)$ are non-negative on $[a, b]$. The proof that W_n is positive is similar.

Solution to Exercise 10.8

We use direct construction, using the condition that the quadrature formula must be exact for polynomials of degree 3. This gives the conditions

$$\begin{aligned} A_0 + A_1 + A_2 &= \int_{-1}^{1} \mathrm{d}x = 2 \\ -A_0 + A_1 x_1 + A_2 &= \int_{-1}^{1} x \mathrm{d}x = 0 \\ A_0 + A_1 x_1^2 + A_2 &= \int_{-1}^{1} x^2 \mathrm{d}x = 2/3 \\ -A_1 + A_1 x_1^3 + A_2 &= \int_{-1}^{1} x^3 \mathrm{d}x = 0. \end{aligned}$$

From the second and fourth equation we find that $A_1 x_1 (1 - x_1^2) = 0$, so that either $x_1 = 0$ or $x_1 = 1$ or $A_1 = 0$. Since the quadrature points must be distinct we reject the possibility that $x_1 = 1$. From the first and third equations we find that

$$A_1(1 - x_1^2) = 4/3,$$

so that A_1 cannot be zero. We therefore have $x_1 = 0$ and $A_1 = 4/3$. It is then easy to find that $A_0 = A_2 = 1/3$, and the quadrature formula is Simpson's rule.

Solution to Exercise 10.9

Using the given notation, and writing $c = b - a$,

$$I(m) = \frac{b-a}{m}\left[\tfrac{1}{2}f(a) + f(a+\tfrac{1}{m}c) + \ldots + f(a+\tfrac{m-1}{m}c) + \tfrac{1}{2}f(b)\right] \quad (2.1)$$

$$\begin{aligned} S(m) &= \frac{b-a}{6m}[f(a) + 4f(a+\tfrac{1}{2m}c) + 2f(a+\tfrac{1}{m}c) + 4f(a+\tfrac{3}{2m}c) + \ldots \\ &\qquad +2f(a+\tfrac{m-1}{m}c) + 4f(a+\tfrac{2m-1}{2m}c) + f(b)]. \end{aligned}$$

$$M(m) = \frac{b-a}{m}[f(a+\tfrac{1}{2m}c) + \ldots + f(a+\tfrac{2m-1}{2m}c)]. \quad (2.2)$$

Hence

$$\begin{aligned} 2I(2m) - I(m) &= 2\frac{b-a}{2m}[\tfrac{1}{2}f(a) + f(a+\tfrac{1}{2m}c) + f(a+\tfrac{1}{m}c) \\ &\quad +f(a+\tfrac{3}{2m}c) + \ldots + \tfrac{2m-1}{2m}c) + \tfrac{1}{2}f(b)] \\ &\quad -\frac{b-a}{m}\left[\tfrac{1}{2}f(a) + f(a+\tfrac{1}{m}c) + \ldots + \tfrac{1}{2}f(b)\right] \\ &= \frac{b-a}{m}[f(a+\tfrac{1}{2m}c) + \ldots + f(a+\tfrac{2m-1}{2m}c)] \\ &= M(m). \end{aligned}$$

In the same way

$$\begin{aligned} \tfrac{4}{3}I(2m) - \tfrac{1}{3}I(m) &= \tfrac{4}{3}\frac{b-a}{2m}[\tfrac{1}{2}f(a) + f(a+\tfrac{1}{2m}c) + f(a+\tfrac{1}{m}c) \\ &\quad +f(a+\tfrac{3}{2m}c) + \ldots + \tfrac{2m-1}{2m}c) + \tfrac{1}{2}f(b)] \\ &\quad -\tfrac{1}{3}\frac{b-a}{m}\left[\tfrac{1}{2}f(a) + f(a+\tfrac{1}{m}c) + \ldots + \tfrac{1}{2}f(b)\right] \\ &= \frac{b-a}{6m}[f(a) + 4f(a+\tfrac{1}{2m}c) + 2f(a+\tfrac{1}{m}c) + \\ &\quad \ldots + 4f(a+\tfrac{2m-1}{2m}c) + f(b)] \\ &= S(m). \end{aligned}$$

Finally, using these relations,

$$\begin{aligned} \tfrac{2}{3}M(m) + \tfrac{1}{3}I(m) &= \tfrac{4}{3}I(2m) - \tfrac{2}{3}I(m) + \tfrac{1}{3}I(m) \\ &= \tfrac{4}{3}I(2m) - \tfrac{1}{3}I(m) \\ &= S(m). \end{aligned}$$

Solution to Exercise 11.1

Suppose that on the interval $[a, b]$ the knots are at

$$a = x_0 < x_1 < \ldots < x_m = b.$$

On each of the m intervals $[x_i, x_{i+1}]$ the spline is a polynomial of degree n; the spline is therefore determined by $m(n+1)$ independent coefficients. On each of these intervals the value of the polynomial is given at the two ends; this gives $2m$ interpolation conditions. At each of the internal knots $x_1, \ldots, x_{m-1}$ the derivatives of orders $1, 2, \ldots, n-1$ must be continuous; this gives $(m-1)(n-1)$ smoothness conditions.

The number of additional conditions required to define the spline uniquely is therefore

$$m(n+1) - 2m - (m-1)(n-1) = n - 1.$$

Solution to Exercise 11.2

(i) According to Theorem 11.1 applied on the interval $[x_{i-1}, x_i]$,

$$f(x) - s_L(x) = \tfrac{1}{2} f''(\xi)(x - x_{i-1})(x - x_i)\,, \qquad x \in [x_{i-1}, x_i]\,.$$

But $f''(x) \equiv 0$, so $s_L(x) \equiv f(x)$.

(ii) Similar to part (i), but using Theorem 6.4 with $n = 1$.

(iii) According to Definition 11.2, the natural cubic spline must satisfy the end conditions $s_2''(x_0) = s_2''(x_m) = 0$. The polynomial f does not in general satisfy these conditions, so s_2 and f are not in general identical.

Solution to Exercise 11.3

The quantities σ_i are determined by (11.7), which in this case become

$$\begin{aligned} h(\sigma_{i-1} + 4\sigma_i + \sigma_{i+1}) &= (6/h)\{(i+1)^3h^3 - 2i^3h^3 + (i-1)^3h^3\} \\ &= 36ih^2, \quad i = 1, \ldots, m-1, \end{aligned}$$

together with $\sigma_0 = \sigma_m = 0$.

Substituting $f''(x_i) = 6ih$ for σ_i we see at once that the equations are satisfied. Moreover $\sigma_0 = 0$ as required, but $f''(x_m) = 6mh = 6 \neq 0$, so the final equation is not satisfied. Hence these values do not satisfy all the equations, and s_2 is not identical to f.

However, if the two additional equations are replaced by $\sigma_0 = f''(0)$ and $\sigma_m = f''(1)$ then all the equations determining σ_i are satisfied. Since the system of equations is nonsingular this means that $\sigma_i = f''(ih), i = 0, \ldots, m$; hence s_2 and f are identical,

Solution to Exercise 11.4

By definition

$$\begin{aligned} \|f-s\|_2^2 &= \|f-\sum \alpha_k\varphi_k\|_2^2 \\ &= \int_0^1 [f(x)-\sum_k \alpha_k\varphi_k(x)]^2 \mathrm{d}x. \end{aligned}$$

To minimise this we require that

$$\frac{\partial}{\partial \alpha_j}\left(\|f-s\|_2^2\right)=0,$$

which gives

$$-2\int_0^1 [f(x)-\sum_k \alpha_k\varphi_k(x)]\varphi_j(x)\mathrm{d}x=0, \quad j=0,\ldots,m.$$

This yields the required system of equations.

Now $\varphi_k(x)$ is nonzero only on the interval $[x_{k-1},x_{k+1}]$, and so

$$\int_0^1 \varphi_i(x)\varphi_j(x)\,\mathrm{d}x=0 \quad \text{if} \ \ |i-j|>1.$$

Hence the matrix A is tridiagonal. The diagonal elements are given by

$$\begin{aligned} A_{ii} &= \int_0^1 [\varphi_i(x)]^2\,\mathrm{d}x \\ &= \int_{(i-1)h}^{(i+1)h} [\varphi_i(x)]^2\,\mathrm{d}x \\ &= \int_{(i-1)h}^{ih} \frac{(x-(i-1)h)^2}{h^2}\mathrm{d}x + \int_{ih}^{(i+1)h} \frac{(x-(i+1)h)^2}{h^2}\mathrm{d}x \\ &= h\int_0^1 t^2\,\mathrm{d}t + h\int_0^1 t^2\,\mathrm{d}t \\ &= \frac{2}{3}h, \quad i=1,\ldots,m-1 \end{aligned}$$

after an obvious change of variable.

At the two ends the same argument shows that

$$A_{0,0}=A_{m,m}=\frac{1}{3}h.$$

For the nonzero off-diagonal elements we get in the same way

$$A_{i,i+1} = \int_0^1 \varphi_i(x)\varphi_{i+1}(x)\,\mathrm{d}x$$

$$
\begin{aligned}
&= \int_{ih}^{(i+1)h} \frac{(i+1)h - x}{h} \cdot \frac{x - ih}{h} \, \mathrm{d}x \\
&= h \int_0^1 (1-t)t \mathrm{d}t \\
&= \frac{1}{6}h, \quad i = 0, \ldots, m-1,
\end{aligned}
$$

and in the same way

$$A_{i-1,i} = \frac{1}{6}h, \qquad i = 1, \ldots, m.$$

The matrix A is evidently symmetric.

Hence all the elements of A are non-negative, and

$$\frac{2}{3}h = A_{ii} > A_{i,i-1} + A_{i,i+1} = \frac{1}{6}h + \frac{1}{6}h = \frac{1}{3}h, \quad i = 1, \ldots, m-1,$$

with

$$A_{0,0} > A_{0,1} \quad \text{and} \quad A_{m,m} + A_{m,m-1}.$$

These are the conditions required for the success of the Thomas algorithm.

Solution to Exercise 11.5

The elements of the vector b are

$$\begin{aligned} b_i &= \int_0^1 x\varphi_i(x)\,\mathrm{d}x \\ &= \int_{(i-1)h}^{ih} x\frac{x-(i-1)h}{h}\mathrm{d}x + \int_{ih}^{(i+1)h} x\frac{(i+1)h-x}{h}\mathrm{d}x \\ &= h^2\int_0^1 (i-1+t)\,t\,\mathrm{d}t + h^2\int_0^1 (i+1)-t)\,t\,\mathrm{d}t \\ &= ih^2, \quad i=1,\ldots,m-1. \end{aligned}$$

In the same way

$$b_0 = h^2\int_0^1 (1-t)t\,\mathrm{d}t = \frac{1}{6}h^2.$$

Substituting the values $\alpha_i = ih$ we get

$$\begin{aligned} \sum_j A_{ij}\,jh &= \frac{1}{6}h^2[(i-1)+4i+(i+1)] \\ &= ih^2 \\ &= b_i, \end{aligned}$$

and

$$\begin{aligned} \sum_j A_{0,j}\,jh &= \frac{1}{6}h^2[0+1] \\ &= \frac{1}{6}h^2 \\ &= b_0. \end{aligned}$$

Thus the quantities $\alpha_i = ih$ satisfy the system of equations, and the solution is therefore $s(x_i) = \alpha_i = x_i$.

When $f(x) = x^2$ we find in the same way that

$$\begin{aligned} b_i &= \int_0^1 x^2\,\varphi_i(x)\,\mathrm{d}x \\ &= h^3\int_0^1 (j-1+t)^2\,t\,\mathrm{d}t + h^3\int_0^1 (j+1-t)^2\,t\,\mathrm{d}t \\ &= h^3(i^2+1/6). \end{aligned}$$

Using $\alpha_k = (kh)^2 + Ch^2$ gives

$$\begin{aligned} \sum_j A_{ij}\alpha_j &= \frac{1}{6}h^3[(i-1)^2 + 4i^2 + (i+1)^2 + 6C] \\ &= h^3[i^2 + 1/3 + C], \quad i = 1, \ldots, m-1, \end{aligned}$$

so the equations are satisfied if

$$h^3[i^2 + 1/3 + C] = b_i = h^3[i^2 + 1/6]$$

i.e., if $C = -1/6$. A similar argument shows that this value of C also satisfies the equations for $i = 0$ and $i = m$. Hence the solution of the system of equations is $\alpha_k = (kh)^2 - h^2/6$, and so $s(x_k) = \alpha_k = f(x_k) - h^2/6$.

Solution to Exercise 11.6

If $x \leq a$ then $(x-a)_+^n = 0$ and $(x-a)_+^{n+1} = 0$; if $x > a$ then $(x-a)_+ = (x-a)$. So in both cases

$$(x-a)_+^n (x-a) = (x-a)_+^{n+1}.$$

From the definition

$$\begin{aligned}
[(n+2)h - x]S_n(x-h) &= \sum_{k=0}^{n+1} (-1)^k \left(\begin{array}{c} n+1 \\ k \end{array} \right) [(n+2)h - x](x - h - kh)_+^n \\
&= \sum_{k=1}^{n+2} (-1)^{k-1} \left(\begin{array}{c} n+1 \\ k-1 \end{array} \right) [(n+2)h - x](x - kh)_+^n,
\end{aligned}$$

giving

$$\begin{aligned}
& xS_n(x) + [(n+2)h - x]S_n(x-h) \\
&= \sum_{k=1}^{n+1} (-1)^k \left\{ \left(\begin{array}{c} n+1 \\ k \end{array} \right) x - \left(\begin{array}{c} n+1 \\ k-1 \end{array} \right) [(n+2)h - x] \right\} (x - kh)_+^n \\
&\quad + \left(\begin{array}{c} n+1 \\ 0 \end{array} \right) x_+^n \, x + (-1)^{n-1} \left(\begin{array}{c} n+1 \\ n+1 \end{array} \right) [(n+2)h - x](x - (n+2)h)_+^n \\
&= \sum_{k=1}^{n+1} (-1)^k \left(\begin{array}{c} n+2 \\ k \end{array} \right) (x - kh)_+^{n+1} + x_+^{n+1} + (-1)^{n+2} (x - (n+2)h)_+^{n+1} \\
&= S_{n+1}(x),
\end{aligned}$$

where we have used the additive property of the binomial coefficient

$$\left(\begin{array}{c} n+1 \\ k \end{array} \right) + \left(\begin{array}{c} n+1 \\ k-1 \end{array} \right) = \left(\begin{array}{c} n+2 \\ k \end{array} \right).$$

Now to show that $S_n(x) \geq 0$; it is clear that $S_1(x) \geq 0$ for all x. Suppose that, for some positive integer k, $S_k(x) \geq 0$ for all x. Notice that $S_k(x) = 0$ when $x \leq 0$, and when $x \geq (k+1)h$. It then follows that $xS_k(x) \geq 0$, and that $[(k+2)h - x]S_k(x) \geq 0$. Thus $S_{k+1}(x) \geq 0$ for all x, and the required result follows by induction.

