Walter Gautschi

Solutions Manual to Numerical Analysis

2d edition

EXERCISES AND MACHINE ASSIGNMENTS TO CHAPTER 1

EXERCISES

1. Represent all elements of $\mathbb{R}_+(3,2) = \{x \in \mathbb{R}(3,2) : \ x > 0, \ x \text{ normalized}\}$ as dots on the real axis. For clarity, draw two axes, one from 0 to 8, the other from 0 to $\frac{1}{2}$.

2. (a) What is the distance $d(x)$ of a positive normalized floating-point number $x \in \mathbb{R}(t,s)$ to its next larger floating-point number:

$$d(x) = \min_{\substack{y \in \mathbb{R}(t,s) \\ y > x}} (y - x) \, ?$$

   (b) Determine the relative distance $r(x) = d(x)/x$, with x as in (a), and give upper and lower bounds for it.

3. The identity $\mathrm{fl}(1+x) = 1$, $x \geq 0$, is true for $x = 0$ and for x sufficiently small. What is the largest machine number x for which the identity still holds?

4. Consider a miniature binary computer whose floating-point words consist of 4 binary digits for the mantissa and 3 binary digits for the exponent (plus sign bits). Let

$$x = (.1011)_2 \times 2^0, \qquad y = (.1100)_2 \times 2^0.$$

   Mark in the following table whether the machine operation indicated (with the result z assumed normalized) is exact, rounded (i.e., subject to a nonzero rounding error), overflows, or underflows.

operation	exact	rounded	overflow	underflow
$z = \mathrm{fl}(x - y)$				
$z = \mathrm{fl}((y - x)^{10})$				
$z = \mathrm{fl}(x + y)$				
$z = \mathrm{fl}(y + (x/4))$				
$z = \mathrm{fl}(x + (y/4))$				

5. The Matlab "machine precision" `eps` is twice the unit roundoff (2×2^{-t}, $t = 53$; cf. Sect. 1.1.3). It can be computed by the following Matlab program (attributed to Cleve Moler):

```
%EI_5  Matlab machine precision
%
a=4/3;
b=a-1;
c=b+b+b;
eps0=abs(c-1)
```

Run the program and prove its validity.

6. Prove (1.12).

7. A set S of real numbers is said to possess a metric if there is defined a distance function $d(x, y)$ for any two elements $x, y \in S$ that has the following properties:

   (i) $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$ (positive definiteness);

   (ii) $d(x, y) = d(y, x)$ (symmetry);

   (iii) $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality).

   Discuss which of the following error measures is, or is not, a distance function on what set S of real numbers:

   (a) absolute error: $\quad \mathrm{ae}(x, y) = |x - y|$;

   (b) relative error: $\quad \mathrm{re}(x, y) = \left|\frac{x-y}{x}\right|$;

   (c) relative precision (F.W.J. Olver, 1978): $\quad \mathrm{rp}(x, y) = |\ \ln|x| - \ln|y|\ |$.

   If $y = x(1 + \varepsilon)$, show that $\mathrm{rp}(x, y) = O(\varepsilon)$ as $\varepsilon \to 0$.

8. Assume that x_1^*, x_2^* are approximations to x_1, x_2 with relative errors E_1 and E_2, respectively, and that $|E_i| \leq E$, $i = 1, 2$. Assume further that $x_1 \neq x_2$.

   (a) How small must E (in dependence of x_1 and x_2) be in order to ensure that $x_1^* \neq x_2^*$?

   (b) Taking $\dfrac{1}{x_1^* - x_2^*}$ to approximate $\dfrac{1}{x_1 - x_2}$, obtain a bound on the relative error committed, assuming (1) exact arithmetic; (2) machine arithmetic with machine precision eps. (In both cases, neglect higher-order terms in E_1, E_2, eps.)

9. Consider the quadratic equation $x^2 + px + q = 0$ with roots x_1, x_2. As seen in the second Example of Sect. 1.2.2, the absolutely larger root must be computed first, whereupon the other can be accurately obtained from $x_1 x_2 = q$. Suppose one incorporates this idea in a program such as

```
x1=abs(p/2)+sqrt(p*p/4-q);
if p>0, x1=-x1; end
x2=q/x1;
```

   Find two serious flaws with this program as a "general-purpose quadratic equation solver." Take into consideration that the program will be executed in floating-point machine arithmetic. Be specific and support your arguments by examples, if necessary.

10. Suppose, for $|x|$ small, one has an accurate value of $y = \mathrm{e}^x - 1$ (obtained, e.g., by Taylor expansion). Use this value to compute accurately $\sinh x = \frac{1}{2}(\mathrm{e}^x - \mathrm{e}^{-x})$ for small $|x|$.

11. Let $f(x) = \sqrt{1 + x^2} - 1$.

    (a) Explain the difficulty of computing $f(x)$ for a small value of $|x|$ and show how it can be circumvented.

    (b) Compute $(\mathrm{cond}\, f)(x)$ and discuss the conditioning of $f(x)$ for small $|x|$.

    (c) How can the answers to (a) and (b) be reconciled?

12. The nth power of some positive (machine) number x can be computed

    (i) either by repeated multiplication by x, or

    (ii) as $x^n = \mathrm{e}^{n \ln x}$.

    In each case, derive bounds for the relative error due to machine arithmetic, neglecting higher powers of the machine precision against the first power. (Assume that exponentiation and taking logarithms both involve a relative error ε with $|\varepsilon| \leq \mathrm{eps}$.) Based on these bounds, state a criterion (involving x and n) for (i) to be more accurate than (ii).

13. Let $f(x) = (1 - \cos x)/x$, $x \neq 0$.

    (a) Show that direct evaluation of f is inaccurate if $|x|$ is small; assume $\mathrm{fl}(f(x)) = \mathrm{fl}((1 - \mathrm{fl}(\cos x))/x)$, where $\mathrm{fl}(\cos x) = (1 + \varepsilon_c)\cos x$, and estimate the relative error ε_f of $\mathrm{fl}(f(x))$ as $x \to 0$.

    (b) A mathematically equivalent form of f is $f(x) = \sin^2 x/(x(1 + \cos x))$. Carry out a similar analysis as in (a), based on $\mathrm{fl}(f(x)) = \mathrm{fl}([\mathrm{fl}(\sin x)]^2/\mathrm{fl}(x(1 + \mathrm{fl}(\cos x))))$, assuming $\mathrm{fl}(\cos x) = (1 + \varepsilon_c)\cos x$, $\mathrm{fl}(\sin x) = (1 + \varepsilon_s)\sin x$ and retaining only first-order terms in ε_s and ε_c. Discuss the result.

    (c) Determine the condition of $f(x)$. Indicate for what values of x (if any) $f(x)$ is ill-conditioned. ($|x|$ is no longer small, necessarily.)

14. If $z = x + \mathrm{i}y$, then $\sqrt{z} = \left(\dfrac{r + x}{2}\right)^{1/2} + \mathrm{i}\left(\dfrac{r - x}{2}\right)^{1/2}$, where $r = (x^2 + y^2)^{1/2}$. Alternatively, $\sqrt{z} = u + \mathrm{i}v$, $u = \left(\dfrac{r + x}{2}\right)^{1/2}$, $v = y/2u$. Discuss the computational merits of these two (mathematically equivalent) expressions. Illustrate with $z = 4.5 + .025\mathrm{i}$, using 8 significant decimal places. {*Hint*: you may assume $x > 0$ without restriction of generality. Why?}

15. Consider the numerical evaluation of

$$f(t) = \sum_{n=0}^{\infty} \frac{1}{1 + n^4(t-n)^2(t-n-1)^2} ,$$

say, for $t = 20$, and 7-digit accuracy. Discuss the danger involved.

16. Let X_+ be the largest positive machine representable number, and X_- the absolute value of the smallest negative one (so that $-X_- \le x \le X_+$ for any machine number x). Determine, approximately, all intervals on $\mathbb{R}$ on which the tangent function overflows.

17. (a) Use Matlab to determine the first value of the integer n for which $n!$ overflows. {*Hint*: use Stirling's formula for $n!$.}

    (b) Do the same as (a), but for x^n, $x = 10, 20, \ldots, 100$.

    (c) Discuss how $x^n \mathrm{e}^{-x}/n!$ can be computed for large x and n without unnecessarily incurring overflow. {*Hint*: use logarithms and an asymptotic formula for $\ln n!$.}

18. Consider a decimal computer with 3 (decimal) digits in the floating-point mantissa.

    (a) Estimate the relative error committed in symmetric rounding.

    (b) Let $x_1 = .982$, $x_2 = .984$ be two machine numbers. Calculate in machine arithmetic the mean $m = \frac{1}{2}(x_1 + x_2)$. Is the computed number between x_1 and x_2 ?

    (c) Derive sufficient conditions for $x_1 < \mathrm{fl}(m) < x_2$ to hold, where x_1, x_2 are two machine numbers with $0 < x_1 < x_2$.

19. For this problem, assume a binary computer with 12 bits in the floating-point mantissa.

    (a) What is the machine precision eps?

    (b) Let $x = 6/7$ and x^* be the correctly rounded machine approximation to x (symmetric rounding). Exhibit x and x^* as binary numbers.

    (c) Determine (exactly!) the relative error ε of x^* as an approximation to x, and calculate the ratio $|\varepsilon|/\mathrm{eps}$.

20. The distributive law of algebra states that

$$(a + b)c = ac + bc.$$

Discuss to what extent this is violated in machine arithmetic. Assume a computer with machine precision eps and assume that a, b, c are machine-representable numbers.

(a) Let y_1 be the floating-point number obtained by evaluating $(a + b)c$ (as written) in floating-point arithmetic, and let $y_1 = (a + b)c(1 + e_1)$. Estimate $|e_1|$ in terms of eps (neglecting second-order terms in eps).

(b) Let y_2 be the floating-point number obtained by evaluating $ac + bc$ (as written) in floating-point arithmetic, and let $y_2 = (a + b)c(1 + e_2)$. Estimate $|e_2|$ (neglecting second-order terms in eps) in terms of eps (and a, b, and c).

(c) Identify conditions (if any) under which one of the two y_i is significantly less accurate than the other.

21. Let $x_1, x_2, \ldots, x_n$, $n > 1$, be machine numbers. Their product can be computed by the algorithm

$$p_1 = x_1,$$
$$p_k = \mathrm{fl}(x_k p_{k-1}), \qquad k = 2, 3, \ldots, n.$$

(a) Find an upper bound for the relative error $(p_n - x_1 x_2 \cdots x_n)/(x_1 x_2 \cdots x_n)$ in terms of the machine precision eps and n.

(b) For any integer $r \geq 1$ not too large so as to satisfy $r \cdot \mathrm{eps} < \frac{1}{10}$, show that

$$(1 + \mathrm{eps})^r - 1 < 1.06 \cdot r \cdot \mathrm{eps}.$$

Hence, for n not too large, simplify the answer given in (a). {*Hint*: use the binomial theorem.}

22. Analyze the error propagation in exponentiation, x^α $(x > 0)$:

(a) assuming x exact and α subject to a small relative error ε_α;

(b) assuming α exact and x subject to a small relative error ε_x.

Discuss the possibility of any serious loss of accuracy.

23. Indicate how you would accurately compute

$$(x + y)^{1/4} - y^{1/4}, \qquad x > 0, \qquad y > 0.$$

24. (a) Let $a = .23371258 \times 10^{-4}$, $b = .33678429 \times 10^{2}$, $c = -.33677811 \times 10^{2}$. Assuming an 8-decimal-digit computer, determine the sum $s = a + b + c$ either as (i) $\mathrm{fl}(s) = \mathrm{fl}(\mathrm{fl}(a + b) + c)$ or as (ii) $\mathrm{fl}(s) = \mathrm{fl}(a + \mathrm{fl}(b + c))$. Explain the discrepancy between the two answers.