Solution to Exercise 11.7

The proof is by induction, starting from the fact that $S_1(x)$ is clearly symmetric. Now suppose that, for some positive integer n, $S_n(x)$ is symmetric. Then use of the recurrence relation from Exercise 6 shows that

$$\begin{aligned} S_{n+1}((n+2)h/2+x) &= [(n+2)h/2+x]S_n((n+2)h/2+x) \\ &\quad +[(n+2)h-(n+2)h/2-x]S_n((n+2)h/2+x-h) \\ &= [(n+2)h/2+x]S_n((n+2)h/2+x) \\ &\quad +[(n+2)h/2-x]S_n(nh/2+x) \\ &= [(n+2)h/2+x]S_n((n+1)h/2+x+h/2) \\ &\quad +[(n+2)h/2-x]S_n((n+1)h+x-h/2). \end{aligned}$$

Using the symmetry of $S_n(x)$ we now see that changing the sign of x merely interchanges the two terms in this expression, leaving the sum unchanged. Hence $S_{n+1}(x)$ is symmetric, and so by induction every $S_n(x)$ is symmetric.

Solution to Exercise 12.1

a)

$$|f(x,y) - f(x,z)| = 2x^{-4}|y-z| \leq 2|y-z|$$

for all $x \in [1,\infty)$ and all $y, z \in \mathbb{R}$. The Lipschitz constant is therefore $L = 2$.

b) By the Mean Value Theorem,

$$|f(x,y) - f(x,z)| = |\frac{\partial f}{\partial y}(x,\eta)|\,|y-z|$$

where η lies between y and z. In our case,

$$\frac{\partial f}{\partial y}(x,\eta) = \mathrm{e}^{-x^2}\,\frac{1}{1+\eta^2} \leq \frac{1}{\mathrm{e}}\,,$$

for all $x \in [1,\infty)$ and all $y, z \in \mathbb{R}$. Therefore, the Lipschitz constant is $L = 1/\mathrm{e}$.

c)

$$\begin{aligned} |f(x,y) - f(x,z)| &= (1+\mathrm{e}^{-|x|})\left|\frac{2y}{1+y^2} - \frac{2z}{1+z^2}\right| \\ &= 2(1+\mathrm{e}^{-|x|})\frac{|1-yz|}{(1+y^2)(1+z^2)}\,|y-z| \\ &\leq 2 \times 2 \times 1 \times |y-z|\,, \end{aligned}$$

for all $x \in (-\infty,\infty)$ and for all $y, z \in \mathbb{R}$. The Lipschitz constant is therefore $L = 4$.

Solution to Exercise 12.2

Observe that $y(x) \equiv 0$ is the trivial solution to the initial value problem. To find nontrivial solutions to the initial value problem we separate the variables (assuming now that $y \neq 0$):

$$y^{-\frac{2m}{2m+1}} \mathrm{d}y = \mathrm{d}x \,.$$

Integrating this yields

$$y(x) = \left(\frac{x + C}{2m+1} \right)^{2m+1} ,$$

which is a solution to the differential equation $y' = y^{2m/(2m+1)}$ for any choice of the constant C. Hence,

$$y_b(x) = \begin{cases} \left(\frac{x+b}{2m+1}\right)^{2m+1} , & x \leq -b \\ 0 \,, & x \in [-b, b] \\ \left(\frac{x-b}{2m+1}\right)^{2m+1} , & x \geq b \end{cases}$$

is a solution to the given initial value problem for any $b \geq 0$. Thus we have infinitely many solutions to the initial value problem.

This does not contradict Picard's Theorem, since $f(x, y) = y^{2m/(2m+1)}$ does not satisfy a Lipschitz condition in any neighbourhood of a point $(x, 0)$ for any $x \in \mathbb{R}$. This can be seen by showing that the opposite of the Lipschitz condition holds: for any $L > 0$ there exists $(x, y) \neq (x, 0)$ such that

$$|f(x, y) - f(x, 0)| = |y^{2m/(2m+1)}| = |y|^{2m/(2m+1)} > L|y| \,.$$

Such is for example (x, y) with any $x \in \mathbb{R}$ and

$$0 < |y| < \left(\frac{1}{L} \right)^{2m+1} .$$

Solution to Exercise 12.3

The solution to the initial value problem is

$$y(x) = 1 + \frac{p+q}{p}(\mathrm{e}^{px} - 1)\,.$$

Picard's iteration has the form: $y_0(x) \equiv 1$,

$$y_{n+1}(x) = 1 + \int_0^x (py_n(t) + q)\mathrm{d}t = 1 + qx + p\int_0^x y_n(t)\mathrm{d}t\,, \qquad n = 0, 1, 2, \ldots$$

The function y_0 is a polynomial of degree 0, and therefore, by induction, y_n is a polynomial of degree n. In particular,

$$\begin{aligned}
y_0(x) &\equiv 1\,,\\
y_1(x) &= 1 + (p+q)x\,,\\
y_2(x) &= 1 + (p+q)x + p(p+q)\frac{x^2}{2!}\,,\\
y_3(x) &= 1 + (p+q)x + p(p+q)\frac{x^2}{2!} + p^2(p+q)\frac{x^3}{3!}\\
&\text{etc.}\\
y_n(x) &= 1 + \frac{p+q}{p}\left[px + \frac{(px)^2}{2!} + \ldots + \frac{(px)^n}{n!}\right]\\
&= 1 + \frac{p+q}{p}\left[\sum_{k=0}^{n}\frac{(px)^k}{k!} - 1\right]\,.
\end{aligned}$$

Passing to the limit as $n \to \infty$, we have that

$$\lim_{n\to\infty} y_n(x) = 1 + \frac{p+q}{p}(\mathrm{e}^{px} - 1) = y(x)\,,$$

as required.

Solution to Exercise 12.4

Euler's method for the initial value problem has the form

$$y_{n+1} = y_n + h y_n^{1/5}\,, \quad n = 0, 1, 2, \ldots\,, \qquad y_0 = 0\,.$$

Hence $y_n = 0$ for all $n \geq 0$. Therefore Euler's method approximates the trivial solution $y(x) \equiv 0$, rather than $y(x) = (4x/5)^{5/4}$.

The implicit Euler method for the initial value problem is

$$y_{n+1} = y_n + h y_{n+1}^{1/5}\,, \quad n = 0, 1, 2, \ldots\,, \qquad y_0 = 0\,.$$

Put $y_n = (C_n h)^{5/4}$; then,

$$C_{n+1}^{5/4} - C_{n+1}^{1/4} = C_n^{5/4}\,, \qquad n \geq 0\,.$$

$y_0 = 0$ implies that $C_0 = 0$. Thus, $C_1^{5/4} - C_1^{1/4} = C_1^{1/4}(C_1 - 1) = 0$, which means that either $C_1 = 0$ or $C_1 = 1$. Taking $C_1 = 1$, we shall prove by induction the existence of $C_n > 1$ for all $n \geq 2$ such that $y_n = (C_n h)^{5/4}$ is a solution of the implicit Euler scheme.

We begin by observing that

$$C_{n+1}^{1/4}((C_{n+1}^{1/4})^4 - 1) = (C_n^{1/4})^5\,.$$

Putting $t = C_{n+1}^{1/4}$, this gives the polynomial equation

$$p(t) \equiv t(t^4 - 1) - (C_n^{1/4})^5 = 0\,.$$

Suppose that $C_n \geq 1$ for some $n \geq 1$ (this is certainly true for $n = 1$). As $p(1) < 0$ and $\lim_{t \to +\infty} p(t) = +\infty$, the polynomial p has a root, t_*, say, in the interval $(1, \infty)$; therefore, $C_{n+1} = t_*^4 > 1$. By induction, then, $C_n > 1$ for all $n \geq 2$.

Solution to Exercise 12.5

The exact solution of the initial value problem is

$$y(x) = \frac{1}{2}x^2 \mathrm{e}^{-5x}\,.$$

The Euler approximation is

$$y_{n+1} = y_n + h(x_n \mathrm{e}^{-5x_n} - 5y_n)\,, \quad n = 0, 1, 2, \ldots, \qquad y_0 = 0\,.$$

Clearly, $y_1 = 0$, $y_2 = h^2\mathrm{e}^{-5h}$, etc.

$$\begin{aligned} y_{n+1} &= h^2\mathrm{e}^{-5h}\sum_{k=0}^{n-1}(k+1)(\mathrm{e}^{-5h})^k(1-5h)^{n-1-k} \\ &= h^2\mathrm{e}^{-5h}(1-5h)^{n-1}\sum_{k=0}^{n-1}(k+1)\left(\frac{\mathrm{e}^{-5h}}{1-5h}\right)^k \end{aligned}$$

Taking $n = N-1$ where $h = 1/N$, we have

$$y_N = \frac{1}{N^2}\mathrm{e}^{-5/N}\left(1-\frac{5}{N}\right)^{N-2}\sum_{k=0}^{N-2}(k+1)\left(\frac{\mathrm{e}^{-5/N}}{1-(5/N)}\right)^k\,.$$

Passing to the limit,

$$\begin{aligned} \lim_{N\to\infty} y_N &= \lim_{N\to\infty}\frac{1}{N^2}\mathrm{e}^{-5/N}\left(1-\frac{5}{N}\right)^{N-2}\sum_{k=0}^{N-2}(k+1) \\ &= \mathrm{e}^{-5}\lim_{N\to\infty}\frac{1}{N^2}\frac{N(N-1)}{2} \\ &= \frac{1}{2}\mathrm{e}^{-5} \\ &= y(1)\,, \end{aligned}$$

as required.

Solution to Exercise 12.6

a) The truncation error of Euler's method is

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n)) = \frac{y(x_{n+1}) - y(x_n)}{h} - y'(x_n) = \frac{1}{2} h y''(\xi_n)$$

for some $\xi_n \in (x_n, x_{n+1})$. In our case, $y' = \ln\ln(4 + y^2)$. Therefore,

$$y'' = \frac{\mathrm{d}}{\mathrm{d}x} y' = \frac{\mathrm{d}}{\mathrm{d}x} \ln\ln(4+y^2) = \frac{1}{\ln(4+y^2)} \frac{1}{4+y^2} 2yy' = \frac{\ln\ln(4+y^2)}{\ln(4+y^2)} \frac{2y}{4+y^2},$$

and hence $|y''(x)| \le 1 \times \frac{1}{2} = \frac{1}{2}$. Thus, $|T_n| \le \frac{1}{4} h$.

b) We have

$$\begin{aligned} y_{n+1} &= y_n + h f(x_n, y_n), \\ y(x_{n+1}) &= y(x_n) + h f(x_n, y(x_n)) + h T_n. \end{aligned}$$

By subtraction and using a Lipschitz condition for $f(x, y) = \ln\ln(4+y^2)$, with Lipschitz constant L, and letting $e_n = y(x_n) - y_n$, we have $e_0 = 0$ and

$$|e_{n+1}| \le |e_n| + hL|e_n| + h|T_n|, \qquad n = 0, 1, \ldots, N-1.$$

To calculate the Lipschitz constant of f, note that by the Mean Value Theorem,

$$|f(x, y) - f(x, z)| = \left|\frac{\partial f}{\partial y}(x, \eta)\right| |y - z|$$

for some η which lies between y and z. In our case,

$$\frac{\partial f}{\partial y} = \frac{1}{\ln(4+y^2)} \frac{2y}{4+y^2}$$

and therefore

$$\left|\frac{\partial f}{\partial y}\right| \le \frac{1}{\ln 4} \times \frac{1}{2}.$$

Hence, $L = 1/(2\ln 4)$.

c) From part b) we have $e_0 = 0$ and

$$|e_{n+1}| \le |e_n| + hL|e_n| + h|T_n|, \qquad n = 0, 1, \ldots, N-1.$$

Therefore, letting T be such that $|T_n| \le T$ for all $n \ge 0$ (e.g., $T = \frac{1}{4} h$),

$$\begin{aligned} |e_1| &\le hT, \\ |e_2| &\le (1 + hL)hT + hT, \\ |e_3| &\le (1 + hL)^2 hT + (1 + hL)hT + hT, \end{aligned}$$

etc.

$$
\begin{aligned}
|e_n| \quad &\leq \quad (1+hL)^{n-1}hT + \ldots + hT = hT\left\{(1+hL)^{n-1} + \ldots + 1\right\} \\
&= hT\frac{(1+hL)^n - 1}{1+hL-1} = \frac{T}{L}[(1+hL)^n - 1] \leq \frac{T}{L}\left((\mathrm{e}^{hL})^n - 1\right) \\
&= \frac{T}{L}\left(\mathrm{e}^{nhL} - 1\right) \leq \frac{T}{L}(\mathrm{e}^{L} - 1)\,,
\end{aligned}
$$

as in our case $nh \leq Nh = 1$ for all $n = 0, 1, \ldots, N-1$. This gives,

$$\max_{0\leq n\leq N} |y(x_n) - y_n| \leq \frac{h}{4} \times (2\ln 4) \times (\mathrm{e}^{1/(2\ln 4)} - 1) = h \times 0.30103067384$$

which is less than 10^{-4} provided that $N = 1/h \geq 3011 = N_0$.