(b) For arbitrary machine numbers a, b, c, on a computer with machine precision eps, find a criterion on a, b, c for the result of (ii) in (a) to be more accurate than the result of (i). {*Hint*: Compare bounds on the relative errors, neglecting higher-order terms in eps and assuming $a + b + c \neq 0$; see also MA 7.}

25. Write the expression $a^2 - 2ab\cos\gamma + b^2$ $(a > 0, b > 0)$ as the sum of two positive terms in order to avoid cancellation errors. Illustrate the advantage gained in the case $a = 16.5$, $b = 15.7$, $\gamma = 5°$, using 3-decimal-digit arithmetic. Is the method foolproof?

26. Determine the condition number for the following functions.

    (a) $f(x) = \ln x, \quad x > 0;$
    (b) $f(x) = \cos x, \quad |x| < \frac{1}{2}\pi;$
    (c) $f(x) = \sin^{-1} x, \quad |x| < 1;$
    (d) $f(x) = \sin^{-1} \dfrac{x}{\sqrt{1+x^2}}.$

    Indicate the possibility of ill-conditioning.

27. Compute the condition number of the following functions, and discuss any possible ill-conditioning:

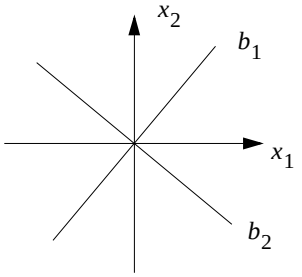
    (a) $f(x) = x^{1/n} \quad (x > 0,\ n > 0$ an integer$);$
    (b) $f(x) = x - \sqrt{x^2 - 1} \quad (x > 1);$
    (c) $\boldsymbol{f}(x_1, x_2) = \sqrt{x_1^2 + x_2^2};$
    (d) $\boldsymbol{f}(x_1, x_2) = x_1 + x_2.$

28. (a) Consider the composite function $h(t) = g(f(t))$. Express the condition of h in terms of the condition of g and f. Be careful to state at which points the various condition numbers are to be evaluated.
    (b) Illustrate (a) with $h(t) = \frac{1+\sin t}{1-\sin t}$, $t = \frac{1}{4}\pi$.

29. Show that $(\operatorname{cond} f \cdot g)(x) \le (\operatorname{cond} f)(x) + (\operatorname{cond} g)(x)$. What can be said about $(\operatorname{cond} f/g)(x)$?

30. Let $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}$ be given by $y = x_1 + x_2$. Define $(\operatorname{cond} \boldsymbol{f})(\boldsymbol{x}) = (\operatorname{cond}_{11} \boldsymbol{f})(\boldsymbol{x}) + (\operatorname{cond}_{12} \boldsymbol{f})(\boldsymbol{x})$, where $\boldsymbol{x} = [x_1, x_2]^T$ (cf. (1.27)).

    (a) Derive a formula for $\kappa(x_1, x_2) = (\operatorname{cond} \boldsymbol{f})(\boldsymbol{x})$.
    (b) Show that $\kappa(x_1, x_2)$ as a function of x_1, x_2 is symmetric with respect to both bisectors b_1 and b_2 (see figure).



(c) Determine the lines in $\mathbb{R}^2$ on which $\kappa(x_1, x_2) = c$, $c \geq 1$ a constant. (Simplify the analysis by using symmetry; cf. part (b).)

31. Let $\|\cdot\|$ be a vector norm in $\mathbb{R}^n$ and denote by the same symbol the associated matrix norm. Show for arbitrary matrices $\boldsymbol{A}$, $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ that

(a) $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$;

(b) $\operatorname{cond}(\boldsymbol{AB}) \leq \operatorname{cond}\boldsymbol{A} \cdot \operatorname{cond}\boldsymbol{B}$.

32. Prove (1.32). {*Hint*: let $m_\infty = \max_\nu \sum_\mu |a_{\nu\mu}|$. Show that $\|\boldsymbol{A}\|_\infty \leq m_\infty$ as well as $\|\boldsymbol{A}\|_\infty \geq m_\infty$, the latter by taking a special vector $\boldsymbol{x}$ in (1.30).}

33. Let the L_1 norm of a vector $\boldsymbol{y} = [y_\lambda]$ be defined by $\|\boldsymbol{y}\|_1 = \sum_\lambda |y_\lambda|$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, show that

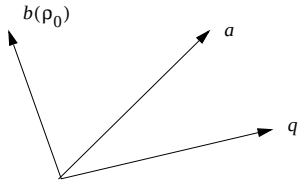
$$\|\boldsymbol{A}\|_1 := \max_{\substack{\boldsymbol{x}\in\mathbb{R}^m \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{Ax}\|_1}{\|\boldsymbol{x}\|_1} = \max_\mu \sum_\nu |a_{\nu\mu}|;$$

that is, $\|\boldsymbol{A}\|_1$ is the "maximum column sum." {*Hint*: let $m_1 = \max_\mu \sum_\nu |a_{\nu\mu}|$. Show that $\|\boldsymbol{A}\|_1 \leq m_1$ as well as $\|\boldsymbol{A}\|_1 \geq m_1$, the latter by taking for $\boldsymbol{x}$ in (1.30) an appropriate coordinate vector.}

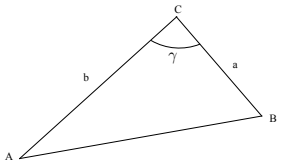
34. Let $\boldsymbol{a}$, $\boldsymbol{q}$ be linearly independent vectors in $\mathbb{R}^n$ of (Euclidean) length 1. Define $\boldsymbol{b}(\rho) \in \mathbb{R}^n$ as follows:

$$\boldsymbol{b}(\rho) = \boldsymbol{a} - \rho\boldsymbol{q}, \quad \rho \in \mathbb{R}.$$

Compute the condition number of the angle $\alpha(\rho)$ between $\boldsymbol{b}(\rho)$ and $\boldsymbol{q}$ at the value $\rho = \rho_0 = \boldsymbol{q}^T\boldsymbol{a}$. (Then $\boldsymbol{b}(\rho_0) \perp \boldsymbol{q}$; see figure.) Discuss the answer.



35. The area Δ of a triangle ABC is given by $\Delta = \frac{1}{2}ab\sin\gamma$ (see figure). Discuss the numerical condition of Δ.



36. Define, for $x \neq 0$,

$$f_n = f_n(x) = (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(\frac{\mathrm{e}^{-x}}{x}\right), \quad n = 0, 1, 2, \ldots .$$

(a) Show that $\{f_n\}$ satisfies the recursion

$$y_k = \frac{k}{x}\, y_{k-1} + \frac{\mathrm{e}^{-x}}{x}, \quad k = 1, 2, 3, \ldots; \quad y_0 = \frac{\mathrm{e}^{-x}}{x}.$$

{*Hint*: differentiate k times the identity $\mathrm{e}^{-x} = x \cdot (\mathrm{e}^{-x}/x)$.}

(b) Why do you expect the recursion in (a), without doing any analysis, to be numerically stable if $x > 0$? How about $x < 0$?

(c) Support and discuss your answer to (b) by considering y_n as a function of y_0 (which for $y_0 = f_0(x)$ yields $f_n = f_n(x)$) and by showing that the condition number of this function at f_0 is

$$(\mathrm{cond}\, y_n)(f_0) = \frac{1}{|e_n(x)|},$$

where $e_n(x) = 1 + x + x^2/2! + \cdots + x^n/n!$ is the nth partial sum of the exponential series. {*Hint*: use Leibniz's formula to evaluate $f_n(x)$.}

37. Consider the algebraic equation

$$x^n + ax - 1 = 0, \quad a > 0, \quad n \geq 2.$$

(a) Show that the equation has exactly one positive root $\xi(a)$.

(b) Obtain a formula for $(\mathrm{cond}\ \xi)(a)$.

(c) Obtain (good) upper and lower bounds for $(\mathrm{cond}\ \xi)(a)$.

38. Consider the algebraic equation

$$x^n + x^{n-1} - a = 0, \quad a > 0, \quad n \geq 2.$$

(a) Show that there is exactly one positive root $\xi(a)$.

(b) Show that $\xi(a)$ is well conditioned as a function of a. Indeed, prove

$$(\mathrm{cond}\, \xi)(a) < \frac{1}{n-1}.$$

39. Consider Lambert's equation

$$x\mathrm{e}^x = a$$

for real values of x and a.

(a) Show graphically that the equation has exactly one root $\xi(a) \geq 0$ if $a \geq 0$, exactly two roots $\xi_2(a) < \xi_1(a) < 0$ if $-1/e < a < 0$, a double root -1 if $a = -1/\mathrm{e}$, and no root if $a < -1/\mathrm{e}$.

(b) Discuss the condition of $\xi(a)$, $\xi_1(a)$, $\xi_2(a)$ as a varies in the respective intervals.

40. Given the natural number n, let $\xi = \xi(a)$ be the unique positive root of the equation $x^n = a\mathrm{e}^{-x}$ $(a > 0)$. Determine the condition of ξ as a function of a; simplify the answer as much as possible. In particular, show that $(\mathrm{cond}\ \xi)(a) < 1/n$.

41. Let $\boldsymbol{f}(x_1, x_2) = x_1 + x_2$ and consider the algorithm A given as follows,

$$\boldsymbol{f}_A:\ \mathbb{R}^2(t,s) \to \mathbb{R}(t,s) \qquad y_A = \mathrm{fl}(x_1 + x_2).$$

Estimate $\gamma(x_1, x_2) = (\mathrm{cond}\, A)(\boldsymbol{x})$, using any of the norms

$$\|\boldsymbol{x}\|_1 = |x_1| + |x_2|, \quad \|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + x_2^2}, \quad \|\boldsymbol{x}\|_\infty = \max\ (|x_1|, |x_2|).$$

Discuss the answer in the light of the conditioning of $\boldsymbol{f}$.

42. This problem deals with the function $f(x) = \sqrt{1-x} - 1$, $-\infty < x < 1$.

    (a) Compute the condition number $(\mathrm{cond}\, f)(x)$.

    (b) Let A be the algorithm that evaluates $f(x)$ in floating-point arithmetic on a computer with machine precision eps, given an (error-free) floating-point number x. Let ε_1, ε_2, ε_3 be the relative errors due, respectively, to the subtraction in $1 - x$, to taking the square root, and to the final subtraction of 1. Assume $|\varepsilon_i| \le \mathrm{eps}$ $(i = 1, 2, 3)$. Letting $f_A(x)$ be the value of $f(x)$ so computed, write $f_A(x) = f(x_A)$ and $x_A = x(1 + \varepsilon_A)$. Express ε_A in terms of x, ε_1, ε_2, ε_3 (neglecting terms of higher order in the ε_i). Then determine an upper bound for $|\varepsilon_A|$ in terms of x and eps, and finally an estimate of $(\mathrm{cond}\, A)(x)$.

    (c) Sketch a graph of $(\mathrm{cond}\, f)(x)$ (found in (a)) and a graph of the estimate of $(\mathrm{cond}\, A)(x)$ (found in (b)) as functions of x on $(-\infty, 1)$. Discuss your results.

43. Consider the function $f(x) = 1 - \mathrm{e}^{-x}$ on the interval $0 \le x \le 1$.

    (a) Show that $(\mathrm{cond}\ f)(x) \le 1$ on [0,1].

    (b) Let A be the algorithm that evaluates $f(x)$ for the machine number x in floating-point arithmetic (with machine precision eps). Assume that the exponential routine returns a correctly rounded answer. Estimate $(\mathrm{cond}\, A)(x)$ for $0 \le x \le 1$, neglecting terms of $O(\mathrm{eps}^2)$. {*Point of information*: $\ln(1 + \varepsilon) = \varepsilon + O(\varepsilon^2)$, $\varepsilon \to 0$.}

    (c) Plot $(\mathrm{cond}\, f)(x)$ and your estimate of $(\mathrm{cond}\, A)(x)$ as functions of x on [0,1]. Comment on the results.