Solution to Exercise 12.7

The truncation error is

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \frac{1}{2}h\left(f(x_{n+1}, y(x_{n+1})) + f(x_n, y(x_n))\right) .$$

Using repeated integration by parts,

$$\begin{aligned}
\int_{x_n}^{x_{n+1}} (x - x_{n+1})(x - x_n)y'''(x)\mathrm{d}x &= (x - x_{n+1})(x - x_n)y''(x) \,|_{x=x_n}^{x=x_{n+1}} \\
&\quad - \int_{x_n}^{x_{n+1}} (2x - x_n - x_{n+1})y''(x)\mathrm{d}x \\
&= - \int_{x_n}^{x_{n+1}} (2x - x_n - x_{n+1})y''(x)\mathrm{d}x \\
&\quad - (2x - x_n - x_{n+1})y'(x) \,|_{x=x_n}^{x=x_{n+1}} + \int_{x=x_n}^{x_{n+1}} 2y'(x)\mathrm{d}x \\
&= -(x_{n+1} - x_n)y'(x_{n+1}) + (x_n - x_{n+1})y'(x_n) + 2[y(x_{n+1}) - y(x_n)] .
\end{aligned}$$

Therefore,

$$y(x_{n+1}) - y(x_n) = \frac{h}{2}[y'(x_{n+1}) + y'(x_n)] - \frac{1}{2}\int_{x_n}^{x_{n+1}} (x_{n+1} - x)(x - x_n)y'''(x)\mathrm{d}x .$$

Using the Integral Mean Value Theorem on the right-hand side,

$$y(x_{n+1}) - y(x_n) = \frac{h}{2}[y'(x_{n+1}) + y'(x_n)] - \frac{1}{2}y'''(\xi_n)\int_{x_n}^{x_{n+1}} (x_{n+1} - x)(x - x_n)\mathrm{d}x$$

with $\xi_n \in (x_n, x_{n+1})$. Now, using the change of variable $x = x_n + sh$,

$$\int_{x_n}^{x_{n+1}} (x_{n+1} - x)(x - x_n)\mathrm{d}x = h^3 \int_0^1 (1 - s)s\mathrm{d}s = h^3 \int_0^1 (s - s^2)\mathrm{d}s = \tfrac{1}{6}h^3 .$$

Thus, we have

$$y(x_{n+1}) - y(x_n) = \frac{h}{2}[f(x_{n+1}, y(x_{n+1})) + f(x_n, y(x_n))] - \frac{1}{12}h^3 y'''(\xi_n) ,$$

with $\xi_n \in (x_n, x_{n+1})$. Hence,

$$T_n = -\frac{1}{12}h^2 y'''(\xi_n) .$$

In particular,

$$|T_n| \le \frac{1}{12}h^2 M ,$$

where $M = \max_{\xi \in \mathbb{R}} |y'''(\xi)|$.

To derive a bound on the global error $e_n = y(x_n) - y_n$, note that

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + \frac{h}{2}[f(x_{n+1}, y(x_{n+1})) + f(x_n, y(x_n))] + hT_n\,, \\ y_{n+1} &= y_n + \frac{h}{2}[f(x_{n+1}, y_{n+1}) + f(x_n, y_n)]\,. \end{aligned}$$

Subtracting and using a Lipschitz condition on f, we have

$$|e_{n+1}| \le |e_n| + \frac{h}{2}(L|e_{n+1}| + L|e_n|) + h|T_n|$$

and therefore,

$$|e_{n+1}| \le |e_n| + \frac{h}{2}(L|e_{n+1}| + L|e_n|) + \frac{1}{12}h^3M\,.$$

This can be rewritten as follows:

$$|e_{n+1}| \le \left(\frac{1+\frac{1}{2}hL}{1-\frac{1}{2}hL}\right)|e_n| + \frac{\frac{1}{12}h^3M}{1-\frac{1}{2}hL}\,, \quad n = 0, 1, 2, \ldots\,,$$

with $e_0 = 0$ (assuming that $0 < h < 1/L$). By induction, this implies that

$$|e_n| \le \frac{h^2M}{12L}\left[\left(\frac{1+\frac{1}{2}hL}{1-\frac{1}{2}hL}\right)^n - 1\right]\,.$$

Solution to Exercise 12.8

We shall perform a Taylor series expansion of the truncation error

$$
\begin{aligned}
T_n &= \frac{y(x_{n+1})-y(x_n)}{h} - \tfrac{1}{2}[f(x_n,y(x_n)) + f(x_{n+1},y(x_n)+hf(x_n,y(x_n)))] \\
&= \frac{y(x_{n+1})-y(x_n)}{h} - \tfrac{1}{2}[y'(x_n) + f(x_{n+1},y(x_n)+hy'(x_n))] \\
&= y'(x_n) + \tfrac{1}{2}hy''(x_n) + \tfrac{1}{6}h^2y'''(x_n) + \mathcal{O}(h^3) \\
&\quad -\tfrac{1}{2}\{y'(x_n) + f(x_n,y(x_n)) + hf_x(x_n,y(x_n)) + hy'(x_n)f_y(x_n,y(x_n)) \\
&\quad +\tfrac{1}{2}[h^2f_{xx}(x_n,y(x_n)) + 2h^2y'(x_n)f_{xy}(x_n,y(x_n)) + h^2(y'(x_n))^2f_{yy}(x_n,y(x_n))]\} + \mathcal{O}(h^3) \\
&= \tfrac{1}{6}h^2[f_y(f_x+f_yf) - \tfrac{1}{2}(f_{xx}+2f_{xy}f+f_{yy}f^2)] \mid_{x=x_n} + \mathcal{O}(h^3)
\end{aligned}
$$

where, in the transition to the last line, we used that

$$
\begin{aligned}
y' &= f\,, \\
y'' &= f_x + f_yy' = f_x + f_yf\,, \\
y''' &= f_{xx} + f_{xy}y' + f_yy'' + (f_{xy}+f_{yy}y')y' \\
&\quad = f_{xx} + 2f_{xy}f + f_xf_y + f(f_y)^2 + f_{yy}f^2\,.
\end{aligned}
$$

Solution to Exercise 12.9

The classical 4th-order Runge–Kutta method is

$$y_{n+1} = y_n + \tfrac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4)$$

where, in our case,

$$\begin{aligned}
k_1 &= f(x_n, y_n) = \lambda y_n \\
k_2 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_1) \\
&= \lambda(y_n + \tfrac{1}{2}\lambda h y_n) = \lambda y_n + \tfrac{1}{2}\lambda^2 h y_n \\
k_3 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_2) \\
&= \lambda(y_n + \tfrac{1}{2}\lambda h y_n + \tfrac{1}{4}\lambda^2 h^2 y_n) \\
&= \lambda y_n + \tfrac{1}{2}\lambda^2 h y_n + \tfrac{1}{4}\lambda^3 h^2 y_n \\
k_4 &= f(x_n + h, y_n + hk_3) \\
&= \lambda(y_n + \lambda h y_n + \tfrac{1}{2}\lambda^2 h^2 y_n + \tfrac{1}{4}\lambda^3 h^3 y_n) \\
&= \lambda y_n + \lambda^2 h y_n + \tfrac{1}{2}\lambda^3 h^2 y_n + \tfrac{1}{4}\lambda^4 h^3 y_n \,.
\end{aligned}$$

Therefore,

$$y_{n+1} = \left(1 + (\lambda h) + \tfrac{1}{2}(\lambda h)^2 + \tfrac{1}{6}(\lambda h)^3 + \tfrac{1}{24}(\lambda h)^4\right) y_n \,.$$

On the other hand, for the exact solution,

$$y(x_{n+1}) = \mathrm{e}^{\lambda x_{n+1}} = \mathrm{e}^{\lambda x_n}\mathrm{e}^{\lambda h} = \mathrm{e}^{\lambda h} y(x_n) \,,$$

and therefore,

$$y(x_{n+1}) = \left(1 + (\lambda h) + \tfrac{1}{2}(\lambda h)^2 + \tfrac{1}{6}(\lambda h)^3 + \tfrac{1}{24}(\lambda h)^4 + \ldots\right) y(x_n) \,.$$

The factor multiplying y_n in the numerical method coincides with the factor multiplying $y(x_n)$ in the exact solution, up to terms of order $\mathcal{O}(h^4)$.

Solution to Exercise 12.10

Taylor series expansion of the truncation error yields

$$\begin{aligned} T_n &= (1-\alpha-\beta)y'(x_n) + h\left(\tfrac{1}{2}-\beta\gamma\right)y''(x_n) + \tfrac{1}{6}h^2 y'''(x_n) \\ &\quad -\frac{1}{2}h^2\beta\gamma^2(f_{xx}+2y'f_{xy}+(y')^2 f_{yy})\,|_{x=x_n} + \mathcal{O}(h^3)\,. \end{aligned}$$

For consistency, we require that $\alpha+\beta=1$. For second order accuracy, we demand, in addition, that $\beta\gamma=\frac{1}{2}$.

Suppose that $\alpha+\beta=1$ and $\beta\gamma=\frac{1}{2}$ and apply the resulting numerical method to $y'=y$, $y(0)=1$ whose exact solution in $y(x)=\mathrm{e}^x$. Then,

$$T_n = \tfrac{1}{6}h^2\mathrm{e}^{x_n} - \tfrac{1}{2}h^2\beta\gamma^2(0+0+0) = \tfrac{1}{6}h^2\mathrm{e}^{x_n}$$

which cannot be made equal to zero for any choice of α, β and γ. Therefore, there is no choice of these parameters for which the order of the method exceeds 2.

Suppose that $\alpha+\beta=1$ and $\beta\gamma=\frac{1}{2}$, and apply the method to $y'=-\lambda y$, $y(0)=1$, where $\lambda>0$. Thus, $y_0=1$ and

$$y_{n+1} = (1-\lambda h+\tfrac{1}{2}(\lambda h)^2)y_n \quad n=0,1,\ldots\,.$$

Therefore,

$$y_n = (1-\lambda h+\tfrac{1}{2}(\lambda h)^2))^n\,, \qquad n=0,1,\ldots\,.$$

The sequence (y_n) is bounded if, and only if,

$$|1-\lambda h+\tfrac{1}{2}(\lambda h)^2| \le 1$$

which holds, if and only if, $0<h\le 2/\lambda$. Now, assuming this restriction on h,

$$\begin{aligned} y(x_n)-y_n &= \mathrm{e}^{-\lambda x_n} - (1-\lambda h+\tfrac{1}{2}(\lambda h)^2)^n \\ &= (\mathrm{e}^{-\lambda h})^n - (1-\lambda h+\tfrac{1}{2}(\lambda h)^2)^n \\ &= \left(\mathrm{e}^{-\lambda h} - (1-\lambda h+\tfrac{1}{2}(\lambda h)^2\right) \\ &\quad \times\left[(\mathrm{e}^{-\lambda h})^{n-1} + (\mathrm{e}^{-\lambda h})^{n-2}(1-\lambda h+\tfrac{1}{2}\lambda^2h^2) + \ldots + (1-\lambda h+\tfrac{1}{2}\lambda^2h^2)^{n-1}\right]. \end{aligned}$$

Each of the n terms in the last line is bounded by 1, and therefore

$$|(\mathrm{e}^{-\lambda h})^{n-1}+(\mathrm{e}^{-\lambda h})^{n-2}(1-\lambda h+\tfrac{1}{2}\lambda^2h^2)+\ldots+(1-\lambda h+\tfrac{1}{2}\lambda^2h^2)^{n-1}| \le n\,.$$

On the other hand, (for example, by the Leibniz criterion for the remainder in a convergent alternating series)

$$|\mathrm{e}^{-\lambda h} - (1-\lambda h+\tfrac{1}{2}(\lambda h)^2| \le \tfrac{1}{6}(\lambda h)^3$$

and therefore,

$$|y(x_n) - y_n| \leq \tfrac{1}{6}(h\lambda)^3 n = \tfrac{1}{6}\lambda^3 h^2 x_n$$

where $x_n = nh$, $n \geq 0$.

Solution to Exercise 12.11

For this method, $\alpha_3 = 1$, $\alpha_2 = \alpha$, $\alpha_1 = -\alpha$, $\alpha_0 = -1$, $\beta_3 = 0$, $\beta_2 = \beta$, $\beta_1 = \beta$, $\beta_0 = 0$. Therefore,

$$\begin{aligned} C_0 &= 0\,, \\ C_1 &= 3 + \alpha - 2\beta\,, \\ C_2 &= \tfrac{9}{2} + \tfrac{3\alpha}{2} - 3\beta\,, \\ C_3 &= \tfrac{27}{6} + \tfrac{7\alpha}{6} - \tfrac{5\beta}{2}\,, \\ C_4 &= \tfrac{81}{24} + \tfrac{51\alpha}{24} - \tfrac{9\beta}{6}\,, \\ C_5 &= \tfrac{243}{120} + \tfrac{31\alpha}{120} - \tfrac{17\beta}{24}\,. \end{aligned}$$

Setting $C_1 = 0$ implies that $\alpha - 2\beta = -3$ and therefore $C_2 = 0$ also. Setting $C_3 = 0$ implies that $7\alpha - 15\beta = -27$. Solving the linear system $\alpha - 2\beta = -3$, $7\alpha - 15\beta = -27$ gives $\alpha = 9$, $\beta = 6$; with this choice, we have $C_0 = C_1 = C_2 = C_3 = 0$, and also, $C_4 = 0$ and $C_5 = 1/10$.

Therefore, with $\alpha = 9$, $\beta = 6$ the method is 4th-order accurate. For such α and β, the first characteristic polynomial of the method is

$$\rho(z) = z^3 + 9(z^2 - z) - 1 = (z - 1)(z^2 + 10z + 1)$$

which has the roots $z_1 = 1$, $z_{2/3} = -5 \pm \sqrt{24}$. Since the Root Condition is violated by z_3, the method is not zero-stable.

Solution to Exercise 12.12

For this method, $\alpha_3 = 1$, $\alpha_2 = 0$, $\alpha_1 = b$, $\alpha_0 = a$, $\beta_3 = 0$, $\beta_2 = 1$, $\beta_1 = 0$, $\beta_0 = 0$. These give

$$\begin{aligned} C_0 &= 1 + b + a\,, \\ C_1 &= 2 + b\,. \end{aligned}$$

Setting $C_0 = 0$, $C_1 = 0$, to achieve consistency of the method, implies that $a = 1$, $b = -2$. Next,

$$C_2 = \frac{9+b}{2} - 2 = \tfrac{7}{2} - 2 = \tfrac{3}{2} \neq 0\,.$$

Therefore, the method is consistent and at most first order accurate.

To investigate the zero-stability of the method for $a = 1$ and $b = -2$, consider the first characteristic polynomial of the method:

$$\rho(z) = z^3 - 2z + 1 = (z-1)(z^2 + z - 1)\,.$$

This has roots

$$z_1 = 1\,, \qquad z_{2/3} = \tfrac{1}{2}(-1 \pm \sqrt{5})$$

one of which (z_3) is outside the closed unit disc. By the Root Condition, the method is not zero-stable. Further, by Dahlquist's Equivalence Theorem the method is not convergent.