44. (a) Suppose A is an algorithm that computes the (smooth) function $f(x)$ for a given machine number x, producing $f_A(x) = f(x)(1 + \varepsilon_f)$, where

$|\varepsilon_f| \leq \varphi(x)$eps (eps = machine precision). If $0 < (\operatorname{cond} f)(x) < \infty$, show that

$$(\operatorname{cond} A)(x) \leq \frac{\varphi(x)}{(\operatorname{cond} f)(x)}$$

if second-order terms in eps are neglected. {*Hint*: set $f_A(x) = f(x_A)$, $x_A = x(1+\varepsilon_A)$, and expand in powers of ε_A, keeping only the first.}

(b) Apply the result of (a) to $f(x) = \frac{1-\cos x}{\sin x}$, $0 < x < \frac{1}{2}\pi$, when evaluated as shown. (You may assume that $\cos x$ and $\sin x$ are computed within a relative error of eps.) Discuss the answer.

(c) Do the same as (b), but for the (mathematically equivalent) function $f(x) = \frac{\sin x}{1+\cos x}$, $0 < x < \frac{1}{2}\pi$.

MACHINE ASSIGNMENTS

1. Let $x = 1+\pi/10^6$. Compute the nth power of x for $n = 100\,000, 200\,000, \ldots, 1\,000\,000$ once in single, and once in double Matlab precision. Let the two results be p_n and dp_n. Use the latter to determine the relative errors r_n of the former. Print $n, p_n, dp_n, r_n, r_n/(n \cdot \text{eps0})$, where eps0 is the single-precision `eps`. What should x^n be, approximately, when $n = 1\,000\,000$? Comment on the results.

2. Compute the derivative $\mathrm{d}y/\mathrm{d}x$ of the exponential function $y = \mathrm{e}^x$ at $x = 0$ from the difference quotients $d(h) = (\mathrm{e}^h - 1)/h$ with decreasing h. Use

   (a) $h = h1 := 2^{-i}$, $i = 5:5:50$;

   (b) $h = h2 := (2.2)^{-i}$, $i = 5:5:50$.

   Print the quantities $i, h1, h2, d1 := d(h1), d2 := d(h2)$, the first and two last ones in `f`-format, the others in `e`-format. Explain what you observe.

3. Consider the following procedure for determining the limit $\lim_{h\to 0}(\mathrm{e}^h - 1)/h$ on a computer. Let

$$d_n = \mathrm{fl}\left(\frac{\mathrm{e}^{2^{-n}} - 1}{2^{-n}}\right) \quad \text{for } n = 0, 1, 2, \ldots$$

   and accept as the machine limit the first value satisfying $d_n = d_{n-1}$ $(n \geq 1)$.

   (a) Write and run a Matlab routine implementing the procedure.

   (b) In $\mathbb{R}(t,s)$-floating-point arithmetic, with rounding by chopping, for what value of n will the correct limit be reached, assuming no underflow (of 2^{-n}) occurs? {*Hint*: use $e^h = 1 + h + \frac{1}{2}h^2 + \cdots$.} Compare with the experiment made in (a).

(c) On what kind of computer (i.e., under what conditions on s and t) will underflow occur before the limit is reached?

4. Euler's constant $\gamma = .57721566490153286\ldots$ is defined as the limit

$$\gamma = \lim_{n\to\infty} \gamma_n, \quad \text{where} \quad \gamma_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \ln n.$$

Assuming that $\gamma - \gamma_n \sim cn^{-d}$, $n \to \infty$, for some constants c and $d > 0$, try to determine c and d experimentally on the computer.

5. Letting $\Delta u_n = u_{n+1} - u_n$, one has the easy formula

$$\sum_{n=1}^{N} \Delta u_n = u_{N+1} - u_1.$$

With $u_n = \ln(1+n)$, compute each side (as it stands) for $N = 1\,000 : 1\,000 : 10\,000$, the left-hand side in Matlab single precision, and the right-hand side in double precision. Print the relative discrepancy of the two results. Repeat with $\sum_{n=1}^{N} u_n$: compute the sum in single and double precision and compare the results. Try to explain what you observe.

6. (a) Write a program to compute

$$S_N = \sum_{n=1}^{N} \left[\frac{1}{n} - \frac{1}{n+1}\right] = \sum_{n=1}^{N} \frac{1}{n(n+1)},$$

once using the first summation, and once using the (mathematically equivalent) second summation. For $N = 10^k$, $k = 1 : 7$, print the respective absolute errors. Comment on the results.

(b) Write a program to compute

$$p_N = \prod_{n=1}^{N} \frac{n}{n+1}\,.$$

For the same values of N as in part (a), print the relative errors. Comment on the results.

7. (a) Prove: *based on best possible relative error bounds, the floating-point addition* $\mathrm{fl}(\mathrm{fl}(x+y)+z)$ *is more accurate than* $\mathrm{fl}(x+\mathrm{fl}(y+z))$ *if and only if* $|x+y| < |y+z|$. As applications, formulate addition rules in the cases

(a1) $0 < x < y < z$;
(a2) $x > 0$, $y < 0$, $z > 0$;
(a3) $x < 0$, $y > 0$, $z < 0$.

(b) Consider the nth partial sums of the series defining the zeta function $\zeta(s)$, resp., eta function $\eta(s)$,

$$z_n = \sum_{k=1}^{n} \frac{1}{k^s}\,, \quad e_n = \sum_{k=1}^{n} (-1)^{k-1} \frac{1}{k^s}\,.$$

For $s = 2, 11/3, 5, 7.2, 10$ and $n = 50, 100, 200, 500, 1000$, compute these sums in Matlab single precision, once in forward direction, and once in backward direction, and compare the results with Matlab double-precision evaluations. Interpret the results in the light of your answers to part (a), especially (a2) and (a3).

8. Let $n = 10^6$ and

$$s = 10^{11} n + \sum_{k=1}^{n} \ln\, k.$$

(a) Determine s analytically and evaluate to 16 decimal digits.

(b) The following Matlab program computes s in three different (but mathematically equivalent) ways:

```
%MAI_8B
%
n=10^6; s0=10^11*n;
s1=s0;
for k=1:n
  s1=s1+log(k);
end
s2=0;
for k=1:n
  s2=s2+log(k);
end
s2=s2+s0;
i=1:n;
s3=s0+sum(log(i));
[s1 s2 s3]'
```

Run the program and discuss the results.

9. Write a Matlab program that computes the Euclidean condition number of the Hilbert matrix $\boldsymbol{H}_n$ following the prescription given in footnote 3 of the text.

(a) The inverse of the Hilbert matrix $\boldsymbol{H}_n$ has elements

$$\left(\boldsymbol{H}_n^{-1}\right)_{ij} = (-1)^{i+j}(i+j-1)\binom{n+i-1}{n-j}\binom{n+j-1}{n-i}\binom{i+j-2}{i-1}^2$$

(cf. Note 3 to Sect. 1.3.2). Simplify the expression to avoid factorials of large numbers. {*Hint*: express all binomial coefficients in terms of factorials and simplify.}

(b) Implement in Matlab the formula obtained in (a) and reproduce Table 1.1 of the text.

10. The (symmetrically truncated) cardinal series of a function f is defined by

$$C_N(f,h)(x) = \sum_{k=-N}^{N} f(kh)\,\mathrm{sinc}\left(\frac{x-kh}{h}\right),$$

where $h > 0$ is the spacing of the data and the sinc function is defined by

$$\mathrm{sinc}(u) = \begin{cases} \dfrac{\sin(\pi u)}{\pi u} & \text{if } u \neq 0, \\ 1 & \text{if } u = 0 \end{cases}$$

Under appropriate conditions, $C_N(f,h)(x)$ approximates $f(x)$ on $[-Nh, Nh]$.

(a) Show that

$$C_N(f,h)(x) = \frac{h}{\pi}\sin\frac{\pi x}{h}\sum_{k=-N}^{N}\frac{(-1)^k}{x-kh}\,f(kh).$$

Since this requires the evaluation of only one value of the sine function, it provides a more efficient way to evaluate the cardinal series than the original definition.

(b) While the form of C_N given in (a) may be more efficient, it is numerically unstable when x is near one of the abscissae kh. Why?

(c) Find a way to stabilize the formula in (a). {*Hint*: introduce the integer k_0 and the real number t such that $x = (k_0 + t)h$, $|t| \leq \frac{1}{2}$.}

(d) Write a program to compute $C_N(f,h)(x)$ according to the formula in (a) and the one developed in (c) for $N = 100$, $h = .1$, $f(x) = x\exp(-x^2)$, and $x = .55$, $x = .5 + 10^{-8}$, $x = .5 + 10^{-15}$. Print $C_N(f,h)(x)$, $f(x)$, and the error $|C_N(f,h)(x) - f(x)|$ in either case. Compare the results.

11. In the theory of Fourier series the numbers

$$\lambda_n = \frac{1}{2n+1} + \frac{2}{\pi}\sum_{k=1}^{n}\frac{1}{k}\tan\frac{k\pi}{2n+1}, \quad n = 1, 2, 3, \ldots\ ,$$

known as *Lebesgue constants*, are of some importance.

(a) Show that the terms in the sum increase monotonically with k. How do the terms for k near n behave when n is large?

(b) Compute λ_n for $n = 1, 10, 10^2, \ldots, 10^5$ in Matlab single and double precision and compare the results. Do the same with n replaced by $\lceil n/2 \rceil$. Explain what you observe.

12. Sum the series

$$\text{(a)} \quad \sum_{n=0}^{\infty} (-1)^n / n!^2 \qquad \text{(b)} \quad \sum_{n=0}^{\infty} 1/n!^2$$

until there is no more change in the partial sums to within the machine precision. Generate the terms recursively. Print the number of terms required and the value of the sum. (Answers in terms of Bessel functions: (a) $J_0(2)$; cf. Abramowitz and Stegun [1964, (9.1.18)] or Olver et al. [2010, (10.9.1)] and (b) $I_0(2)$; cf. Abramowitz and Stegun [1964, (9.6.16)] or Olver et al. [2010, (10.32.1)].)

13. (P.J. Davis, 1993) Consider the series $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k^{3/2} + k^{1/2}}$. Try to compute the sum to three correct decimal digits.

14. We know from calculus that

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = \mathrm{e}.$$

What is the "machine limit"? Explain.

15. Let $f(x) = (n+1)x - 1$. The iteration

$$x_k = f(x_{k-1}), \quad k = 1, 2, \ldots, K; \quad x_0 = 1/n,$$

in exact arithmetic converges to the fixed point $1/n$ in one step (why?). What happens in machine arithmetic? Run a program with $n = 1 : 5$ and $K = 10 : 10 : 50$ and explain quantitatively what you observe.

16. Compute the integral $\displaystyle\int_0^1 \mathrm{e}^x \mathrm{d}x$ from Riemann sums with n equal subintervals, evaluating the integrand at the midpoint of each. Print the Riemann sums for $n = 5\,000 : 5\,000 : 100\,000$ (to 15 decimal digits after the decimal point), together with absolute errors. Comment on the results.

17. Let $y_n = \displaystyle\int_0^1 t^n \mathrm{e}^{-t} \mathrm{d}t$, $n = 0, 1, 2, \ldots$.

(a) Use integration by parts to obtain a recurrence formula relating y_k to y_{k-1} for $k = 1, 2, 3, \ldots$, and determine the starting value y_0.

(b) Write and run a Matlab program that generates $y_0, y_1, \ldots, y_{20}$, using the recurrence of (a), and print the results to 15 decimal digits after the decimal point. Explain in detail (quantitatively, using mathematical analysis) what is happening.

(c) Use the recursion of (a) in reverse order, starting (arbitrarily) with $y_N = 0$. Place into five consecutive columns of a (21×5) matrix Y the values $y_0^{(N)}, y_1^{(N)}, \ldots, y_{20}^{(N)}$ thus obtained for $N = 22, 24, 26, 28, 30$. Determine how much consecutive columns of Y differ from one another by printing

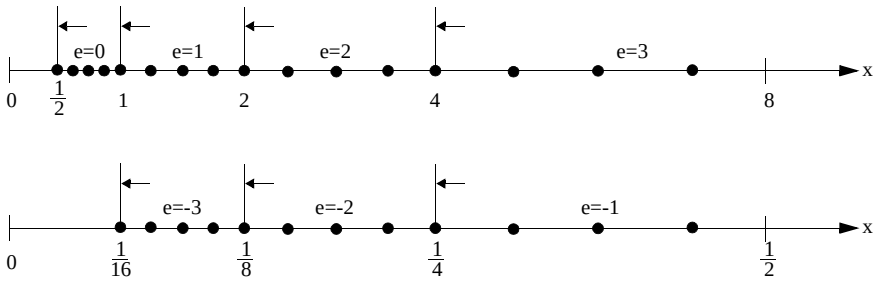
$$e_i = \max |\, (Y(:, i+1) - Y(:, i))\, ./ Y(:, i+1)|, \quad i = 1, 2, 3, 4.$$

Print the last column $Y(:, 5)$ of Y and explain why this represents accurately the column vector of the desired quantities $y_0, y_1, \ldots, y_{20}$.

ANSWERS TO THE EXERCISES AND MACHINE ASSIGNMENTS OF CHAPTER 1

ANSWERS TO EXERCISES

1. The plots for the two pieces of the real axis are as follows:



2. (a) Given $x = \left(\sum_{k=1}^{t} b_{-k}2^{-k}\right) \cdot 2^e$, $b_{-1} = 1$, one has $x_{\text{next}} = x + 2^{-t} \cdot 2^e$, thus $d(x) = x_{\text{next}} - x = 2^{-t} \cdot 2^e$.

   (b) We have $d(x)/x = 2^{-t} \cdot 2^e/x$. Since x is normalized, there holds $\frac{1}{2} \cdot 2^e \leq x < 2^e$, hence $2^{-t} < d(x)/x < 2 \cdot 2^{-t}$.

3. Assume first rounding by *chopping*. We claim that

$$x = (.11\ldots1) \times 2^{-t+1}.$$

(All binary numbers have t binary digits in single precision, and $2t$ in double precision.) We have

$$\begin{aligned} 1 &= (.100\ldots0) \times 2 \\ x &= (.000\ldots0\ 11\ldots1) \times 2 \quad \text{(double precision)} \\ \text{fl}(1+x) &= \text{fl}[(.100\ldots011\ldots1) \times 2] \\ &= (.100\ldots0) \times 2 \quad \text{(chopping!)} \\ &= 1. \end{aligned}$$

On the other hand, for the next larger number,

$$\begin{aligned} x_{\text{next}} &= (.10\ldots0) \times 2^{-t+2} = (.00\ldots01) \times 2 \\ \text{fl}(1+x_{\text{next}}) &= \text{fl}[(.100\ldots01) \times 2] \\ &= (.100\ldots01) \times 2 \\ &> 1. \end{aligned}$$

For *symmetric rounding*, the answer is clearly half the previous answer, that is, $x = (.11\ldots1) \times 2^{-t}$.

4. The marked table should look as follows:

operation	exact	rounded	overflow	underflow
$z = \mathrm{fl}(x - y)$	$\times$			
$z = \mathrm{fl}((x - y)^{10})$				$\times$
$z = \mathrm{fl}(x + y)$		$\times$		
$z = \mathrm{fl}(y + (x/4))$		$\times$		
$z = \mathrm{fl}(x + (y/4))$	$\times$			

5. The output (on a Sparc workstation) is

   `eps0=2.220446049250313-16`,

   which is precisely `eps=2.220446049250313-16`.

   Analysis: In exact arithmetic,

$$4/3 = .1010\overline{10}\ldots\ .$$

   Therefore, since $t = 53$ is odd,

$$\mathtt{a} = \mathrm{fl}(4/3) = .1010\ldots101 \times 2,$$

   where the last digit "1" of the mantissa is in position 53. There follows

$$\mathtt{b} = \mathtt{a} - \mathtt{1} = (.1010\ldots101 - .1) \times 2 = (.0010\ldots101) \times 2,$$
$$\mathtt{c} = \mathtt{b} + \mathtt{b} + \mathtt{b} = (.0111\ldots111) \times 2,$$
$$\mathtt{abs(c - 1)} = \mathtt{1} - \mathtt{c} = (.1 - .0111\ldots111) \times 2 = (.0000\ldots001) \times 2 = 2 \times 2^{-53},$$

   as was to be shown.

6. Let

$$x = f \cdot 2^e, \quad f = \pm\, .b_{-1}b_{-2}\cdots b_{-t}b_{-(t+1)}\cdots\ .$$

   Case 1: $b_{-(t+1)} = 0$. Here,

$$\mathrm{rd}(x) = \pm(.b_{-1}b_{-2}\cdots b_{-t})2^e,$$

$$|x - \mathrm{rd}(x)| = \left|\sum_{k=t+2}^{\infty} b_{-k}2^{-k}\right| 2^e \quad (\text{since } b_{-(t+1)} = 0)$$

$$\leq \sum_{k=t+2}^{\infty} 2^{-k} \cdot 2^e = 2^{-(t+1)} \cdot 2^e,$$

   so that

$$\left|\frac{x - \mathrm{rd}(x)}{x}\right| \leq \frac{2^{-(t+1)} \cdot 2^e}{\frac{1}{2} \cdot 2^e} = 2^{-t}.$$

Case 2: $b_{-(t+1)} = 1$. Here,

$$\mathrm{rd}(x) = \pm\{(.b_{-1}b_{-2}\cdots b_{-t}) + 2^{-t}\}2^e,$$

$$\begin{aligned}|x - \mathrm{rd}(x)| &= \left|\pm \sum_{k=t+1}^{\infty} b_{-k}2^{-k} \mp 2^{-t}\right| 2^e \\ &= \left(2^{-t} - 2^{-(t+1)} - \sum_{k=t+2}^{\infty} b_{-k}2^{-k}\right) 2^e \quad (\text{since } b_{-(t+1)} = 1) \\ &= \left(2^{-(t+1)} - \sum_{k=t+2}^{\infty} b_{-k}2^{-k}\right) 2^e \le 2^{-(t+1)} \cdot 2^e,\end{aligned}$$

and the conclusion follows as in Case 1.

7. (a) The absolute error *is* a distance function on $S = \mathbb{R}$, since it clearly satisfies (i) – (iii).

   (b) The relative error is *not* a distance function on the set of nonzero real numbers, since it is not symmetric. Nor is the triangle inequality always satisfied, since

$$\left|\frac{x-y}{x}\right| \le \left|\frac{x-z}{x}\right| + \left|\frac{z-y}{z}\right| \qquad (xz \neq 0)$$

   is equivalent to

$$|z|\,|x-y| \le |z|\,|x-z| + |x|\,|z-y|,$$

   which, when $x \to 0$, yields $|z|\,|y| \le |z|^2$, which is false if $|y| > |z|$.

   (c) On the set $S = \{x \in \mathbb{R} : \ x \neq 0\}$, the relative precision is *not* a distance function, since $\mathrm{rp}(-x, x) = 0$ even though $-x \neq x$ (if $x \neq 0$). However, rp *is* a distance function on the set of positive numbers $S = \{x \in \mathbb{R}, x > 0\}$. Property (i) is then satisfied, as well as (ii). Also the triangle inequality holds for any $x > 0$, $y > 0$, $z > 0$:

$$\begin{aligned}|\ln x - \ln y| &= |\ln x - \ln z + \ln z - \ln y| \\ &\le |\ln x - \ln z| + |\ln z - \ln y|.\end{aligned}$$

   We have

$$\mathrm{rp}(x, y) = \left|\ln\left|\tfrac{x}{y}\right|\right| = \left|\ln \tfrac{1}{|1+\varepsilon|}\right| = |\ln(1+\varepsilon)| = O(\varepsilon) \quad \text{as} \ \ \varepsilon \to 0.$$

8. (a) Since

$$|x_1^* - x_2^*| = |x_1(1+E_1) - x_2(1+E_2)| = |x_1 - x_2 + x_1E_1 - x_2E_2|$$

$$\geq |x_1 - x_2| - |x_1|\ |E_1| - |x_2|\ |E_2| \geq |x_1 - x_2| - (|x_1| + |x_2|)E,$$

we certainly have $x_1^* \neq x_2^*$ if

$$|x_1 - x_2| - (|x_1| + |x_2|)E > 0,$$

that is, if

$$E < \frac{|x_1 - x_2|}{|x_1| + |x_2|}\,.$$

(b(1)) We have

$$\frac{1}{x_1^* - x_2^*} = \frac{1}{x_1(1+E_1) - x_2(1+E_2)} = \frac{1}{x_1 - x_2 + x_1E_1 - x_2E_2}$$

$$= \frac{1}{(x_1 - x_2)\left(1 + \frac{x_1}{x_1 - x_2}\,E_1 - \frac{x_2}{x_1 - x_2}\,E_2\right)}$$

$$\approx \frac{1}{x_1 - x_2}\left(1 - \frac{x_1}{x_1 - x_2}\,E_1 + \frac{x_2}{x_1 - x_2}\,E_2\right),$$

so that the relative error, in absolute value, is approximately equal to

$$\left| - \frac{x_1}{x_1 - x_2}\,E_1 + \frac{x_2}{x_1 - x_2}\,E_2 \right| \leq \frac{|x_1| + |x_2|}{|x_1 - x_2|}\,E.$$

(b(2)) Additional errors caused by subtraction and division amount to multiplication by

$$\frac{1+\varepsilon_2}{1+\varepsilon_1} \approx (1+\varepsilon_2)(1-\varepsilon_1) \approx 1 - \varepsilon_1 + \varepsilon_2, \qquad |\varepsilon_i| \leq \text{eps},$$

so that the result in (b(1)) must be multiplied by $1 - \varepsilon_1 + \varepsilon_2$, giving

$$\frac{1}{x_1 - x_2}\left(1 - \frac{x_1}{x_1 - x_2}\,E_1 + \frac{x_2}{x_1 - x_2}\,E_2\right)(1 - \varepsilon_1 + \varepsilon_2)$$
$$= \frac{1}{x_1 - x_2}\left(1 - \frac{x_1}{x_1 - x_2}\,E_1 + \frac{x_2}{x_1 - x_2}\,E_2 - \varepsilon_1 + \varepsilon_2\right).$$

Therefore, the relative error, in absolute value, is now approximately equal to

$$\left| - \frac{x_1}{x_1 - x_2}\,E_1 + \frac{x_2}{x_1 - x_2}\,E_2 - \varepsilon_1 + \varepsilon_2 \right| \leq \frac{|x_1| + |x_2|}{|x_1 - x_2|}\,E + 2\,\text{eps}.$$

9. The program has the following flaws:

flaw #1: $p = q = 0$ ($x_1 = x_2 = 0$) yields division by zero in the last statement, and thus `x2 = NaN`.

flaw #2: even when p and q, as well as x_1 and x_2, are within exponent range, the same may not be true for $p \times p$. Example:

$$x^2 - 10^{200}x + 10^{50} = 0, \quad x_1 \approx 10^{200}, \quad x_2 \approx 10^{-150}.$$

Here, on a machine with, say, exponent range $[-308, 308]$ the numbers p, q, x_1, x_2 are representable, but $p^2 = 10^{400}$ is not.

10. Express e^x in terms of y and obtain

$$\sinh x = \frac{1}{2}\left(y + 1 - \frac{1}{y+1}\right) = \frac{1}{2}\, y\, \frac{y+2}{y+1}\,.$$

11. (a) If $|x|$ is small, there will be a large cancellation error since the square-root term, in general, is not machine-representable and has to be rounded. An easy way to avoid cancellation is to use

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

which requires only benign arithmetic operations.

(b) We have

$$(\mathrm{cond}\, f)(x) = \left| \frac{x \cdot \frac{1}{2}\frac{2x}{\sqrt{1+x^2}}}{\sqrt{1+x^2}-1} \right| = \frac{x^2}{\sqrt{1+x^2}(\sqrt{1+x^2}-1)}$$

$$= \frac{x^2(\sqrt{1+x^2}+1)}{\sqrt{1+x^2}\cdot x^2} = 1 + \frac{1}{\sqrt{1+x^2}}.$$

Hence, $(\mathrm{cond}\, f)(x) \le 2$ for all real x, and f is well conditioned (even for small $|x|$).