Let us apply the method to $y' = 0$, $y(0) = 1$ whose exact solution is $y(x) \equiv 1$. Then,

$$y_{n+3} - 2y_{n+1} + y_n = 0\,.$$

The general solution of this third-order linear recurrence relation is

$$y_n = A \cdot 1^n + B \cdot \left(\frac{-1+\sqrt{5}}{2}\right)^n + C \cdot \left(\frac{-1-\sqrt{5}}{2}\right)^n\,,$$

where A, B, C are arbitrary constants. For suitable initial conditions $A = 0$, $B = 0$, $C = 1$, and therefore

$$y_n = (-1)^n \left(\frac{1+\sqrt{5}}{2}\right)^n\,,$$

which means that the numerical solution oscillates between positive and negative values whose absolute value increases exponentially with n, exhibiting 'zero-instability'. Clearly, the solution bears no resemblance to the exact solution $y(x) \equiv 1$.

Solution to Exercise 12.13

The first characteristic polynomial of the method is $\rho(z) = z^2 - \alpha$, $\alpha > 0$, whose two roots are $z_{1/2} = \pm\sqrt{\alpha}$. To ensure zero-stability (using the Root Condition), we need to assume that $0 < \alpha \leq 1$.

To explore the accuracy of the method, note that $\alpha_2 = 1$, $\alpha_1 = 0$, $\alpha_0 = -\alpha$, $\beta_2 = \frac{1}{3}$, $\beta_1 = \frac{4}{3}$, $\beta_0 = \frac{1}{3}$. Therefore,

$$C_0 = 1 - \alpha = 0 \quad \text{with } \alpha = 1\,.$$

With $\alpha = 1$ we also have $C_1 = C_2 = C_3 = C_4 = 0$, while $C_5 = -\frac{1}{90}$. Therefore, the method is 4th-order accurate.

Since the method is zero-stable and 4th-order accurate (in particular it is consistent) when $\alpha = 1$, it follows from Dahlquist's Equivalence Theorem that it is 4th order convergent when $\alpha = 1$.

Solution to Exercise 12.14

Let us explore the zero-stability of the methods.
a) $\rho(z) = z - 1$. This has only one root, $z = 1$. By the Root Condition, the method is zero-stable.
b) $\rho(z) = z^2 + z - 1$. This has roots $z_1 = 1$ and $z_2 = -2$. The second root violates the Root Condition; the method is not zero-stable.
c) $\rho(z) = z^2 - 1$. This has roots $z_{1/2} = \pm 1$. By the Root Condition the method is zero-stable.
d) $\rho(z) = z^2 - z = z(z-1)$, whose roots are $z_1 = 0$ and $z_2 = 1$, so the method is zero-stable.
e) $\rho(z) = z^2 - z = z(z-1)$, so the method is zero-stable.

Let us explore the absolute stability of the methods a) and c).
a) The stability polynomial of the method is $\pi(z, \lambda h) = z - 1 - \lambda h$ whose only root is $z_1 = 1 + \lambda h$. For the method to be absolutely stable, it is necessary and sufficient that $|z_1| < 1$, *i.e.,* $-1 < 1 + \lambda h < 1$. Hence, the method is absolutely stable if, and only if, $0 < h < 2/(-\lambda)$.
c) Here $\pi(z, \lambda h) = z^2 - 1 - \frac{1}{3}\lambda h(z^2 + 4z + 1)$. Now $\pi(z, \lambda h) = 0$ if, and only if,

$$z^2 - \frac{4\lambda h}{3 - \lambda h} z - \frac{3 + \lambda h}{3 - \lambda h} = 0\,.$$

Note that since $\lambda < 0$ and $h > 0$, we have $3 - \lambda h \neq 0$.

Given a quadratic polynomial of the form $z^2 - az + b$, for both roots to lie in the open interval $(-1, 1)$ it is necessary and sufficient that the polynomial is positive at $z = 1$ and $z = -1$, it is negative at the stationary point $z = a/2$ (*i.e.,* $a^2/4 - a^2/2 + b = b - a^2/4 < 0$) and $-1 < a/2 < 1$ (*i.e.,* $-2 < a < 2$).

In our case, the first two requirements yield

$$\begin{aligned} 1 - \frac{4\lambda h}{3 - \lambda h} - \frac{3 + \lambda h}{3 - \lambda h} &> 0\,, \\ 1 + \frac{4\lambda h}{3 - \lambda h} - \frac{3 + \lambda h}{3 - \lambda h} &> 0\,, \end{aligned}$$

with $\lambda < 0$. These inequalities yield the contradictory requirements that $3 - \lambda h > 0$ and $3 - \lambda h < 0$. The method is not absolutely stable for any value of $h > 0$.

Solution to Exercise 12.15

For this method, $\alpha_2 = 1$, $\alpha_1 = -(1+a)$, $\alpha_0 = 1$, $\beta_2 = (3-a)/4$, $\beta_1 = 0$, $\beta_0 = (1-3a)/4$.

With these values, $C_0 = 1-a$ is equal to zero if, and only if $a = 1$. Let us suppose that $a = 1$. Then, $C_1 = C_2 = C_3 = 0$ and $C_4 = -1/12 \neq 0$, which means that the method is 3rd-order accurate when $a = 1$. For $a \neq 0$ the method is not consistent.

To check zero stability, for $a = 1$, we consider the first characteristic polynomial $\rho(z) = z^2 - 2z + 1 = (z-1)^2$. This has double root $z = 1$ on the unit circle. The Root Condition implies that the method is not zero-stable.

For absolute stability for $a = 1$, consider the stability polynomial $\pi(z, \lambda h) = z^2 - 2z + 1 - \frac{1}{2}\lambda h(z^2 - 1)$. Proceeding as in part (c) of the previous question, we find that there is no $h > 0$ for which the method is absolutely stable.

Solution to Exercise 12.16

We seek a method of the form

$$ay_{n+2} + by_{n+1} + cy_{n+1} = hf_{n+2}\,.$$

For this method, $\alpha_2 = a$, $\alpha_1 = b$, $\alpha_0 = c$, $\beta_2 = 1$, $\beta_1 = 0$, $\beta_0 = 0$. We demand that

$$\begin{aligned} C_0 &= a + b + c = 0\,, \\ C_1 &= 2a + b - 1 = 0\,, \\ C_2 &= \tfrac{4a}{2} + \tfrac{b}{2} - 2 = 0\,. \end{aligned}$$

Solving the resulting linear system for a, b, c gives

$$a = \tfrac{3}{2}\,, \quad b = -2\,, \quad c = \tfrac{1}{2}\,.$$

This choice ensures that the method is 2nd-order accurate.

The first characteristic polynomial of the resulting method is $\rho(z) = \frac{3}{2}z^2 - 2z + \frac{1}{2}$ which has roots $z_1 = 1$ and $z_2 = 1/3$. By the Root Condition the method is zero-stable, and Dahlquist's Equivalence Theorem implies that the method is 2nd-order convergent.

The stability polynomial of the method is

$$\pi(z, \lambda h) = \tfrac{3}{2}z^2 - 2z + \tfrac{1}{2} - \lambda h z^2\,.$$

The roots of this have absolute value less than one if and only if

$$\pi(1, \lambda h) > 0\,, \quad \pi(-1, \lambda h) > 0\,,$$

at the stationary point $z_0 = 2/(3 - 2\lambda h)$ where $\pi'_z(z_0, \lambda h) = 0$ we have $\pi(z_0, \lambda h) < 0$ and $-1 < z_0 < 1$.

These requirements yield $-\lambda h > 0$, $4 - \lambda h > 0$, $(1/2) - 1/((3/2) - \lambda h) > 0$, and $-1 < 2/(3 - 2\lambda h) < 1$, each of which holds for all $h > 0$ and all $\lambda < 0$. The method is absolutely stable for all $h \in (-\infty, 0)$.

Solution to Exercise 12.17

The θ-method has stability polynomial

$$\pi(z, \lambda h) = z - 1 - \lambda h[(1-\theta) + \theta z]$$

whose only root is

$$z = \frac{1 + \lambda h(1-\theta)}{1 - \lambda h\theta},$$

where $\lambda = a + \imath b$ is a complex number with negative real part, $a < 0$. Now,

$$|z|^2 = \frac{(1 + ah(1-\theta))^2 + b^2h^2(1-\theta)^2}{(1 - ah\theta)^2 + h^2\theta^2b^2}.$$

We have $|z| < 1$ if, and only if,

$$\frac{2a}{a^2 + b^2} < h(2\theta - 1).$$

For the method to be A-stable we need this to be true for all λ with negative real part a, which is true if, and only if, $2\theta - 1 \geq 0$, that is, when $\frac{1}{2} \leq \theta \leq 1$.

Solution to Exercise 12.18

The quadratic Lagrange interpolation polynomial of y is

$$\begin{aligned} p_2(x) &= \frac{(x-x_{n+1})(x-x_{n+2})}{(x_n-x_{n+1})(x_n-x_{n+2})}y(x_n) \\ &\quad + \frac{(x-x_n)(x-x_{n+2})}{(x_{n+1}-x_n)(x_{n+1}-x_{n+2})}y(x_{n+1}) \\ &\quad + \frac{(x-x_n)(x-x_{n+1})}{(x_{n+2}-x_n)(x_{n+2}-x_{n+1})}y(x_{n+2})\,. \end{aligned}$$

Differentiation of this gives

$$p_2'(x_{n+2}) = \frac{1}{2h}[3y(x_{n+2}) - 4y(x_{n+1}) + y(x_n)]\,.$$

By expanding $y(x_n)$ and $y(x_{n+1})$ into a Taylor series about x_{n+2} we get

$$p_2'(x_{n+2}) = y'(x_{n+2}) + \mathcal{O}(h^2)\,.$$

The truncation error of BDF2 is

$$T_n = \frac{3y(x_{n+2}) - 4y(x_{n+1}) + y(x_n)}{2h} - f(x_{n+2}, y(x_{n+2}))\,.$$

Noting that the last term is equal to $y'(x_{n+2})$ and expanding each of $y(x_n)$ and $y(x_{n+1})$ into a Taylor series about x_{n+2}, we get

$$T_n = \frac{1}{3}h^2 y'''(\xi_n) - \frac{2}{3}h^2 y'''(\eta_n)\,,$$

where $\xi_n \in (x_{n+1}, x_{n+2})$ and $\eta_n \in (x_n, x_{n+2})$. Therefore $T_n = \mathcal{O}(h^2)$.

Solution to Exercise 12.19

The implicit Runge–Kutta method has the form

$$y_{n+1} = y_n + h(b_1 k_1 + b_2 k_2)$$

where

$$\begin{aligned} k_1 &= f(x_n + c_1 h, y_n + a_{11} k_1 h + a_{12} k_2 h)\,, \\ k_2 &= f(x_n + c_2 h, y_n + a_{21} k_1 h + a_{22} k_2 h)\,. \end{aligned}$$

When applied to $y' = \lambda y$, we get

$$\begin{aligned} k_1 &= \lambda(y_n + a_{11} k_1 h + a_{12} k_2 h)\,, \\ k_2 &= \lambda(y_n + a_{21} k_1 h + a_{22} k_2 h)\,. \end{aligned}$$

Rearranging this,

$$\begin{aligned} (1 - \lambda a_{11} h) k_1 \qquad - \lambda a_{12} h k_2 &= \lambda y_n \\ -\lambda a_{21} h k_1 + (1 - \lambda a_{21} h) k_2 &= \lambda y_n\,. \end{aligned}$$

The matrix of this linear system is $I - \lambda h A$, and the corresponding determinant is $\Delta = \det(I - \lambda h A)$. In expanded form,

$$\Delta = 1 - \lambda h(a_{11} + a_{22}) + (\lambda h)^2 (a_{11} a_{22} - a_{12} a_{21})\,.$$

Let us suppose that $\Delta \neq 0$. Solving the linear system for k_1 and k_2 gives

$$k_1 = \frac{\lambda y_n (1 + \lambda h(a_{12} - a_{22}))}{\Delta}\,, \qquad k_2 = \frac{\lambda y_n (1 + \lambda h(a_{21} - a_{11}))}{\Delta}\,.$$

For the method with the given Butcher tableau,

$$c_1 = \tfrac{1}{6}(3 - \sqrt{3})\,, \qquad c_2 = \tfrac{1}{6}(3 + \sqrt{3})\,,$$

$$b_1 = b_2 = \tfrac{1}{2}$$

$$a_{11} = \tfrac{1}{4}\,, \qquad a_{12} = \tfrac{1}{12}(3 - 2\sqrt{3})\,,$$

$$a_{21} = \tfrac{1}{12}(3 + 2\sqrt{3})\,, \qquad a_{22} = \tfrac{1}{4}\,.$$

Therefore,

$$\begin{aligned} k_1 &= \lambda y_n (1 - \tfrac{1}{6}\lambda\sqrt{3})/\Delta \\ k_2 &= \lambda y_n (1 + \tfrac{1}{6}\lambda\sqrt{3})/\Delta \end{aligned}$$

which then gives

$$y_{n+1} = (1 + \frac{\lambda h}{\Delta}) y_n$$

with

$$\Delta = 1 - \tfrac{1}{2}\lambda h + \tfrac{1}{12}(\lambda h)^2 \,.$$

Thus,

$$y_{n+1} = \frac{1 + \frac{1}{2}(\lambda h) + \frac{1}{12}(\lambda h)^2}{1 - \frac{1}{2}(\lambda h) + \frac{1}{12}(\lambda h)^2} \,.$$

Writing the quadratic polynomials in the numerator and the denominator in factorised form, we have

$$y_{n+1} = \frac{(\lambda h + 3 + \imath\sqrt{3})(\lambda h + 3 - \imath\sqrt{3})}{(\lambda h - 3 - \imath\sqrt{3})(\lambda h - 3 + \imath\sqrt{3})} y_n \,.$$

The numerical solution will exhibit (exponential) decay for complex $\lambda = a + \imath b$ with negative real part a, $a < 0$, provided that

$$\left| \frac{(\lambda h + 3 + \imath\sqrt{3})(\lambda h + 3 - \imath\sqrt{3})}{(\lambda h - 3 - \imath\sqrt{3})(\lambda h - 3 + \imath\sqrt{3})} \right|^2 < 1 \,.$$

Writing $p = 3 + \imath\sqrt{3}$, this can be written as follows:

$$\left| \left(\frac{\lambda h + p}{\lambda h - p} \right) \left(\frac{\lambda h + \bar{p}}{\lambda h - \bar{p}} \right) \right|^2 < 1 \,,$$

The expression on the left can be rewritten as

$$\frac{|\lambda h|^2 + |p|^2 + 2\mathrm{Re}(p\lambda h)}{|\lambda h|^2 + |p|^2 - 2\mathrm{Re}(p\lambda h)} \frac{|\lambda h|^2 + |p|^2 + 2\mathrm{Re}(p\overline{\lambda h})}{|\lambda h|^2 + |p|^2 - 2\mathrm{Re}(p\overline{\lambda h})} \equiv \frac{A + \mathrm{Re}(X)}{A - \mathrm{Re}(X)} \frac{A + \mathrm{Re}(\bar{X})}{A - \mathrm{Re}(\bar{X})}$$

where $A = |\lambda h|^2 + |p|^2$, and $X = 2p\lambda h$.

Now,

$$\frac{A + \mathrm{Re}(X)}{A - \mathrm{Re}(X)} \frac{A + \mathrm{Re}(\bar{X})}{A - \mathrm{Re}(\bar{X})} < 1$$

if, and only if,

$$2A(\mathrm{Re}(X) + \mathrm{Re}(\bar{X})) < 0 \,.$$

In our case $p = 3 + \imath\sqrt{3}$, $\lambda = a + \imath b$, so

$$\mathrm{Re}(X) = \mathrm{Re}(2p\lambda h) = 2h(3a - \sqrt{3}b) \,, \quad \mathrm{Re}(\bar{X}) = \mathrm{Re}(\overline{2p\lambda h}) = 2h(3a + \sqrt{3}b)$$

which means that

$$\mathrm{Re}(X) + \mathrm{Re}(\bar{X}) = 12ha$$

As $A > 0$ and $\mathrm{Re}(X) + \mathrm{Re}(\bar{X}) = 12ha < 0$, we deduce that

$$\frac{A + \mathrm{Re}(X)}{A - \mathrm{Re}(X)} \frac{A + \mathrm{Re}(\bar{X})}{A - \mathrm{Re}(\bar{X})} < 1$$

for all complex λ with negative real part. Consequently,

$$\left| \frac{(\lambda h + 3 + \imath\sqrt{3})(\lambda h + 3 - \imath\sqrt{3})}{(\lambda h - 3 - \imath\sqrt{3})(\lambda h - 3 + \imath\sqrt{3})} \right|^2 < 1$$

for all complex λ with negative real part. This implies that the method is A-stable.

Solution to Exercise 13.1

Expanding $y(x+h)$ and $y(x-h)$ in Taylor series and adding the results gives

$$y(x+h)+y(x-h) = 2y(x)+h^2y''(x)+\frac{2}{4!}h^4y^{IV}(x)+\frac{1}{6!}h^6[y^{VI}(\xi_1)+y^{VI}(\xi_2)],$$

where $x-h < \xi_2 < x < \xi_1 < x+h$. Since y^{VI} is continuous, there is an η such that

$$y^{VI}(\xi_1) + y^{VI}(\xi_2) = 2y^{VI}(\eta), \quad x-h < \eta < x+h.$$

The required result is then immediate.

Solution to Exercise 13.2

The matrix M has elements

$$M_{jj} = \frac{2}{h^2} + r_j, \quad M_{j,j-1} = M_{j,j+1} = -1/h^2.$$

Since $r(x) > 0$ this matrix satisfies the conditions of Theorem 3.6 since the diagonal elements are positive, the off-diagonal elements are negative, and in each row the diagonal element is at least as large as the sum of the magnitudes of the off-diagonal elements. Hence the matrix M is monotone.

Now define M^* to be a matrix identical to M, except that each diagonal element is $M^*_{jj} = 2/h^2$. Then evidently M^* is also monotone, and since $M \geq M^*$,

$$||M^{-1}||_\infty \leq ||M^* - 1||_\infty.$$

But we know from Exercise 4 that $||M^{-1}||_\infty \leq \frac{1}{8}$.

Solution to Exercise 13.3

To determine the local truncation we substitute the values of the exact solution of the problem $y(x_j)$ in place of y_j and find the residual in the difference approximation. Since we are using the exact solution, we may replace $r(x)y(x)-f(x)$ by $y''(x)$. Expanding in Taylor series and writing $x_{j-1} = x-h$, $x_{j+1} = x+h$, on assuming y''' is continuous we get

$$\begin{aligned} y(x-h)+y(x+h)-2y(x) &= h^2y''(x)+\frac{1}{3!}h^3[y'''(\xi_1)-y'''(\xi_2)] \\ \beta_{-1}y''(x-h)+\beta_0y''(x)+\beta_1y''(x+h) &= (\beta_{-1}+\beta_0+\beta_1)y'' \\ &\quad +h[-\beta_{-1}y'''(\xi_3)+\beta_1y'''(\xi_4)], \end{aligned}$$

from which the first result follows.

(ii) In the same way, when $\beta_{-1}+\beta_0+\beta_1 = 1$, and assuming y^{IV} is continuous

$$\begin{aligned} y(x-h)+y(x+h)-2y(x) &= h^2y''(x)+\frac{1}{4!}h^4[y^{IV}(\xi_1)+y^{IV}(\xi_2)] \\ \beta_{-1}y''(x-h)+\beta_0y''(x)+\beta_1y''(x+h) &= y''+h[-\beta_{-1}y'''(x)+\beta_1y'''(x)] \\ &\quad +\frac{1}{2}h^2[\beta_{-1}y^{IV}(\xi_3)+\beta_1y^{IV}(\xi_4)], \end{aligned}$$

(iii) When $\beta_{-1}+\beta_0+\beta_1 = 1$, $\beta_{-1} = \beta_1 \neq 1/12$, we get on requiring that y^{VI} is continuous

$$\begin{aligned} y(x-h)+y(x+h)-2y(x) &= h^2y''(x)+\frac{1}{12}h^4y^{IV}(x) \\ &\quad +\frac{1}{6!}h^6[y^{VI}(\xi_{-1}+y^{VI}(\xi_2)] \\ \beta_{-1}y''(x-h)+\beta_0y''(x)+\beta_1y''(x+h) &= y''+\beta_1h^2y^{IV}(x)+ \\ &\quad \frac{\beta_1}{4!}h^4[y^{VI}(\xi_3)+y^{VI}(\xi_4)] \end{aligned}$$

(iv) In the same way, when $\beta_{-1} = \beta_1 = \frac{1}{12}$ and $\beta_0 = \frac{5}{6}$, now requiring that $y^{VIII}($ is continuous,

$$\begin{aligned} y(x-h) + y(x+h) - 2y(x) &= h^2 y''(x) + \frac{1}{12} h^4 y^{IV}(x) \\ &\quad + \frac{2}{6!} h^6 y^{VI}(x) \\ &\quad + \frac{1}{8!} h^8 [y^{VIII}(\xi_{-1}) + y^{VIII}(\xi_2)], \\ \beta_{-1} y''(x-h) + \beta_0 y''(x) + \beta_1 y''(x+h) &= y'' + \frac{1}{12} h^2 y^{IV}(x) \\ &\quad + \frac{1}{12}\,\frac{2}{4!} h^4 y^{VI}(x) \\ &\quad \frac{1}{12}\,\frac{1}{6!} h^6 [y^{VIII}(\xi_3) + y^{VIII}(\xi_4)] \end{aligned}$$

from which the last result follows.

Solution to Exercise 13.4

Taylor expansion, with integral form of the remainder, gives

$$\begin{aligned} y(x+h) &= y(x) + hy'(x) + \frac{1}{2!}h^2y''(x) + \frac{1}{3!}h^3y'''(x) \\ &\quad + \frac{1}{4!}h^4y^{IV}(x) + \frac{1}{5!}h^5y^{V}(x) \\ &\quad + \int_0^h \frac{(h-s)^5}{5!}y^{VI}(x+s)\mathrm{d}s. \end{aligned}$$

The expansion of $y(x-h)$ is obtained by changing the sign of h, so the integral term becomes

$$\int_0^{-h} \frac{(-h-s)^5}{5!}y^{VI}(x+s)\mathrm{d}s = \int_{-h}^0 \frac{(h+s)^5}{5!}y^{VI}(x+s)\mathrm{d}s.$$

In the same way we obtain

$$\begin{aligned} y''(x+h) &= y''(x) + hy'''(x) + \frac{1}{2!}h^2y^{IV}(x) + \frac{1}{3!}h^3y^{V}(x) \\ &\quad \int_0^h \frac{(h-s)^3}{3!}y^{VI}(x+s)\mathrm{d}s. \end{aligned}$$

The expansion of $y''(x-h)$ is similar, with the integral replaced by

$$\int_{-h}^0 \frac{(h+s)^3}{3!}y^{VI}(x+s)\mathrm{d}s.$$

Inserting these expressions into the expression for the truncation error, only the integral terms are left, and we obtain

$$\begin{aligned} h^2T_j &= -\int_0^h \frac{(h-s)^5}{5!}y^{VI}(x+s)\mathrm{d}s \\ &\quad -\int_{-h}^0 \frac{(h+s)^5}{5!}y^{VI}(x+s)\mathrm{d}s \\ &\quad +\frac{1}{12}h^2\int_0^h \frac{(h-s)^3}{3!}y^{VI}(x+s)\mathrm{d}s \\ &\quad +\frac{1}{12}h^2\int_{-h}^0 \frac{(h+s)^3}{3!}y^{VI}(x+s)\mathrm{d}s. \end{aligned}$$

Hence

$$h^2T_j = \int_{-h}^h G(s)y^{VI}(x_j+s)\mathrm{d}s,$$

as required, with $G(s) = G(-s)$. Now G may be written

$$G(s) = \frac{(h-s)^3}{360}[3(h-s)^2 - 5h^2], \quad 0 \le s \le h.$$

The factor $(h-s)^3$ is non-negative on $[0,h]$. The quadratic factor has a minimum at $s = h$, where the value is $-5h^2$, and it takes the value $-2h^2$ at $s = 0$. Hence this factor is negative on $(0,h)$. This shows that $G(s) \le 0$ on $[-h,h]$. The integral mean value theorem then shows that ther exists a value of $\eta \in (-h,h)$ such that

$$h^2 T_j = -y^{VI}(x_j + \eta) \int_{-h}^{h} G(s)\mathrm{d}s.$$

A simple calculation shows that

$$\int_{-h}^{h} G(s)\mathrm{d}s = 2\int_0^h \left[\frac{t^5}{5!} - h^2\frac{t^3}{72}\right] \mathrm{d}t = -\frac{h^6}{240},$$

which is the required result.

Solution to Exercise 13.5

The proof follows the same lines as that of Theorem 13.4, but with two differences.

First, the operator $L(\varphi)$ is now

$$\begin{aligned} L(\varphi)_j &= -\frac{\varphi_{j-1} - 2\varphi_j + \varphi_{j+1}}{h^2} + \frac{1}{12}[r_{j-1}\varphi_{j-1} + 10 r_j \varphi_j + r_{j+1}\varphi_{j+1}] \\ &= -a_j\varphi_{j-1} + b_j\varphi_j - c_{j+1}\varphi_{j+1}, \end{aligned}$$

where

$$\begin{aligned} a_j &= 1/h^2 - r_{j-1}/12 \\ b_j &= 2/h^2 + 10 r_j/12 \\ c_j &= 1/h^2 - r_{j+1}/12. \end{aligned}$$

To apply the maximum principle we require that each of the coefficients a_j, b_j and c_j is non-negative. This requires that $h^2 r(x) < 12$ on [a,b].

The other difference is in the algebraic details. With the same definition of the function φ we now find that

$$L(\varphi)_j = -8C + [r_{j-1}\varphi_{j-1} + 10 r_j \varphi_j + r_{j+1}\varphi_{j+1}/12.$$

Since $r_j\varphi_j \geq 0$ the same argument still applies. We can define $C = T/8$, and deduce that $|e_j| \leq T/8$. Using the exxpression for T_j obtained in Exercise 4, this gives the required result.

Solution to Exercise 13.6

The proof in the text has shown that $e_j + \varphi_j \leq 0$, for $j = 0, 1, \ldots, n$. There are two final steps in the proof.

First, since $C \geq 0$ and $D \geq 0$ we see that

$$\varphi_j \geq E, \quad j = 0, 1, \ldots, n.$$

This means that

$$\begin{aligned} e_j &\leq -E \\ &= C(b-a)^2 + D(b-a) \\ &= \frac{1}{24}h^2 M_4 (b-a)^2 + \frac{1}{6}h^2 M_3 (b-a). \end{aligned}$$

The final step is to prove the same inequality for $-e_j$. With the same definitions it is easy to see that

$$L^*(-e + j + \varphi_j) \leq 0, \quad j = 0, 1, \ldots, n-1.$$

The maximum principle then shows that

$$-e_j + \varphi_j \leq \max(-e_0 + \varphi_0, -e_n + \varphi_n, 0);$$

moreover, $e_n = \varphi_n = 0$. In the last part of the proof in the text we can replace e_j by $-e_j$ throughout, and the argument still holds. This shows that $-e_0 + \varphi_0 \leq 0$. Hence $-e_j \leq -\varphi_j, j = 0, 1, \ldots, n$. As we have shown that $\varphi_j \geq E$, this means that

$$-e_j \leq -E, \quad j = 0, 1, n-1.$$

Putting these two results together we get $|e_j| \leq -E$, which is the required result.

Solution to Exercise 13.7

The function $\cosh ax$ evidently satisfy the differential equation. The factor $1/\cosh a$ then ensures that it also satisfies the boundary conditions. The problem is clearly symmetric about $x = 0$.

With the given function Y_j we see that

$$\begin{aligned}\frac{Y_{j-1} - 2Y_j + Y_{j+1}}{h^2} &= \frac{\cosh\vartheta(x_j - h) + \cosh\vartheta(x_j + h) - 2\cosh\vartheta x_j}{h^2\cosh\vartheta} \\ &= \frac{2\cosh\vartheta x_j(\cosh\vartheta h - 1)}{h^2\cosh\vartheta},\end{aligned}$$

so that Y_j is the solution of the difference approximation provided that

$$-2\frac{\cosh\vartheta h - 1}{h^2} + a^2 = 0.$$

This is equivalent to

$$\vartheta = (1/h)\cosh^{-1}(1 + \tfrac{1}{2}a^2h^2).$$

It is also clear that $Y_0 = Y_n = 1$, so the boundary conditions are satisfied.