(c) The condition of f is a property of the function f alone and does not depend on algorithmic considerations, whereas cancellation does. We have here an example of a "computational crime", where a well-conditioned problem is solved by an "ill-conditioned" algorithm. (See also Problem 42.)

12. (a) Here, x^n is computed by

$$\begin{aligned} &p_1 = x \\ &p_k = \mathrm{fl}(x p_{k-1}), \quad k = 2, 3, \ldots, n, \\ &p_n =: \mathrm{fl}(x^n). \end{aligned}$$

Since $p_2 = x^2(1+\varepsilon_2)$, $p_3 = x^3(1+\varepsilon_2)(1+\varepsilon_3), \ldots$, with $|\varepsilon_k| \le \mathrm{eps}$, we have

$$p_n = x^n(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n),$$

hence

$$\left|\frac{p_n - x^n}{x^n}\right| = |(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n) - 1|$$
$$\approx |\varepsilon_2 + \varepsilon_3 + \cdots + \varepsilon_n| \le (n-1)\ \text{eps}.$$

For a more rigorous answer, see Ex. 21(b).

(b) Here,

$$\text{fl}(x^n) := \mathrm{e}^{n[\ln x](1+\varepsilon_1)(1+\varepsilon_2)}(1+\varepsilon_3), \quad |\varepsilon_i| \le \text{eps}, \quad i = 1,2,3,$$

where ε_1, ε_2, ε_3 are the relative errors committed in computing respectively the logarithm, the product by n, and the exponential. Thus,

$$\text{fl}(x^n) \approx \mathrm{e}^{n[\ln x](1+\varepsilon_1+\varepsilon_2)}(1+\varepsilon_3) = \mathrm{e}^{n[\ln x]}\mathrm{e}^{(\varepsilon_1+\varepsilon_2)n[\ln x]}(1+\varepsilon_3)$$
$$\approx x^n(1 + (\varepsilon_1+\varepsilon_2)n\ln x + \varepsilon_3),$$

and

$$\left|\frac{\text{fl}(x^n) - x^n}{x^n}\right| \approx |(\varepsilon_1+\varepsilon_2)n\ln x + \varepsilon_3| \le (2n|\ln x| + 1)\ \text{eps}.$$

It follows that (a) is more accurate than (b) if

$$n - 1 \le 2n|\ln x| + 1,$$

that is, if

$$n(1 - 2|\ln x|) \le 2.$$

This is always true if $|\ln x| > \frac{1}{2}$, that is, if $0 < x < \mathrm{e}^{-\frac{1}{2}}$ or $x > \mathrm{e}^{\frac{1}{2}}$, and on the interval $\mathrm{e}^{-\frac{1}{2}} < x < \mathrm{e}^{\frac{1}{2}}$ it is true if

$$n \le \frac{2}{1 - 2|\ln x|}\,.$$

13. (a) By assumption, except for a factor $(1+\varepsilon_1)(1+\varepsilon_2)$, $|\varepsilon_i| \le \text{eps}$, accounting for the subtraction and division, there holds

$$\text{fl}(f(x)) = \frac{1 - (1+\varepsilon_c)\cos x}{x} = \frac{1-\cos x}{x}\left\{1 - \frac{\cos x}{1-\cos x}\ \varepsilon_c\right\}.$$

The relative error is thus

$$\varepsilon_f := -\frac{\cos x}{1-\cos x}\ \varepsilon_c = -\frac{\cos x}{2\sin^2\frac{x}{2}}\ \varepsilon_c \sim -\frac{2}{x^2}\ \varepsilon_c \quad \text{as} \quad x \to 0.$$

(b) Here, again with unimportant factors $1+\varepsilon_i$ removed, we have

$$\text{fl}(f(x)) = \frac{(1+\varepsilon_s)^2\sin^2 x}{x(1+(1+\varepsilon_c)\cos x)} = \frac{\sin^2 x}{x(1+\cos x)}\ \frac{(1+\varepsilon_s)^2}{1 + \frac{\cos x}{1+\cos x}\ \varepsilon_c}$$
$$\approx \frac{\sin^2 x}{x(1+\cos x)}\ (1+2\varepsilon_s)\left(1 - \frac{\cos x}{1+\cos x}\ \varepsilon_c\right)$$
$$\approx \frac{\sin^2 x}{x(1+\cos x)}\left(1 + 2\varepsilon_s - \frac{\cos x}{1+\cos x}\ \varepsilon_c\right).$$

This yields

$$\varepsilon_f \approx 2\varepsilon_s - \frac{\cos x}{1+\cos x}\,\varepsilon_c \sim 2\varepsilon_s - \frac{1}{2}\,\varepsilon_c \quad \text{as} \quad x \to 0.$$

Evidently, the error ε_f is now of the same order as ε_s and ε_c. The reason is that only benign arithmetic operations are required, if x is small, in contrast to the form in (a), where a large cancellation error occurs in the calculation of $1 - \cos x$.

(c) We have

$$(\operatorname{cond} f)(x) = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\,\frac{x\sin x-(1-\cos x)}{x^2}}{\frac{1-\cos x}{x}}\right|$$

$$= \left|\frac{x\sin x-(1-\cos x)}{1-\cos x}\right| = \left|\frac{x\sin x}{1-\cos x} - 1\right|.$$

Note that $(\operatorname{cond} f)(x) \sim 1$ as $x \to 0$, so that f is very well conditioned for small $|x|$. However, if $x \approx 2\pi n$, $n \neq 0$, say $x = 2\pi n + \varepsilon$, then

$$(\operatorname{cond} f)(2\pi n + \varepsilon) \sim \frac{4\pi|n|}{|\varepsilon|} \quad \text{as} \quad \varepsilon \to 0, \quad n \neq 0,$$

and f is ill conditioned, more so the larger n.

14. See the text.

15. There could be premature termination of the series, prior to the terms (equal to 1) for $n = t - 1$ and $n = t$ (when t is a large integer), which dramatically alters the sum. This is illustrated by the first evaluation in the routine below; the second applies the termination criterion only after the term with $n = t$ has passed, and the third sums the series directly with twice the number of terms used in the second evaluation.

```
 PROGRAM

%EI_15
%
f0='%8.0f %12.7f\n';
disp('       n      sum')
t=20;
%
% first evaluation
%
n=0; s1=1; s=0;
while abs(s-s1)>.5e-7
  n=n+1; s=s1;
  s1=s+1/(1+n^4*(t-n)^2*(t-n-1)^2);
```

```
end
fprintf(f0,n,s1)
%
% second evaluation
%
k0=ceil(t); k=0:k0;
s1=sum(1./(1+k.^4.*(t-k).^2.*(t-k-1).^2));
n=k0; s=0;
while abs(s-s1)>.5e-7
  n=n+1; s=s1;
  s1=s+1/(1+n^4*(t-n)^2*(t-n-1)^2);
end
fprintf(f0,n,s1)
%
% third evaluation
%
k=0:2*n;
s1=sum(1./(1+k.^4.*(t-k).^2.*(t-k-1).^2));
fprintf(f0,2*n,s1)

OUTPUT

>> EI_15
       n      sum
       5    1.0000095
      23    3.0000139
      46    3.0000140
>>
```

The results speak for themselves. Similar results are obtained for noninteger t in $[20, 21]$.

16. Overflow occurs if either $\tan x > X_+$ or $\tan x < -X_-$. By the periodicity of the tangent function, it suffices to look at x near $\frac{\pi}{2}$.

    Let $x = \frac{\pi}{2} - \varepsilon$ ($|\varepsilon|$ small). Then

$$\tan x = \frac{\sin\left(\frac{\pi}{2} - \varepsilon\right)}{\cos\left(\frac{\pi}{2} - \varepsilon\right)} = \frac{\cos\varepsilon}{\sin\varepsilon} \sim \frac{1}{\varepsilon} \quad \text{for} \quad \varepsilon \to 0.$$

    So, overflow occurs (approximately) if $\frac{1}{\varepsilon} > X_+$ or $\frac{1}{\varepsilon} < -X_-$, i.e., if

$$-\frac{1}{X_-} < \varepsilon < \frac{1}{X_+}\,,$$

    or, equivalently, if

$$-\frac{1}{X_+} < x - \frac{\pi}{2} < \frac{1}{X_-}\,.$$

The answer, in general, is

$$-\frac{1}{X_+} < x - (2k+1)\frac{\pi}{2} < \frac{1}{X_-}\,, \quad k = 0, \pm 1, \pm 2, \dots\ .$$

17. (a) With X_+ the largest machine-representable number, and using Stirling's approximation $n! \sim \sqrt{2\pi n}(n/\mathrm{e})^n$, we look for the first n such that $\ln n! > \ln X_+$, i.e., approximately

$$\frac{n}{\mathrm{e}} \ln \frac{n}{\mathrm{e}} > \frac{\ln X_+ - \ln \sqrt{2\pi n}}{\mathrm{e}}\,.$$

In Matlab, $X_+ = \texttt{realmax} = 1.797693\ldots \times 10^{308}$ and the routine

```
%EI_17A
%
e=exp(1); y=0; n=0;
while y<=0
  n=n+1;
  y=(n/e)*log(n/e)-(log(realmax) ...
    -log(sqrt(2*pi*n)))/e;
end
n
```

yields $n = 171$. Indeed,

$\texttt{factorial}(170) = 7.257415615307994\texttt{e}(+306), \quad \texttt{factorial}(171) = \texttt{Inf}.$

(b) We seek the smallest n such that $n \ln x > \ln X_+$. The routine

```
%EI_17B
%
f0='%8.0f %7.0f\n';
disp('      x       n')
for x=10:10:100
  y=0; n=0;
  while y<=0
    n=n+1;
    y=n*log(x)-log(realmax);
  end
  fprintf(f0,x,n)
end
```

yields

```
>> EI_17B
      x       n
      10     309
```

```
    20      237
    30      209
    40      193
    50      182
    60      174
    70      168
    80      162
    90      158
   100      155
>>
```

(c) Compute as follows:

$$\frac{x^n \mathrm{e}^{-x}}{n!} = \exp(n \ln x - x - \ln n!),$$

where an asymptotic formula can be used for $\ln n!$, say

$$\ln n! = \left(n + \frac{1}{2}\right) \ln(n+1) - (n+1) + \frac{1}{2} \ln(2\pi) + \frac{1}{12(n+1)} - \frac{1}{360(n+1)^3} + \frac{1}{1260(n+1)^5} - \frac{1}{1680(n+1)^7} + \cdots$$

for $n > 15$, and

`lnfact=sum(log(k))` where `k=2:n`

for $2 \le n \le 15$. The integer 15 is the smallest value of n for which the relative error of the above asymptotic approximation for $\ln n!$ is less than $.5 \times 10^{-15}$.

18. (a) Assume one rounds *up* if the first discarded digit is 5. Then, if d is the exponent of x, and $x^\star$ the rounded value of x, one has

$$|x - x^\star| \le 5 \times 10^{-4} \times 10^d,$$
$$\frac{|x - x^\star|}{|x|} \le \frac{5 \times 10^{-4} \times 10^d}{\frac{1}{10} \times 10^d} = .5 \times 10^{-2}.$$

If instead one rounds *down*, then .5 should be replaced by .6.

(b) For $x_1 = .982$, $x_2 = .984$, we have $\mathrm{fl}(x_1 + x_2) = \mathrm{fl}(.1966 \times 10^1) = .197 \times 10^1$, $\mathrm{fl}(\frac{1}{2}\mathrm{fl}(x_1 + x_2)) = \mathrm{fl}(.985) = .985 \notin [x_1, x_2]$.