Expanding in Taylor series we find that

$$\cosh\vartheta = \cosh a - \frac{a^3h^2}{24}\sinh a + O(h^4),$$

$$\cosh\vartheta x = \cosh ax - \frac{a^3xh^2}{24}\sinh ax + O(h^4),$$

$$\frac{\cosh\vartheta x}{\cosh\vartheta} = \frac{\cosh ax}{\cosh a} + \frac{h^2a^3}{24}\,\frac{\cosh ax\sinh a - x\sinh ax\cosh a}{(\cosh(a)^2} + O(h^4).$$

The term in h^2 has a maximum when its derivative vanishes. This leads to either $x = 0$ or

$$\frac{\tanh ax}{ax} = \frac{\cosh a}{a\sinh a - \cosh a}.$$

Now $\tanh a < a$ when $a > 0$, so the right hand side is negative, but the left hand side is positive. Hence the only maximum is at $x = 0$, and here the term in h^2 is

$$\frac{1}{24}h^2a^3\sinh a/(\cosh a)^2 \le \frac{a^4}{24}h^2,$$

since $\tanh a \le a$ and $\cosh a \ge 1$.

The truncation error is

$$T_j = \frac{h^2}{12}y^{IV}(\xi) = \frac{h^2}{12}\,\frac{a^4\cosh ax}{\cosh a} \le \frac{a^4}{12}h^2$$

Theorem 13.4 then shows that

$$|Y_j - y(x_j)| \leq \frac{(b-a)^2}{8} \frac{a^4}{12} h^2 = \frac{a^4}{24} h^2$$

also.

Solution to Exercise 13.8

The analysis is the same as in the previous exercise, but with a^2 replaced by $-a^2$, or with a replaced by ia. The result is

$$y(x) = \frac{\cos ax}{\cos a},$$

$$\begin{aligned} Y_j - y(x_j) &= \frac{1}{24}h^2a^3(\cos ax \sin a - x \sin ax \cos a)/(\cos a)^2 \\ &\quad + O(h^4). \end{aligned}$$

Theorem 13.4 cannot be applied to this problem, as it requires that $r(x) > 0$; here $r(x) = -a^2 < 0$. The analysis requires that $\cos a \neq 0$; if $\cos a = 0$ the boundary value problem has no solution. Hence a must not be an odd multiple of $\pi/2$.

Note also that it is more difficult to obtain a simple bound on the term of order h^2 in this case. In general it may have any number of maxima and minima in $[-1, 1]$, and the largest of them will not be at $x = 0$.

Solution to Exercise 13.9

Evidently the function $y = \sin m\pi x$ satisfies the differential equation $-y'' = m^2\pi^2 y$, and also satisfies the boundary conditions $y(0) = y(1) = 0$, when m is an integer.

The numerical approximation $Y_j = \sin m\pi x_j$ also satisfies the boundary conditions, and

$$\begin{aligned} Y_{j-1} + Y_{j+1} &= \sin m\pi(x_j - h) + \sin m\pi(x_j + h) \\ &= 2\sin m\pi x_j \cos m\pi h \\ &= 2Y_j \cos m\pi h. \end{aligned}$$

Hence

$$-\frac{\delta^2 Y_j}{h^2} = -\frac{2\cos m\pi h - 2}{h^2} Y_j = \mu Y_j,$$

showing that the difference equations are satisfied, with

$$\mu = \frac{2(1 - \cos m\pi h)}{h^2}.$$

Expanding in Taylor series gives

$$\begin{aligned} \mu &= \frac{2}{h^2}\left[\tfrac{1}{2}(m\pi h)^2 - \tfrac{1}{24}(m\pi h)^4 \cos\xi\right] \\ &= m^2\pi^2 - \tfrac{1}{12} m^4\pi^4 h^2 \cos\xi, \end{aligned}$$

where $-m\pi h < \xi < m\pi h$. This gives

$$|\lambda - \mu| \le \tfrac{1}{12} m^4\pi^4 h^2.$$

The truncation error of the approximation is

$$T_j = \tfrac{1}{12} h^2 y^{IV} = \tfrac{1}{12} h^2 m^4\pi^4 y_j + O(h^4)$$

So the error bound in (13.23) gives

$$|\lambda - \mu| \le \frac{\frac{1}{12} h^2 m^4\pi^4 ||y||}{||y||} = \tfrac{1}{12} h^2 m^4\pi^4 + O(h^4),$$

which agrees with the result just obtained.

Solution to Exercise 14.1

(a)

$$\begin{aligned} v^2(x) &= \left(\int_a^x 1 \cdot v'(t) \mathrm{d}t \right)^2 \\ &\leq \int_a^x 1^2 \mathrm{d}t \int_a^x |v'(t)|^2 \mathrm{d}t \\ &= (x-a) \int_a^x |v'(t)|^2 \mathrm{d}t \\ &\leq (x-a) \int_a^b |v'(t)|^2 \mathrm{d}t \end{aligned}$$

for all $x \in [a,b]$, using the Cauchy–Schwarz inequality. On integrating both sides,

$$\int_a^b |v(x)|^2 \mathrm{d}x \leq \frac{1}{2}(b-a)^2 \int_a^b |v'(x)|^2 \mathrm{d}x \,,$$

as required.

(b)

$$\begin{aligned} |v(x)|^2 &= \int_a^x \frac{\mathrm{d}}{\mathrm{d}t}[v(t)]^2 \mathrm{d}t \\ &= 2 \int_a^x v(t)\, v'(t) \mathrm{d}t \\ &\leq 2 \left(\int_a^x |v(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}} \left(\int_a^x |v'(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}} \\ &\leq 2 \left(\int_a^b |v(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}} \left(\int_a^b |v'(t)|^2 \mathrm{d}t \right)^{\frac{1}{2}} \\ &= 2 \|v\|_{\mathrm{L}^2(a,b)} \, \|v'\|_{\mathrm{L}^2(a,b)} \end{aligned}$$

for all $x \in [a,b]$, and therefore

$$\max_{x \in [a,b]} |v(x)|^2 \leq 2 \|v\|_{\mathrm{L}^2(a,b)} \, \|v'\|_{\mathrm{L}^2(a,b)} \,,$$

as required.

Solution to Exercise 14.2

(a) The weak formulation of the problem is: find $u \in \mathrm{H}^1_0(0,1)$ such that $\mathcal{A}(u,v) = \ell(v)$ for all v in $\mathrm{H}^1_0(0,1)$, where

$$\begin{aligned} \mathcal{A}(u,v) &= \int_0^1 (u'v' + uv)\mathrm{d}x\,, \\ \ell(v) &= \int_0^1 fv\mathrm{d}x\,. \end{aligned}$$

(b) To derive the weak formulation of the problem, multiply the differential equation by an arbitrary function $v \in \mathrm{H}^1(0,1)$, and integrate by parts in the term involving $u''v$. Then,

$$\begin{aligned} \int_0^1 (-u'' + u)v\mathrm{d}x &= \int_0^1 (u'v' + uv)\mathrm{d}x - u'v|_{x=0}^{x=1} \\ &= \int_0^1 (u'v' + uv)\mathrm{d}x - u'(1)v(1) + u'(0)v(0)\,. \end{aligned}$$

Since we have no information about $u'(0)$, we eliminate this term by selecting $v \in \mathrm{H}^1(0,1)$ such that $v(0) = 0$. Let us therefore define

$$\mathrm{H}^1_{\mathrm{E}_0}(0,1) = \{v \in \mathrm{H}^1(0,1) \,:\, v(0) = 0\}\,.$$

Hence, noting also that $u'(1) = 1$, we have that

$$\int_0^1 (-u'' + u)v\mathrm{d}x = \int_0^1 (u'v' + uv)\mathrm{d}x - v(1) \qquad \forall v \in \mathrm{H}^1_{\mathrm{E}_0}(0,1)\,,$$

and therefore

$$\int_0^1 (u'v' + uv)\mathrm{d}x = v(1) + \int_0^1 fv\mathrm{d}x \qquad \forall v \in \mathrm{H}^1_{\mathrm{E}_0}(0,1)\,.$$

On observing that u needs to satisfy the same homogeneous boundary condition at $x = 0$ as the one we have required for v, we conclude that the weak formulation of the problem is: find $u \in \mathrm{H}^1_{\mathrm{E}_0}(0,1)$ such that $\mathcal{A}(u,v) = \ell(v)$ for all $v \in \mathrm{H}^1_{\mathrm{E}_0}(0,1)$, where

$$\begin{aligned} \mathcal{A}(u,v) &= \int_0^1 (u'v' + uv)\mathrm{d}x\,, \\ \ell(v) &= v(1) + \int_0^1 fv\mathrm{d}x\,. \end{aligned}$$

(c) Proceeding in the same way as in part (b), we multiply the differential

equation by $v \in \mathrm{H}^1(0,1)$ and integrate by parts in the term involving $u''v$. This yields

$$\int_0^1 (u'v' + uv)\mathrm{d}x - u'(1)v(1) + u'(0)v(0) = \int_0^1 fv\mathrm{d}x$$

for all $v \in \mathrm{H}^1(0,1)$. As $-u'(1) = u(1) - 2$ from the boundary condition at $x = 1$, we can use this in the identity above; also, since we have no information about $u'(0)$, again we choose $v \in \mathrm{H}^1(0,1)$ such that $v(0) = 0$ to eliminate the term involving $u'(0)$. Thus, the weak formulation is: find $u \in \mathrm{H}^1_{\mathrm{E}_0}(0,1)$ such that $\mathcal{A}(u,v) = \ell(v)$ for all $v \in \mathrm{H}^1_{\mathrm{E}_0}(0,1)$, where

$$\begin{aligned} \mathcal{A}(u,v) &= u(1)v(1) + \int_0^1 (u'v' + uv)\mathrm{d}x\,, \\ \ell(v) &= 2v(1) + \int_0^1 fv\mathrm{d}x\,. \end{aligned}$$

Uniqueness of weak solution. In each of the three examples, the weak formulation of the problem has the same general form: find $u \in V$ such that $\mathcal{A}(u,v) = \ell(v)$ for all $v \in V$, with a suitable choice of the space V (namely, $V = \mathrm{H}^1_0(0,1)$ in (a), and $V = \mathrm{H}^1_{\mathrm{E}_0}(0,1)$ in parts (b) and (c)).

Suppose that there are two weak solutions u and $\tilde{u}$. Then, on subtraction, $\mathcal{A}(u - \tilde{u}, v) = 0$ for all $v \in V$. As $u - \tilde{u}$ is an element of V, we can take $v = u - \tilde{u}$, which gives

$$\mathcal{A}(u - \tilde{u}, u - \tilde{u}) = 0\,.$$

In (a) and (b) this means

$$\int_0^1 \left(|(u - \tilde{u})'|^2 + |u - \tilde{u}|^2\right) \mathrm{d}x = 0\,,$$

and in part (c)

$$|u(1)|^2 + \int_0^1 \left(|(u - \tilde{u})'|^2 + |u - \tilde{u}|^2\right) \mathrm{d}x = 0\,,$$

both of which imply that $u - \tilde{u} \equiv 0$, *i.e.*, $u \equiv \tilde{u}$, and hence uniqueness.

Solution to Exercise 14.3

The proof of Theorem 14.4 is identical to the proof of Theorem 14.1 on replacing u by u^h, $\mathrm{H}^1_{\mathrm{E}}(a,b)$ by S^h_{E}, $w \in \mathrm{H}^1_{\mathrm{E}}(a,b)$ by $w^h \in S^h_{\mathrm{E}}$, and $v \in \mathrm{H}^1_0(a,b)$ by $v^h \in S^h_0$ throughout.

Solution to Exercise 14.4

From Corollary 14.1,

$$\begin{aligned}\|u - \mathcal{I}^h u\|_{\mathcal{A}}^2 &\leq \max_{1\leq i\leq n}\left(\frac{h_i}{\pi}\right)^2\left[P_i + \left(\frac{h_i}{\pi}\right)^2 R_i\right]\sum_{i=1}^{n}\|u''\|_{L^2(x_{i-1},x_i)}^2 \\ &\leq \frac{h^2}{\pi^2}\left(P + \frac{h^2}{\pi^2}R\right)\|u''\|_{L^2(a,b)}^2 .\end{aligned}$$

By Ceá's Lemma,

$$\|u - u^h\|_{\mathcal{A}} \leq \|u - \mathcal{I}^h u\|_{\mathcal{A}} \leq \frac{h}{\pi}\left(P + \frac{h^2}{\pi^2}R\right)^{1/2}\|u''\|_{\mathrm{L}^2(a,b)} .$$

Solution to Exercise 14.5

The finite element approximation of the boundary value problem is: find $u^h \in S_0^h$ such that

$$p_0 \int_0^1 (u^h)'(v^h)' \mathrm{d}x + r_0 \int_0^1 u^h v^h \mathrm{d}x = \int_0^1 f v^h \mathrm{d}x \qquad \forall v^h \in S_0^h$$

where $S_0^h = \mathrm{span}\{\varphi_1, \ldots, \varphi_{n-1}\}$. Seek

$$u^h(x) = \sum_{j=1}^{n-1} U_j \varphi_j(x)$$

and choose $v^h = \varphi_i$ for $i = 1, 2, \ldots, n-1$; hence,

$$p_0 \sum_{j=1}^{n-1} U_j \int_0^1 \varphi_j' \varphi_i' \mathrm{d}x + r_0 \sum_{j=1}^{n-1} U_j \int_0^1 \varphi_j \varphi_i \mathrm{d}x = \int_0^1 f \varphi_i \mathrm{d}x\,, \qquad i = 1, 2, \ldots, n-1\,.$$

For $|i-j| > 1$ each of the integrals on the left-hand side of this identity is equal to 0, since the supports of φ_j and φ_i are then disjoint. For $i = j-1, j, j+1$, after calculating the integrals involved we have

$$-p_0 \frac{U_{i-1} - 2U_i + U_{i+1}}{h} + r_0 h \frac{U_{i-1} + 4U_i + U_{i+1}}{6} = \int_0^1 f \varphi_i \mathrm{d}x\,, \quad i = 1, 2, \ldots, n-1\,,$$

and we put $U_0 = 0$ and $U_n = 0$, given that $u^h(0) = 0$ and $u^h(1) = 0$. Equivalently,

$$-p_0 \frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} + r_0 \frac{U_{i-1} + 4U_i + U_{i+1}}{6} = \frac{1}{h} \int_0^1 f \varphi_i \mathrm{d}x\,, \quad i = 1, 2, \ldots, n-1\,,$$

with $U_0 = 0$ and $U_n = 0$.