(c) Since $\mathrm{fl}(m) = \frac{1}{2}(x_1 + x_2)(1+\varepsilon_1)(1+\varepsilon_2)$, where $|\varepsilon_i| \le \varepsilon$, $\varepsilon = .5 \times 10^3$, one has $\mathrm{fl}(m) - x_1 = \frac{1}{2}[(x_1 + x_2)(1 + \varepsilon_1)(1 + \varepsilon_2) - 2x_1] > \frac{1}{2}[(x_1 + x_2)(1 - \varepsilon)^2 - 2x_1]$. Therefore,

$$\mathrm{fl}(m) - x_1 > 0 \ \text{ if } \ (x_1 + x_2)(1 - \varepsilon)^2 - 2x_1 > 0.$$

The latter is equivalent to

$$\frac{x_2 - x_1}{(x_1 + x_2)\left(1 + \sqrt{\frac{2x_1}{x_1+x_2}}\right)} > \varepsilon. \tag{1}$$

Similarly, $\mathrm{fl}(m) - x_2 < 0$ if

$$\frac{x_2 - x_1}{(x_1 + x_2)\left(1 + \sqrt{\frac{2x_2}{x_1+x_2}}\right)} > \varepsilon. \tag{2}$$

Since $x_2 > x_1$, condition (2) implies (1), so that (2) is the desired sufficient condition.

If one neglects ε^2 against ε, then instead of (1) and (2), one obtains

$$\frac{x_2 - x_1}{2(x_1 + x_2)} > \varepsilon.$$

19. (a) $\mathrm{eps} = 2^{-12}$

(b) Let $\lambda = 3/4 = (.110)_2$. Then

$$\begin{aligned} x = 6/7 &= \lambda \cdot 8/7 = \lambda/(1 - 1/8) \\ &= \lambda(1 + 8^{-1} + 8^{-2} + 8^{-3} + \cdots) \\ &= (.110\ 110\ \overline{110}\ldots)_2 \end{aligned}$$

and

$$x^* = (.110\ 110\ 110\ 111)_2.$$

(c) We have

$$x^\star - 2^{-12} = (.110\ 110\ 110\ 110)_2,$$

hence

$$x - (x^\star - 2^{-12}) = (.000\ 000\ 000\ 000\ \overline{110}\ldots)_2 = 2^{-12}x.$$

Therefore,

$$x^\star - x = 2^{-12}(1 - x),$$

and

$$\varepsilon = \frac{x^\star - x}{x} = 2^{-12}\ \frac{1 - x}{x} = \frac{1}{6} \cdot 2^{-12},$$

$$\frac{|\varepsilon|}{\mathrm{eps}} = \frac{1}{6}\,.$$

20. (a) We have

$$y_1 = [(a + b)(1 + \varepsilon_1) \cdot c](1 + \varepsilon_2) \ \text{ where } \ |\varepsilon_i| \le \mathrm{eps}, \quad i = 1, 2.$$

Thus,

$$y_1 = (a + b)c(1 + \varepsilon_1)(1 + \varepsilon_2) \approx (a + b)c(1 + \varepsilon_1 + \varepsilon_2),$$

and

$$|e_1| \approx |\varepsilon_1 + \varepsilon_2| \le |\varepsilon_1| + |\varepsilon_2| \le 2\,\mathrm{eps}.$$

(b) We have

$$y_2 = [ac(1+\varepsilon_1) + bc(1+\varepsilon_2)](1+\varepsilon_3), \quad |\varepsilon_i| \le \text{eps}, \quad i = 1,2,3.$$

Thus, if $a + b \neq 0$,

$$\begin{aligned} y_2 &= [(a+b)c + ac\varepsilon_1 + bc\varepsilon_2](1+\varepsilon_3) \\ &= (a+b)c\left[1 + \frac{a}{a+b}\,\varepsilon_1 + \frac{b}{a+b}\,\varepsilon_2\right](1+\varepsilon_3) \\ &\approx (a+b)c\left[1 + \frac{a}{a+b}\,\varepsilon_1 + \frac{b}{a+b}\,\varepsilon_2 + \varepsilon_3\right], \end{aligned}$$

and

$$\begin{aligned} |e_2| &\approx \left|\frac{a}{a+b}\ \varepsilon_1 + \frac{b}{a+b}\ \varepsilon_2 + \varepsilon_3\right| \\ &\le \left(\frac{|a|}{|a+b|} + \frac{|b|}{|a+b|} + 1\right)\text{eps}. \end{aligned}$$

(c) The distributive law is valid in floating-point arithmetic if and only if $e_1 = e_2$, which almost surely is not the case. In fact, if $|a+b| \ll \max(|a|,|b|)$, then y_2 is significantly less accurate than y_1, i.e., $|e_2| \gg |e_1|$.

21. See the text.

22. (a) Let $\alpha^* = \alpha(1+\varepsilon_\alpha)$. Then $x^{\alpha^*} = x^{\alpha(1+\varepsilon_\alpha)} = x^\alpha \mathrm{e}^{\alpha\varepsilon_\alpha \ln x}$, so that $x^{\alpha^*} \approx x^\alpha(1+\alpha\varepsilon_\alpha \ln x)$, with a propagated relative error $\varepsilon = (\alpha \ln x)\varepsilon_\alpha$. There is a large error propagation if $|\alpha \ln x|$ is large; but if x^α is machine-representable, say $X_+^{-1} \le x^\alpha \le X_+$, then $-\ln X_+ \le \alpha \ln x \le \ln X_+$, and $|\alpha \ln x| \le \ln X_+$. Therefore, $|\alpha \ln x|$ cannot become very large, typically (in the Matlab environment) at most $\ln 10^{308} = 709.196\ldots$.

(b) Let $x^* = x(1+\varepsilon_x)$. Then $(x^*)^\alpha = x^\alpha(1+\varepsilon_x)^\alpha \approx x^\alpha(1+\alpha\varepsilon_x)$, so that now $\varepsilon = \alpha\varepsilon_x$. There is a large error propagation if $|\alpha|$ is very large, which could be serious if $x \approx 1$ and x^α is still machine-representable.

23. Use

$$(x+y)^{1/4} - y^{1/4} = \frac{(x+y)^{1/2} - y^{1/2}}{(x+y)^{1/4} + y^{1/4}} = \frac{x}{[(x+y)^{1/4} + y^{1/4}][(x+y)^{1/2} + y^{1/2}]}\,.$$

24. (a)

(1) With

$$\begin{aligned} a &= .23371258 \times 10^{-4} = .00000023371258 \times 10^2, \\ b &= .33678429 \times 10^2 \ \ = .33678429 \times 10^2, \end{aligned}$$

the exact sum is

$$a + b = .33678452371258 \times 10^2,$$

which, when rounded, gives

$$\mathrm{fl}(a + b) = .33678452 \times 10^2.$$

With

$$c = -.33677811 \times 10^2,$$

this yields

$$\mathrm{fl}(\mathrm{fl}(a + b) + c) = \mathrm{fl}(.00000641 \times 10^2) = .64100000 \times 10^{-3}.$$

Since $\mathrm{fl}(a + b)$ is contaminated by a rounding error, the final addition incurs a severe cancellation error, resulting in an answer correct to only 3 significant digits.

(2) With

$$b = .33678429 \times 10^2,$$
$$c = -.33677811 \times 10^2$$

we get (exactly!)

$$\mathrm{fl}(b + c) = \mathrm{fl}(.00000618 \times 10^2) = .61800000 \times 10^{-3}.$$

With

$$a = .23371258 \times 10^{-4} = .023371258 \times 10^{-3},$$

this yields

$$\mathrm{fl}(a + \mathrm{fl}(b + c)) = \mathrm{fl}(.641371258 \times 10^{-3}) = .64137126 \times 10^{-3},$$

correctly to machine precision.

(b) We have

$$\mathrm{fl}(\mathrm{fl}(a + b) + c) = ((a + b)(1 + \varepsilon_1) + c)(1 + \varepsilon_2) =: s_1^*,$$
$$\mathrm{fl}(a + \mathrm{fl}(b + c)) = (a + (b + c)(1 + \varepsilon_3))(1 + \varepsilon_4) =: s_2^*,$$

where $|\varepsilon_i| \le \mathrm{eps}$, $i = 1, 2, 3, 4$. Thus, with higher-order terms in the ε_i neglected,

$$s_1^* \approx a + b + c + (a + b)\varepsilon_1 + (a + b + c)\varepsilon_2 = s\left(1 + \frac{a + b}{s}\varepsilon_1 + \varepsilon_2\right),$$
$$s_2^* \approx a + b + c + (b + c)\varepsilon_3 + (a + b + c)\varepsilon_4 = s\left(1 + \frac{b + c}{s}\varepsilon_3 + \varepsilon_4\right),$$

where $s = a + b + c$. Therefore,

$$\left|\frac{s_1^* - s}{s}\right| \approx \left|\frac{a+b}{s}\varepsilon_1 + \varepsilon_2\right| \leq \left(\frac{|a+b|}{|s|} + 1\right) \text{eps},$$
$$\left|\frac{s_2^* - s}{s}\right| = \left|\frac{b+c}{s}\varepsilon_3 + \varepsilon_4\right| \leq \left(\frac{|b+c|}{|s|} + 1\right) \text{eps}.$$

The latter bound is smaller than the former if

$$|b + c| < |a + b|.$$

This is clearly the case in the example of (a).

25. We have

$$\begin{aligned} & a^2 - 2ab\cos\gamma + b^2 \\ &= a^2 - 2ab + b^2 + 2ab(1 - \cos\gamma) \\ &= (a-b)^2 + 4ab\sin^2\tfrac{1}{2}\gamma. \end{aligned}$$

If $a = 16.5$, $b = 15.7$, $\gamma = 5°$, then the left-hand side computes to $272. - 2.00 \times 259. \times .996 + 246. = 272. - 516. + 246. = 2.00$, the right-hand side to $.640 + 4.00 \times 259. \times (4.36 \times 10^{-2})^2 = .640 + (1.04 \times 10^3) \times (19.0 \times 10^{-4}) = 2.62$, whereas the exact value is $2.6115269\ldots$. The procedure is still subject to a cancellation error in the formation of $a - b$, which however will be significant only if the second term in the above expression is of comparable (or smaller) magnitude as the first term.

26. Using $(\text{cond}\, f)(x) = \left|\dfrac{xf'(x)}{f(x)}\right|$ for $f\colon \mathbb{R}^1 \to \mathbb{R}^1$, we find

(a) $(\text{cond}\, f)(x) = \dfrac{x}{x|\ln x|} = \dfrac{1}{|\ln x|}$; ill conditioned near $x = 1$.

(b) $(\text{cond}\, f)(x) = |x\, \tan x|$; ill conditioned near $x = \dfrac{\pi}{2}$.

(c) $(\text{cond}\, f)(x) = \left|\dfrac{x}{\sqrt{1-x^2}\, \sin^{-1} x}\right|$; ill conditioned near $|x| = 1$.

(d) $(\text{cond}\, f)(x) = \left|\dfrac{x}{(1+x^2)\sin^{-1}\frac{x}{\sqrt{1+x^2}}}\right|$; always well conditioned.

27. (a) $(\text{cond}\, f)(x) = \left|\dfrac{x \cdot \frac{1}{n}x^{\frac{1}{n}-1}}{x^{\frac{1}{n}}}\right| = \dfrac{1}{n} \leq 1$

Evidently, f is very *well* conditioned for all $x > 0$.

(b) $(\operatorname{cond} f)(x) = \left| \dfrac{x(1 - \frac{x}{\sqrt{x^2-1}})}{x - \sqrt{x^2-1}} \right| = \left| \dfrac{x(\sqrt{x^2-1} - x)}{\sqrt{x^2-1}(x - \sqrt{x^2-1})} \right| = \dfrac{x}{\sqrt{x^2-1}}.$

Here, f is *ill* conditioned near $x = 1$, *well* conditioned as $x \to \infty$.