Let us expand f into a Taylor series with remainder:

$$f(x) = f(x_i) + (x - x_i) f'(x_i) + \frac{1}{2}(x - x_i)^2 f''(x_i) + \frac{1}{6}(x - x_i)^3 f'''(x_i) + \mathcal{O}((x - x_i)^4)\,.$$

Hence,

$$\begin{aligned}
\frac{1}{h} \int_0^1 f(x) \varphi_i(x) \mathrm{d}x &= \frac{1}{h} \int_{x_{i-1}}^{x_{i+1}} f(x) \varphi_i(x) \mathrm{d}x \\
&= \frac{1}{h} \int_{x_{i-1}}^{x_{i+1}} f(x_i) \varphi_i(x) \mathrm{d}x + \frac{1}{2h} \int_{x_{i-1}}^{x_{i+1}} (x - x_i)^2 f''(x_i) \varphi_i(x) \mathrm{d}x + \mathcal{O}(h^4) \\
&= f(x_i) + \frac{1}{12} h^2 f''(x_i) + \mathcal{O}(h^4)\,.
\end{aligned}$$

The finite difference equations arising from the finite element method are

$$-p_0\frac{U_{i-1}-2U_i+U_{i+1}}{h^2}+r_0\frac{U_{i-1}+4U_i+U_{i+1}}{6}=f(x_i)+\frac{1}{12}h^2f''(x_i)\,,\quad i=1,2,\ldots,n-1\,,$$

with $U_0=0$ and $U_n=0$. The truncation error of the scheme is

$$T_i=-p_0\frac{u_{i-1}-2u_i+u_{i+1}}{h^2}+r_0\frac{u_{i-1}+4u_i+u_{i+1}}{6}-f(x_i)-\frac{1}{12}h^2f''(x_i)\,,\quad i=1,2,\ldots,n-1\,,$$

where $u_i=u(x_i)$ for $i=0,1,\ldots,n$. Therefore, after Taylor series expansion,

$$\begin{aligned}
T_i &= -p_0\left(u''(x_i)+\frac{1}{12}h^2u^{IV}(x_i)+\mathcal{O}(h^4)\right)\\
&\quad +r_0\left(u(x_i)+\frac{1}{6}h^2u''(x_i)+\mathcal{O}(h^4)\right)\\
&\quad -f(x_i)-\frac{1}{12}h^2f''(x_i)\\
&= -\frac{1}{12}h^2p_0u^{IV}(x_i)+\frac{1}{6}h^2r_0u''(x_i)-\frac{1}{12}h^2\left(-p_0u^{IV}(x_i)+r_0u''(x_i)\right)+\mathcal{O}(h^4)\\
&= \frac{1}{12}h^2r_0u''(x_i)+\mathcal{O}(h^4)\,.
\end{aligned}$$

We define the global error $e_i=u(x_i)-U_i=u(x_i)-u^h(x_i)$. Then,

$$-p_0\frac{e_{i-1}-2e_i+e_{i+1}}{h^2}+r_i\frac{e_{i-1}+4e_i+e_{i+1}}{6}=T_i\,,\quad i=1,2,\ldots,n-1\,,$$

with $e_0=0$, $e_n=0$. As in the proof of Theorem 13.4 we have, with $T=\max_{1\le i\le n}|T_i|$, that

$$\begin{aligned}
\max_{0\le i\le n}|e_i| &\le \frac{1}{8}(1-0)^2T\\
&\le \frac{1}{8}\left(\frac{1}{12}h^2r_0\max_{1\le i\le n-1}|u''(x_i)|+\mathcal{O}(h^4)\right)\\
&= \frac{1}{96}h^2r_0\max_{1\le i\le n-1}|u''(x_i)|+\mathcal{O}(h^4)\\
&\le Mh^2\,,
\end{aligned}$$

where M is a positive constant, $M>\frac{1}{96}h^2r_0\max_{1\le i\le n-1}|u''(x_i)|$.

Solution to Exercise 14.6

The difference equation arising from the finite element method is

$$-p_0\frac{U_{i-1}-2U_i+U_{i+1}}{h}+r_0\sum_{j=1}^{n-1}U_j\int_{x_{i-1}}^{x_{i+1}}\varphi_j\varphi_i\mathrm{d}x=\int_{x_{i-1}}^{x_{i+1}}f\varphi_i\mathrm{d}x\,,\quad i=1,2,\ldots,n-1\,,$$

with $U_0=0$, $U_n=0$. Now, using the trapezium rule,

$$\int_{x_{i-1}}^{x_{i+1}}\varphi_{i\pm1}\varphi_i\mathrm{d}x=\int_{x_{i-1}}^{x_i}\varphi_{i\pm1}\varphi_i\mathrm{d}x+\int_{x_i}^{x_{i+1}}\varphi_{i\pm1}\varphi_i\mathrm{d}x\approx\frac{1}{2}h(0+0)+\frac{1}{2}h(0+0)=0\,.$$

Similarly,

$$\int_{x_{i-1}}^{x_{i+1}}\varphi_i^2\mathrm{d}x=\int_{x_{i-1}}^{x_i}\varphi_i^2\mathrm{d}x+\int_{x_i}^{x_{i+1}}\varphi_i^2\mathrm{d}x\approx\frac{1}{2}h(0+1)+\frac{1}{2}h(1+0)=h\,,$$

and

$$\int_{x_{i-1}}^{x_{i+1}}f\varphi_i\mathrm{d}x\approx\frac{1}{2}h(0+f(x_i))+\frac{1}{2}h(f(x_i)+0)=hf(x_i)\,.$$

The difference scheme therefore becomes

$$-p_0\frac{U_{i-1}-2U_i+U_{i+1}}{h}+r_0hU_i=hf(x_i)\,,\quad i=1,2,\ldots,n-1,$$

with $U_0=0$ and $U_n=0$. Equivalently,

$$-p_0\frac{U_{i-1}-2U_i+U_{i+1}}{h^2}+r_0U_i=f(x_i)\,,\quad i=1,2,\ldots,n-1,$$

with $U_0=0$ and $U_n=0$, which is identical to the central difference approximation from Chapter 13. From Theorem 13.4, and noting that $U_i=u^h(x_i)$, we then have that

$$\max_{0\le i\le n}|u(x_i)-u^h(x_i)|\le\frac{1}{96}h^2M_4=\mathcal{O}(h^2)\,.$$

Solution to Exercise 14.7

To establish the weak formulation, we multiply the differential equation by a function $v \in \mathrm{H}^1(a,b)$ and integrate by parts in the term involving $(pu')'v$; hence,

$$\int_a^b (pu'v' + ruv)\mathrm{d}x - p(x)u'(x)v(x)|_{x=a}^{x=b} = \int_a^b fv\mathrm{d}x \qquad \forall v \in \mathrm{H}^1(a,b)\,.$$

Replacing $p(a)u'(a)$ by $\alpha u(a) - A$ and $p(b)u'(b)$ by $B - \beta u(b)$, we have

$$\int_a^b (pu'v'+ruv)\mathrm{d}x+\alpha u(a)v(a)+\beta u(b)v(b) = Av(a)+Bv(b)+\int_a^b fv\mathrm{d}x \quad \forall v \in \mathrm{H}^1(a,b)\,.$$

We thus define

$$\begin{aligned} \mathcal{A}(w,v) &= \int_a^b (pw'v' + rwv)\mathrm{d}x + \alpha w(a)v(a) + \beta w(b)v(b) \\ \ell(v) &= Av(a) + Bv(b) + \int_a^b fv\mathrm{d}x\,. \end{aligned}$$

The weak formulation of the problem is: find $u \in \mathrm{H}^1(a,b)$ such that $\mathcal{A}(u,v) = \ell(v)$ for all $v \in \mathrm{H}^1(a,b)$.

The finite element approximation of the boundary value problem is: find $u^h \in S^h$ such that $\mathcal{A}(u^h, v^h) = \ell(v^h)$ for all $v^h \in S^h$, where $S^h = \mathrm{span}\{\varphi_0, \ldots, \varphi_n\}$.

Writing $u^h(x) = \sum_{j=0}^n U_j\varphi_j(x)$ and choosing $v^h = \varphi_i$, $i = 0, 1, \ldots, n$, we get a system of $n+1$ linear equations for the $n+1$ unknowns $U_0, U_1, \ldots, U_n$.

Uniqueness follows by noting that

$$\mathcal{A}(v^h, v^h) = \int_a^b p(x)|(v^h)'|^2 + r(x)|v^h|^2\mathrm{d}x + \alpha|v^h(a)|^2 + \beta|v^h(b)|^2\,.$$

Thus, if $\mathcal{A}(v^h, v^h) = 0$ then $v^h \equiv 0$ on $[a,b]$. Hence, if u^h and $\tilde{u}^h$ are both solutions of the finite element approximation, then $\mathcal{A}(u^h - \tilde{u}^h, v^h) = 0$ for all $v^h \in S^h$. With $v^h = u^h - \tilde{u}^h$, we have that $\mathcal{A}(u^h - \tilde{u}^h, u^h - \tilde{u}^h) = 0$, and therefore $u^h - \tilde{u}^h \equiv 0$ on $[a,b]$.

Let us now write down the system of linear equations. The (i,j) entry M_{ij} of the matrix M of the linear system is

$$M_{ij} = \mathcal{A}(\varphi_j, \varphi_i) = \int_{x_{i-1}}^{x_{i+1}} (p(x)\varphi_j'\varphi_i' + r(x)\varphi_j\varphi_i)\mathrm{d}x$$

for $i, j = 1, \ldots, n-1$. For $i = 0$,

$$M_{0j} = \mathcal{A}(\varphi_j, \varphi_0) = \int_0^{x_1} (p(x)\varphi_j'\varphi_0' + r(x)\varphi_j\varphi_0)\mathrm{d}x + \alpha\varphi_j(a)\,, \quad j = 0, 1, \ldots, n\,.$$

Note that $\varphi_j(a) = 0$ unless $j = 0$.

For $i = n$,

$$M_{nj} = \mathcal{A}(\varphi_j, \varphi_n) = \int_{x_{n-1}}^{x_n} (p(x)\varphi_j'\varphi_n' + r(x)\varphi_j\varphi_n)\mathrm{d}x + \beta\varphi_j(b)\,, \quad j = 0, 1, \ldots, n\,.$$

We note that $\varphi_j(b) = 0$ unless $j = n$.

The matrix M is tridiagonal. Since, $\mathcal{A}(\varphi_i, \varphi_j) = \mathcal{A}(\varphi_j, \varphi_i)$, the matrix M is symmetric.

Let $v^h(x) = \sum_{i=0}^n V_i\varphi_i(x)$ and $\mathbf{V} = (V_0, \ldots, V_n)^T$. Then,

$$\mathbf{V}^T M \mathbf{V} = (V_0, \ldots, V_n)^T M \begin{pmatrix} V_0 \\ \vdots \\ V_n \end{pmatrix} = \sum_{j=0}^n \sum_{i=0}^n \mathcal{A}(\varphi_j, \varphi_i) V_j V_i = \mathcal{A}(v^h, v^h)$$

$$= \int_0^1 (p(x)|(v^h)'|^2 + r(x)|v^h|^2)\mathrm{d}x + \alpha|v^h(a)|^2 + \beta|v^h(b)|^2 > 0$$

unless $v_h \equiv 0$, *i.e.*, $\mathbf{V} = \mathbf{0}$ in $\mathbb{R}^{n+1}$. Therefore, the matrix M of the linear system is positive definite.

Solution to Exercise 14.8

The weak formulation of the boundary value problem is: find $u \in \mathrm{H}^1_{\mathrm{E}_0}(0,1) = \{v \in \mathrm{H}^1(0,1) \,:\, v(0) = 0\}$ such that

$$\mathcal{A}(u,v) \equiv \int_0^1 (u'v' + uv)\mathrm{d}x + \alpha u(1)v(1) = \int_0^1 fv\mathrm{d}x \equiv \ell(v)$$

for all $v \in \mathrm{H}^1_{\mathrm{E}_0}(0,1)$.

Let $S^h_{\mathrm{E}_0} = \mathrm{span}\{\varphi_1, \ldots, \varphi_n\}$. The finite element approximation of the boundary value problem is: find $u^h \in S^h_{\mathrm{E}_0}$ such that $\mathcal{A}(u^h, v^h) = \ell(v^h)$ for all v^h in $S^h_{\mathrm{E}_0}$. Let us write $u^h(x) = \sum_{j=1}^n U_j \varphi_j(x)$ and take $v^h = \varphi_i$, $i = 1, 2, \ldots, n$. Then, for the case of $\alpha = 0$ and $f(x) \equiv 1$, we obtain the following difference equations: $U_0 = 0$,

$$-\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} + \frac{U_{i-1} + 4U_i + U_{i+1}}{6} = \frac{1}{h}\int_{x_{i-1}}^{x_{i+1}} \varphi_i(x)\mathrm{d}x\,, \quad i = 1, 2, \ldots, n-1\,,$$

and, for $i = n$, we have

$$U_{n-1}\mathcal{A}(\varphi_{n-1}, \varphi_n) + U_n \mathcal{A}(\varphi_n, \varphi_n) = \int_{x_{n-1}}^{x_n} \varphi_n(x)\mathrm{d}x\,.$$

Therefore, $U_0 = 0$,

$$-\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} + \frac{U_{i-1} + 4U_i + U_{i+1}}{6} = 1\,, \quad i = 1, 2, \ldots, n-1\,,$$

and

$$U_{n-1}\left(-\frac{1}{h^2} + \frac{1}{6}\right) + U_n\left(\frac{1}{h^2} + \frac{1}{3}\right) = \frac{1}{2}\,.$$

For $n = 3$, we have $h = 1/3$ and $1/h^2 = 9$ and therefore,

$$\begin{aligned}
\left(18 + \tfrac{4}{6}\right) U_1 + \left(-9 + \tfrac{1}{6}\right) U_2 &= 1 \\
\left(-9 + \tfrac{1}{6}\right) U_1 + \left(18 + \tfrac{4}{6}\right) U_2 + \left(-9 + \tfrac{1}{6}\right) U_3 &= 1 \\
\left(-9 + \tfrac{1}{6}\right) U_2 + \left(9 + \tfrac{2}{6}\right) U_3 &= \tfrac{1}{2}\,.
\end{aligned}$$

Solving this yields $U_1 = 0.2039$, $U_2 = 0.3177$, $U_3 = 0.3543$, together with $U_0 = 0$.