(c) *detailed analysis* (cf. (1.27), (1.28)):

$$(\operatorname{cond}_{11} \boldsymbol{f})(\boldsymbol{x}) = \frac{x_1 \cdot \frac{1}{2}(x_1^2 + x_2^2)^{-1/2} \cdot 2x_1}{(x_1^2 + x_2^2)^{1/2}} = \frac{x_1^2}{x_1^2 + x_2^2} < 1,$$
$$(\operatorname{cond}_{12} \boldsymbol{f})(\boldsymbol{x}) = \frac{x_2^2}{x_1^2 + x_2^2} < 1.$$

Hence, in the 1-norm,

$$(\operatorname{cond} \boldsymbol{f})(\boldsymbol{x}) = (\operatorname{cond}_{11} \boldsymbol{f})(\boldsymbol{x}) + (\operatorname{cond}_{12} \boldsymbol{f})(\boldsymbol{x}) = 1.$$

Thus, $\boldsymbol{f}$ is *well* conditioned for arbitrary x_1, x_2 with $|x_1| + |x_2| > 0$.
global approach (based on Euclidean norm:

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + x_2^2}, \quad \|[a_1, a_2]\|_2 := \sqrt{a_1^2 + a_2^2},$$
$$(\operatorname{cond} \boldsymbol{f})(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|_2 \|\boldsymbol{f}'(\boldsymbol{x})\|_2}{|\boldsymbol{f}(\boldsymbol{x})|}$$
$$= \frac{\|\boldsymbol{x}\|_2 \|[x_1(x_1^2 + x_2^2)^{-1/2}, x_2(x_1^2 + x_2^2)^{-1/2}]\|_2}{\sqrt{x_1^2 + x_2^2}} = \frac{\|\boldsymbol{x}\|_2 \cdot 1}{\|\boldsymbol{x}\|_2} = 1.$$

(d) *detailed analysis*:

$$(\operatorname{cond}_{11} \boldsymbol{f})(\boldsymbol{x}) = \left| \frac{x_1}{x_1 + x_2} \right|, \quad (\operatorname{cond}_{12} \boldsymbol{f})(\boldsymbol{x}) = \left| \frac{x_2}{x_1 + x_2} \right|,$$
$$(\operatorname{cond} \boldsymbol{f})(\boldsymbol{x}) = (\operatorname{cond}_{11} \boldsymbol{f})(\boldsymbol{x}) + (\operatorname{cond}_{12} \boldsymbol{f})(\boldsymbol{x}) = \frac{|x_1| + |x_2|}{|x_1 + x_2|}.$$

As expected, $\boldsymbol{f}$ is *ill* conditioned when $|x_1 + x_2|$ is very small, but $|x_1|$ (and hence $|x_2|$) is not (cancellation!).
global approach (based on L_1-norm):

$$\|\boldsymbol{x}\|_1 := |x_1| + |x_2|,$$
$$(\operatorname{cond} \boldsymbol{f})(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|_1 \|\boldsymbol{f}'(\boldsymbol{x})\|_1}{|\boldsymbol{f}(\boldsymbol{x})|} = \frac{\|\boldsymbol{x}\|_1 \|[1, 1]\|_1}{|x_1 + x_2|}$$
$$= 2 \, \frac{|x_1| + |x_2|}{|x_1 + x_2|}.$$

28. (a) We have

$$\begin{aligned}(\operatorname{cond} h)(t) &= \left|\frac{th'(t)}{h(t)}\right| = \left|\frac{tg'(f(t))f'(t)}{g(f(t))}\right| \\ &= \left|\frac{f(t)g'(f(t))}{g(f(t))}\right| \cdot \left|\frac{tf'(t)}{f(t)}\right| \\ &= (\operatorname{cond} g)(s) \cdot (\operatorname{cond} f)(t), \quad \text{where } s = f(t).\end{aligned}$$

(b) Here,

$$f(t) = \sin t, \quad g(s) = \frac{1+s}{1-s}\,.$$

Thus,

$$\begin{aligned}(\operatorname{cond} g)(s) &= \left|\frac{s\frac{2}{(1-s)^2}}{\frac{1+s}{1-s}}\right| = \left|\frac{2s}{1-s^2}\right|, \\ (\operatorname{cond} f)(t) &= \left|\frac{t\cos t}{\sin t}\right| = |t\cot t|,\end{aligned}$$

so that, with $t = \frac{\pi}{4},\ s = \sin\frac{\pi}{4} = \frac{1}{\sqrt{2}}$, we get

$$(\operatorname{cond} h)(\tfrac{\pi}{4}) = (\operatorname{cond} g)(\tfrac{1}{\sqrt{2}}) \cdot (\operatorname{cond} f)(\tfrac{\pi}{4}) = \frac{2\cdot\frac{1}{\sqrt{2}}}{1-\frac{1}{2}} \cdot \frac{\pi}{4} = \frac{\pi}{\sqrt{2}}\,.$$

29. We have

$$\begin{aligned}(\operatorname{cond} f\cdot g)(x) &= \left|\frac{x(fg)'(x)}{(fg)(x)}\right| = \left|\frac{xf'(x)g(x)}{f(x)g(x)} + \frac{xf(x)g'(x)}{f(x)g(x)}\right| \\ &\le \left|\frac{xf'(x)}{f(x)}\right| + \left|\frac{xg'(x)}{g(x)}\right| = (\operatorname{cond} f)(x) + (\operatorname{cond} g)(x).\end{aligned}$$

The same answer is obtained for $(\operatorname{cond} f/g)(x)$.

30. (a) $\kappa(x_1,x_2) = \left|\dfrac{x_1}{x_1+x_2}\right| + \left|\dfrac{x_2}{x_1+x_2}\right| = \dfrac{|x_1|+|x_2|}{|x_1+x_2|}$

(b) Symmetry with respect to the bisector b_1 means $\kappa(x_1,x_2) = \kappa(x_2,x_1)$; symmetry with respect to the bisector b_2 means $\kappa(x_1,x_2) = \kappa(-x_2,-x_1)$. Both identities are trivially satisfied.

(c) It suffices to look at the sector $S = \{(x_1,x_2):\ x_1 > 0, |x_2| \le x_1\}$. In the upper half of this sector, where $x_2 > 0$, one has $\kappa(x_1,x_2) \equiv 1$. In the lower half ($x_2 < 0$) one has $\kappa(x_1,x_2) = c$ if and only if $\frac{x_1-x_2}{x_1+x_2} = c$, that is, $x_2 = -\frac{c-1}{c+1}x_1$. Thus, in this part of S, we have $\kappa = \text{constant} = c$ along straight lines through the origin having negative slopes $-\frac{c-1}{c+1}$. In the limiting cases $c = 1$ and $c = \infty$, the slopes are 0 and –1, respectively.

31. (a) By definition,

$$\|\boldsymbol{AB}\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|(\boldsymbol{AB})\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}(\boldsymbol{Bx})\|}{\|\boldsymbol{x}\|}\,.$$

Since for any vector $\boldsymbol{y} \in \mathbb{R}^n$ one has $\|\boldsymbol{Ay}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{y}\|$ (again by definition of the matrix norm), one gets (with $\boldsymbol{y} = \boldsymbol{Bx}$)

$$\|\boldsymbol{AB}\| \leq \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\|\|\boldsymbol{Bx}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{A}\| \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Bx}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{A}\|\|\boldsymbol{B}\|.$$

(b) We have

$$\operatorname{cond}(\boldsymbol{AB}) = \|\boldsymbol{AB}\|\|(\boldsymbol{AB})^{-1}\| = \|\boldsymbol{AB}\|\|\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}\|,$$

hence, by part (a),

$$\begin{aligned}\operatorname{cond}(\boldsymbol{AB}) &\leq \|\boldsymbol{A}\|\|\boldsymbol{B}\| \cdot \|\boldsymbol{B}^{-1}\|\|\boldsymbol{A}^{-1}\| \\ &= \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\| \cdot \|\boldsymbol{B}\|\|\boldsymbol{B}^{-1}\| \\ &= \operatorname{cond}\boldsymbol{A} \cdot \operatorname{cond}\boldsymbol{B}.\end{aligned}$$

32. Using the triangle inequality, one gets

$$\begin{aligned}\|\boldsymbol{Ax}\|_\infty = \max_\nu \left|\sum_\mu a_{\nu\mu}x_\mu\right| &\leq \max_\nu \sum_\mu |a_{\nu\mu}||x_\mu| \\ &\leq \max_\mu |x_\mu| \cdot \max_\nu \sum_\mu |a_{\nu\mu}| = \|\boldsymbol{x}\|_\infty m_\infty,\end{aligned}$$

so that

$$\frac{\|\boldsymbol{Ax}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq m_\infty,$$

hence also

$$\max_{\substack{\boldsymbol{x}\in\mathbb{R}^m \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{Ax}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq m_\infty.$$

This proves

$$\|\boldsymbol{A}\|_\infty \leq m_\infty.$$

On the other hand,

$$\|\boldsymbol{A}\|_\infty = \max_{\substack{\boldsymbol{x}\in\mathbb{R}^m \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{Ax}\|_\infty}{\|\boldsymbol{x}\|_\infty} \geq \frac{\|\boldsymbol{A}\overline{\boldsymbol{x}}\|_\infty}{\|\overline{\boldsymbol{x}}\|_\infty}$$

for any $\overline{\boldsymbol{x}} \in \mathbb{R}^m$, $\overline{\boldsymbol{x}} \neq \boldsymbol{0}$. Let

$$m_\infty = \sum_\mu |a_{\overline{\nu}\mu}|.$$

Choose

$$\overline{\boldsymbol{x}} = [\overline{x}_\mu], \quad \overline{x}_\mu = \operatorname{sgn} a_{\overline{\nu}\mu} \text{ if } a_{\overline{\nu}\mu} \neq 0 \text{ and } 0 \text{ otherwise.}$$

Then (unless $\boldsymbol{A} = \boldsymbol{0}$, in which case there is nothing to prove) $\|\overline{\boldsymbol{x}}\|_\infty = 1$, and

$$\|\boldsymbol{A}\overline{\boldsymbol{x}}\|_\infty = \max_\nu \left| \sum_\mu a_{\nu\mu}\overline{x}_\mu \right| \geq \left| \sum_\mu a_{\overline{\nu}\mu}\overline{x}_\mu \right| = \sum_\mu |a_{\overline{\nu}\mu}| = m_\infty,$$

showing that

$$\|\boldsymbol{A}\|_\infty \geq m_\infty.$$

Consequently, combining the two inequalities, one obtains

$$\|\boldsymbol{A}\|_\infty = m_\infty.$$

33. Using the triangle inequality, one gets

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{x}\|_1 &= \sum_\nu \left| \sum_\mu a_{\nu\mu}x_\mu \right| \leq \sum_\nu \sum_\mu |a_{\nu\mu}||x_\mu| \\ &= \sum_\mu |x_\mu| \sum_\nu |a_{\nu\mu}| \leq \sum_\mu |x_\mu| \cdot \max_\mu \sum_\nu |a_{\nu\mu}| = \|\boldsymbol{x}\|_1 m_1, \end{aligned}$$

so that

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} \leq m_1,$$

hence also

$$\max_{\substack{\boldsymbol{x}\in\mathbb{R}^m \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} \leq m_1.$$

This proves

$$\|\boldsymbol{A}\|_1 \leq m_1.$$

Now let

$$m_1 = \sum_\nu |a_{\nu\overline{\mu}}| = \|\boldsymbol{a}_{\overline{\mu}}\|_1,$$

where $\boldsymbol{a}_{\overline{\mu}}$ is the $\overline{\mu}$th column of $\boldsymbol{A}$, and let $\overline{\boldsymbol{x}} = \boldsymbol{e}_{\overline{\mu}}$ be the coordinate vector in direction of the $\overline{\mu}$th coordinate axis. Then $\|\overline{\boldsymbol{x}}\|_1 = 1$ and

$$\|\boldsymbol{A}\|_1 \geq \frac{\|\boldsymbol{A}\overline{\boldsymbol{x}}\|_1}{\|\overline{\boldsymbol{x}}\|_1} = \|\boldsymbol{a}_{\overline{\mu}}\|_1 = m_1,$$

showing that

$$\|\boldsymbol{A}\|_1 \geq m_1.$$

Consequently, combining the two inequalities, one obtains

$$\|\boldsymbol{A}\|_1 = m_1.$$

34. See the text.

35. The area Δ is a homogeneous linear function of both a and b. Hence,

$$(\mathrm{cond}_a\,\Delta)(a) = (\mathrm{cond}_b\,\Delta)(b) = 1$$

and Δ is perfectly conditioned as a function of a and as a function of b. Not so for γ:

$$(\mathrm{cond}_\gamma\,\Delta)(\gamma) = \left|\frac{\gamma\cdot\frac{1}{2}ab\cos\gamma}{\frac{1}{2}ab\sin\gamma}\right| = \gamma|\cot\gamma|.$$

This is well conditioned near $\gamma = 0$, since

$$\lim_{\gamma\to 0}(\mathrm{cond}_\gamma\,\Delta)(\gamma) = 1,$$

but ill conditioned near $\gamma = \pi$ (a very flat triangle!), since

$$\lim_{\gamma\to \pi}(\mathrm{cond}_\gamma\,\Delta)(\gamma) = \infty.$$

36. (a) Using the *Hint*, in conjunction with Leibniz's rule of differentiation, gives

$$(-1)^k\mathrm{e}^{-x} = x(-1)^k f_k(x) + k(-1)^{k-1}f_{k-1}(x),$$

that is,

$$f_k(x) = \frac{1}{x}(k\ f_{k-1}(x) + \mathrm{e}^{-x}), \quad f_0(x) = \frac{\mathrm{e}^{-x}}{x}.$$

(b) If $x > 0$, then $f_0(x) > 0$, hence by (a) applied with $k = 1, 2, 3, \ldots$, we get $f_k(x) > 0$, all $k \geq 0$. Thus, in the recursion, we are always adding two positive terms, which is a benign operation. If $x < 0$, this is no longer necessarily true (certainly not for $k = 1$). Hence, cancellation errors are a potential threat throughout the recursion.