Solution to Exercise 14.9

The energy norm $\|\cdot\|_{\mathcal{A}}$ is defined by

$$\|v\|_{\mathcal{A}} = \left(\int_0^1 p(x)|v'|^2 + r(x)|v|^2 \mathrm{d}x \right)^{1/2} .$$

By Ceá's Lemma, $\|u - u^h\|_{\mathcal{A}} \leq \|u - \mathcal{I}^h u\|_{\mathcal{A}}$. Therefore,

$$c_0 \int_0^1 |(u-u^h)'|^2 \mathrm{d}x \leq P \int_0^1 |(u - \mathcal{I}^h u)'|^2 \mathrm{d}x + R \int_0^1 |u - \mathcal{I}^h u|^2 \mathrm{d}x .$$

This yields

$$c_0 \int_0^1 |(u-u^h)'|^2 \mathrm{d}x \leq \left(P\frac{h^2}{\pi^2} + R\frac{h^4}{\pi^4} \right) \|u''\|^2_{\mathrm{L}^2(0,1)} ,$$

that is,

$$\int_0^1 |(u-u^h)'|^2 \mathrm{d}x \leq \frac{h^2}{\pi^2 c_0} \left(P + R\frac{h^2}{\pi^2} \right) \|u''\|^2_{\mathrm{L}^2(0,1)} .$$

By the Poincaré–Friedrichs inequality [cf. Exercise 1],

$$\int_0^1 |(u-u^h)|^2 \mathrm{d}x \leq \frac{h^2}{2\pi^2 c_0} \left(P + R\frac{h^2}{\pi^2} \right) \|u''\|^2_{\mathrm{L}^2(0,1)} .$$

By adding the last two inequalities, we have

$$\|u - u^h\|^2_{\mathrm{H}^1(0,1)} \leq \frac{3h^2}{2\pi^2 c_0} \left(P + R\frac{h^2}{\pi^2} \right) \|u''\|^2_{\mathrm{L}^2(0,1)} .$$

Therefore,

$$\|u - u^h\|_{\mathrm{H}^1(0,1)} \leq C_1 h \|u''\|_{\mathrm{L}^2(0,1)} .$$

where

$$C_1 = \frac{1}{\pi} \sqrt{\frac{3}{2c_0} \left(P + R\frac{1}{\pi^2} \right)} .$$

Let us now bound $\|u''\|_{\mathrm{L}^2(0,1)}$ by $\|f\|_{\mathrm{L}^2(0,1)}$. Multiplying the differential equation by u, integrating by parts and using the boundary conditions $u(0) = 0$ and $u(1) = 0$ yields

$$\int_0^1 (p(x)|u'|^2 + r(x)|u|^2) \mathrm{d}x = \int_0^1 f(x)u(x) \mathrm{d}x .$$

Therefore, by the Cauchy–Schwarz inequality,

$$c_0 \|u'\|^2_{\mathrm{L}^2(0,1)} \leq \|f\|_{\mathrm{L}^2(0,1)} \|u\|_{\mathrm{L}^2(0,1)} .$$

According to the Poincare–Friedrichs inequality,

$$\|u\|_{\mathrm{L}^2(0,1)} \leq \frac{1}{\sqrt{2}} \|u'\|_{\mathrm{L}^2(0,1)} .$$

Hence,

$$\|u'\|_{\mathrm{L}^2(0,1)} \leq \frac{1}{c_0\sqrt{2}} \|f\|_{\mathrm{L}^2(0,1)} ,$$

which then yields

$$\|u\|_{\mathrm{L}^2(0,1)} \leq \frac{1}{2c_0} \|f\|_{\mathrm{L}^2(0,1)} .$$

Summing the squares of the last two inequalities and taking the square–root gives

$$\|u\|_{\mathrm{H}^1(0,1)} = (\|u'\|^2_{\mathrm{L}^2(0,1)} + \|u\|^2_{\mathrm{L}^2(0,1)})^{1/2} \leq \frac{\sqrt{3}}{2c_0} \|f\|_{\mathrm{L}^2(0,1)} .$$

Now we use this to bound $\|u''\|_{\mathrm{L}^2(0,1)}$. First, observe that from the differential equation,

$$u'' = -\frac{p'}{p} u' + \frac{r}{p} u - \frac{f}{p} .$$

Therefore,

$$\begin{aligned}
\|u''\|_{\mathrm{L}^2(0,1)} &\leq \|\frac{p'}{p}\|_\infty \|u'\|_{\mathrm{L}^2(0,1)} + \|\frac{r}{p}\|_\infty \|u\|_{\mathrm{L}^2(0,1)} + \|\frac{1}{p}\|_\infty \|f\|_{\mathrm{L}^2(0,1)} \\
&\leq \left(\|\frac{p'}{p}\|^2_\infty + \|\frac{r}{p}\|^2_\infty \right)^{1/2} (\|u'\|^2_{\mathrm{L}^2(0,1)} + \|u\|^2_{\mathrm{L}^2(0,1)})^{1/2} + \|\frac{1}{p}\|_\infty \|f\|_{\mathrm{L}^2(0,1)} \\
&\leq \left(\left(\|\frac{p'}{p}\|^2_\infty + \|\frac{r}{p}\|^2_\infty \right)^{1/2} \frac{\sqrt{3}}{2c_0} + \|\frac{1}{p}\|_\infty \right) \|f\|_{\mathrm{L}^2(0,1)} .
\end{aligned}$$

Letting

$$C_2 = \left(\|\frac{p'}{p}\|^2_\infty + \|\frac{r}{p}\|^2_\infty \right)^{1/2} \frac{\sqrt{3}}{2c_0} + \|\frac{1}{p}\|_\infty$$

we then deduce that

$$\|u''\|_{\mathrm{L}^2(0,1)} \leq C_2 \|f\|_{\mathrm{L}^2(0,1)} .$$

Hence,

$$\|u - u^h\|_{\mathrm{H}^1(0,1)} \leq Ch \|f\|_{\mathrm{L}^2(0,1)} ,$$

with $C = C_1 C_2$ and C_1 an C_2 as defined above.

In the special case when $p(x) \equiv 1$, $r(x) \equiv 0$, and $f(x) \equiv 1$, we have

$$C_1 = \frac{1}{\pi}\sqrt{\frac{3}{2}} \quad \text{and} \quad C_2 = 1\,,$$

which yields

$$\|u - u^h\|_{\mathrm{H}^1(0,1)} \leq \frac{h}{\pi}\sqrt{\frac{3}{2}}\,.$$

With $h = 10^{-3}$, we then have

$$\|u - u^h\|_{\mathrm{H}^1(0,1)} \leq \frac{10^{-3}}{\pi}\sqrt{\frac{3}{2}}\,.$$

Solution to Exercise 14.10

When the trapezium rule is used to approximate the integral $\int_0^1 fv^h \mathrm{d}x$ in the finite element method, the method becomes: find $u^h \in S_0^h$ such that

$$\mathcal{A}(u^h, v^h) \equiv \int_0^1 ((u^h)'(v^h)' + u^h v^h)\mathrm{d}x = \int_0^1 \mathcal{I}^h(fv^h)\mathrm{d}x \equiv \ell^h(v^h)\,, \quad v^h \in S_0^h\,.$$

The argument now proceeds in much the same way as discussed in Section 14.5 in the case when the integral $\ell(v^h) = \int_0^1 fv^h \mathrm{d}x$ is computed exactly. We define the dual problem

$$-z'' + z = u - u^h\,, \quad x \in (0,1)\,, \qquad z(0) = 0\,, \quad z(1) = 0\,.$$

Hence,

$$\begin{aligned}
\|u - u^h\|^2_{\mathrm{L}^2(0,1)} &= \langle u - u^h, -z'' + z\rangle \\
&= \mathcal{A}(u - u^h, z) = \mathcal{A}(u - u^h, z - \mathcal{I}^h z) + \mathcal{A}(u - u^h, \mathcal{I}^h z) \\
&= \mathcal{A}(u - u^h, z - \mathcal{I}^h z) + [\ell(\mathcal{I}^h z) - \ell^h(\mathcal{I}^h z)] \\
&= \mathcal{A}(u - u^h, z - \mathcal{I}^h z) + \int_0^1 [f(\mathcal{I}^h z) - \mathcal{I}^h(f\mathcal{I}^h z)]\mathrm{d}x \\
&\equiv \mathrm{T}_1 + \mathrm{T}_2\,.
\end{aligned}$$

Term T_1 is bounded in exactly the same way as shown in Section 14.5:

$$\mathrm{T}_1 \le K_0 \left(\sum_{i=1}^n h_i^4 \|R(u^h)\|^2_{\mathrm{L}^2(x_{i-1},x_i)}\right)^{1/2} \|u - u^h\|_{\mathrm{L}^2(0,1)}\,,$$

where $K_0 = 2/\pi^2$.

As for T_2,

$$\begin{aligned}
(\mathrm{T}_2)^2 &\le \left(\int_0^1 |f(\mathcal{I}^h z) - \mathcal{I}^h(f\mathcal{I}^h z)|\mathrm{d}x\right)^2 \\
&\le \int_0^1 |f(\mathcal{I}^h z) - \mathcal{I}^h(f\mathcal{I}^h z)|^2 \mathrm{d}x \\
&\le \sum_{i=1}^n \frac{h_i^4}{\pi^4} \|(f\mathcal{I}^h z)''\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \\
&= \sum_{i=1}^n \frac{h_i^4}{\pi^4} \|f''(\mathcal{I}^h z) + 2f'(\mathcal{I}^h z)'\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \\
&\le \sum_{i=1}^n \frac{h_i^4}{\pi^4} \left(\max_{x\in[x_{i-1},x_i]} |f''|\, \|\mathcal{I}^h z\|_{\mathrm{L}^2(x_{i-1},x_i)} + 2 \max_{x\in[x_{i-1},x_i]} |f'|\, \|(\mathcal{I}^h z)'\|_{\mathrm{L}^2(x_{i-1},x_i)}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \frac{h_i^4}{\pi^4} \left(\max_{x\in[x_{i-1},x_i]} |f''|^2 + 4 \max_{x\in[x_{i-1},x_i]} |f'|^2 \right) \\
&\quad \times \left(\|\mathcal{I}^h z\|^2_{\mathrm{L}^2(x_{i-1},x_i)} + \|(\mathcal{I}^h z)'\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \right) \\
&\leq \max_{1\leq i\leq n} \frac{h_i^4}{\pi^4} \left(\max_{x\in[x_{i-1},x_i]} |f''|^2 + 4 \max_{x\in[x_{i-1},x_i]} |f'|^2 \right) \\
&\quad \times \sum_{i=1}^{n} \|\mathcal{I}^h z\|^2_{\mathrm{L}^2(x_{i-1},x_i)} + \|(\mathcal{I}^h z)'\|^2_{\mathrm{L}^2(x_{i-1},x_i)} \\
&= K_2^2 \|\mathcal{I}^h z\|^2_{\mathrm{H}^1(0,1)},
\end{aligned}
$$

where

$$
K_2 = \max_{1\leq i\leq n} \frac{h_i^2}{\pi^2} \left(\max_{x\in[x_{i-1},x_i]} |f''|^2 + 4 \max_{x\in[x_{i-1},x_i]} |f'|^2 \right)^{1/2}.
$$

Now,

$$
\begin{aligned}
\|\mathcal{I}^h z\|_{\mathrm{H}^1(0,1)} &= \|\mathcal{I}^h z - z + z\|_{\mathrm{H}^1(0,1)} \leq \|\mathcal{I}^h z - z\|_{\mathrm{H}^1(0,1)} + \|z\|_{\mathrm{H}^1(0,1)} \\
&\leq \frac{h}{\pi} \left(1 + \frac{h^2}{\pi^2}\right)^{1/2} \|z''\|_{\mathrm{L}^2(0,1)} + \|z\|_{\mathrm{H}^1(0,1)}.
\end{aligned}
$$

However, $-z'' + z = u - u^h$, and therefore

$$
\|z''\|_{\mathrm{L}^2(0,1)} \leq \|z\|_{\mathrm{L}^2(0,1)} + \|u - u^h\|_{\mathrm{L}^2(0,1)}.
$$

As

$$
\|z'\|^2_{\mathrm{L}^2(0,1)} + \|z\|^2_{\mathrm{L}^2(0,1)} = \langle u - u^h, z\rangle \leq \|u - u^h\|_{\mathrm{L}^2(0,1)} \|z\|_{\mathrm{L}^2(0,1)},
$$

we conclude that

$$
\|z\|_{\mathrm{L}^2(0,1)} \leq \|u - u^h\|_{\mathrm{L}^2(0,1)},
$$

and then also,

$$
\|z'\|^2_{\mathrm{L}^2(0,1)} + \|z\|^2_{\mathrm{L}^2(0,1)} \leq \|u - u^h\|^2_{\mathrm{L}^2(0,1)}.
$$

Hence

$$
\|z''\|_{\mathrm{L}^2(0,1)} \leq 2\|u - u^h\|_{\mathrm{L}^2(0,1)},
$$

and so

$$
\|\mathcal{I}^h z\|_{\mathrm{H}^1(0,1)} \leq \left[\frac{2h}{\pi} \left(1 + \frac{h^2}{\pi^2}\right)^{1/2} + 1 \right] \|u - u^h\|_{\mathrm{L}^2(0,1)}.
$$

Consequently,

$$\mathrm{T}_2 \leq K_2 \|\mathcal{I}^h z\|_{\mathrm{H}^1(0,1)} \leq K_2 \left[\frac{2h}{\pi} \left(1 + \frac{h^2}{\pi^2} \right)^{1/2} + 1 \right] \|u - u^h\|_{\mathrm{L}^2(0,1)} .$$

Therefore,

$$\mathrm{T}_2 \leq K_1 \max_{1 \leq i \leq n} h_i^2 \left(\max_{x \in [x_{i-1}, x_i]} |f''|^2 + 4 \max_{x \in [x_{i-1}, x_i]} |f'|^2 \right)^{1/2} \|u - u^h\|_{\mathrm{L}^2(0,1)} ,$$

where

$$K_1 = \frac{1}{\pi^2} \left[\frac{2h}{\pi} \left(1 + \frac{h^2}{\pi^2} \right)^{1/2} + 1 \right] \|u - u^h\|_{\mathrm{L}^2(0,1)} .$$

Using the estimates of T_1 and T_2 yields the required *a posteriori* error bound.

Given a positive tolerance `TOL`, the mesh adaptation algorithm proceeds as discussed in Section 14.5.