(c) By repeated application of the recursion of (a), writing y instead of f, one gets

$$y_n = \frac{n!}{x^n}\,y_0 + \cdots, \quad n > 1,$$

where dots indicate some function of x which depends on n but not on y_0. Therefore,

$$(\mathrm{cond}\,y_n)(y_0) = \left|\frac{n!}{x^n}\,\frac{y_0}{y_n}\right|,$$

which for $y_0 = f_0$, hence $y_n = f_n$, becomes

$$(\mathrm{cond}\,y_n)(f_0) = \left|\frac{n!\mathrm{e}^{-x}}{x^{n+1}}\,\frac{1}{f_n(x)}\right|.$$

By Leibniz's rule, and the definition of $f_n(x)$,

$$\begin{aligned} f_n(x) &= (-1)^n \sum_{k=0}^{n} \binom{n}{k} \frac{\mathrm{d}^{n-k}}{\mathrm{d}x^{n-k}} \left(\frac{1}{x}\right) \frac{\mathrm{d}^k}{\mathrm{d}x^k} \left(\mathrm{e}^{-x}\right) \\ &= (-1)^n \sum_{k=0}^{n} \binom{n}{k} \frac{(-1)^{n-k}(n-k)!}{x^{n-k+1}} (-1)^k \mathrm{e}^{-x} \\ &= \frac{1}{x^{n+1}} \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} \frac{(n-k)!}{x^{-k}} \mathrm{e}^{-x} \\ &= \frac{n!\mathrm{e}^{-x}}{x^{n+1}} \sum_{k=0}^{n} \frac{x^k}{k!} = \frac{n!\mathrm{e}^{-x}}{x^{n+1}} e_n(x). \end{aligned}$$

Therefore,

$$(\operatorname{cond} y_n)(f_0) = \frac{1}{|e_n(x)|}.$$

Clearly, when $x > 0$, then

$$e_n(x) > 1, \quad e_n(x) \uparrow \infty,$$

hence $(\operatorname{cond} y_n)(f_0) \downarrow 0$ as $n \to \infty$, and the condition number of y_n at f_0 is less than 1 and gets smaller with increasing n. For $x < 0$, the behavior of $|e_n(x)|$ is more complicated. Note, however, that

$$\lim_{n\to\infty} (\operatorname{cond} y_n)(f_0) = \frac{1}{\mathrm{e}^x} = \mathrm{e}^{|x|} \quad \text{if } x < 0,$$

which is large when $|x|$ is large.

37. Denote $p(x) = x^n + ax - 1$.

(a) Since $p(0) = -1$, $p(+\infty) = +\infty$, and $p'(x) = nx^{n-1} + a > 0$ for positive x, the equation $p(x) = 0$ has exactly one positive root, $\xi(a)$.

(b) From $[\xi(a)]^n + a\xi(a) - 1 \equiv 0$ one gets by differentiation

$$n[\xi(a)]^{n-1}\xi'(a) + \xi(a) + a\xi'(a) = 0,$$

hence

$$\xi'(a) = -\frac{\xi(a)}{a + n[\xi(a)]^{n-1}}.$$

Therefore,

$$(\operatorname{cond} \xi)(a) = \left|\frac{a\xi'(a)}{\xi(a)}\right| = \frac{a\xi(a)}{(a + n[\xi(a)]^{n-1})\xi(a)} = \frac{1}{1 + \frac{n}{a}\,[\xi(a)]^{n-1}}.$$

(c) Since $p(1) = a > 0$, one has $0 < \xi(a) < 1$, which, by the result in (b), implies

$$\frac{1}{1 + \frac{n}{a}} \;<\; \operatorname{cond} \xi(a) < 1.$$

38. (a) Let $p(x) = x^n + x^{n-1} - a$. Then $p'(x) = nx^{n-1} + (n-1)x^{n-2} = x^{n-2}(nx + n - 1) > 0$ for $x > 0$. Since $p(0) = -a < 0$, $p(\infty) > 0$, there is exactly one positive root.

(b) We have

$$[\xi(a)]^n + [\xi(a)]^{n-1} - a \equiv 0.$$

Differentiating, we get

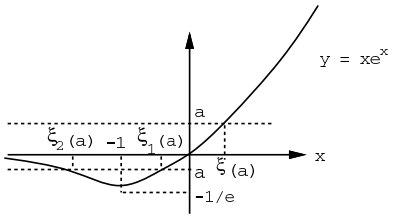
$$n[\xi(a)]^{n-1}\xi'(a) + (n-1)[\xi(a)]^{n-2}\xi'(a) - 1 = 0,$$

$$\begin{aligned}\xi'(a) &= \frac{1}{n[\xi(a)]^{n-1} + (n-1)[\xi(a)]^{n-2}} = \frac{[\xi(a)]}{n[\xi(a)]^n + (n-1)[\xi(a)]^{n-1}} \\ &= \frac{\xi(a)}{n[\xi(a)]^n + (n-1)(a - [\xi(a)]^n)} \\ &= \frac{\xi(a)}{(n-1)a + [\xi(a)]^n}\,.\end{aligned}$$

Therefore,

$$\begin{aligned}(\operatorname{cond}\xi)(a) &= \left|\frac{a\xi'(a)}{\xi(a)}\right| = \frac{a}{(n-1)a + [\xi(a)]^n} \\ &= \frac{1}{n - 1 + \frac{[\xi(a)]^n}{a}} < \frac{1}{n-1}\,.\end{aligned}$$

39. (a) This becomes obvious if one moves up the horizontal line at height a for increasing values of a, beginning with $a = -1/\mathrm{e}$ where the line is tangent to the curve $y = x\mathrm{e}^x$. In the range $-1/\mathrm{e} < a < 0$, there will be exactly two points of intersection with the curve, for $a \geq 0$ exactly one (see figure below).



(b) We have

$$\xi(a)\mathrm{e}^{\xi(a)} \equiv a.$$

Differentiating with respect to a gives

$$\xi'(a)\mathrm{e}^{\xi(a)} + \xi(a)\xi'(a)\mathrm{e}^{\xi(a)} \equiv 1,$$

hence

$$\xi'(a) = \frac{\mathrm{e}^{-\xi(a)}}{1 + \xi(a)}\,.$$

Therefore,

$$(\operatorname{cond}\xi)(a) = \left|\frac{a\xi'(a)}{\xi(a)}\right| = \left|\frac{a\mathrm{e}^{-\xi(a)}}{\xi(a)(1+\xi(a))}\right| = \frac{1}{|1+\xi(a)|}\,.$$

When $a \geq 0$, the unique root $\xi(a)$ is nonnegative, hence $(\operatorname{cond}\xi)(a) < 1$. When $-1/e < a < 0$, the formula holds for both, $\xi_1(a)$ and $\xi_2(a)$. From the figure of part (a), we clearly have

$$(\operatorname{cond}\xi_1)(a) \to \infty \quad (a \downarrow -1/\mathrm{e}), \quad (\operatorname{cond}\xi_1)(a) \to 1 \quad (a \uparrow 0);$$

$$(\operatorname{cond}\xi_2)(a) \to \infty \quad (a \downarrow -1/\mathrm{e}), \quad (\operatorname{cond}\xi_2)(a) \to 0 \quad (a \uparrow 0).$$

40. The equation has a unique positive root since the function $f(x) = x^n - a\mathrm{e}^{-x}$ increases on the interval $[0,\infty]$ from $-a$ to ∞.

Differentiating $[\xi(a)]^n - a\mathrm{e}^{-\xi(a)} \equiv 0$ with respect to a gives

$$n[\xi(a)]^{n-1}\xi'(a) - \mathrm{e}^{-\xi(a)} - a\mathrm{e}^{-\xi(a)}(-\xi'(a)) = 0,$$

$$\xi'(a)(a\mathrm{e}^{-\xi(a)} + n[\xi(a)]^{n-1}) = \mathrm{e}^{-\xi(a)},$$

hence

$$\xi'(a) = \frac{\mathrm{e}^{-\xi(a)}}{a\mathrm{e}^{-\xi(a)} + n[\xi(a)]^{n-1}} = \frac{\mathrm{e}^{-\xi(a)}}{[\xi(a)]^n + n[\xi(a)]^{n-1}} = \frac{\mathrm{e}^{-\xi(a)}}{[\xi(a)]^{n-1}(\xi(a)+n)}\,.$$

Therefore,

$$(\operatorname{cond}\,\xi)(a) = \left|\frac{a\xi'(a)}{\xi(a)}\right| = \frac{a\mathrm{e}^{-\xi(a)}}{[\xi(a)]^n(\xi(a)+n)} = \frac{1}{\xi(a)+n}\,.$$

Since $\xi(a) > 0$, one clearly has

$$(\operatorname{cond}\,\xi)(a) < \frac{1}{n}\,.$$

41. By the law of machine arithmetic,

$$y_A = \mathrm{fl}(x_1 + x_2) = (x_1 + x_2)(1+\varepsilon), \quad |\varepsilon| \leq \mathrm{eps}.$$

Hence, we can write

$$y_A = x_1(1+\varepsilon) + x_2(1+\varepsilon) = \boldsymbol{f}(\boldsymbol{x}_A), \quad \boldsymbol{x}_A = [x_1(1+\varepsilon),\ x_2(1+\varepsilon)]^T.$$

There follows

$$\frac{\|\boldsymbol{x} - \boldsymbol{x}_A\|_1}{\|\boldsymbol{x}\|_1} = \frac{|x_1\varepsilon| + |x_2\varepsilon|}{|x_1| + |x_2|} \leq \frac{|x_1| + |x_2|}{|x_1| + |x_2|}\ \mathrm{eps} = \mathrm{eps},$$

which, by (1.65), means that

$$\gamma(x_1, x_2) = (\operatorname{cond} A)(x) \leq 1.$$

(Taking $\boldsymbol{x}_A = [x_1, x_1\varepsilon + x_2(1+\varepsilon)]$ would yield the sharper estimate $\gamma(x_1, x_2) \leq \frac{|x_1+x_2|}{|x_1|+|x_2|}$.) The same is obtained for the other norms. The algorithm A is thus perfectly conditioned. Nevertheless, the result y_A can be inaccurate if $\boldsymbol{f}$ is ill conditioned (cf. Problem 27(d) or 30):

$$\left|\frac{y_A - y}{y}\right| = \left|\frac{\boldsymbol{f}(\boldsymbol{x}_A) - \boldsymbol{f}(\boldsymbol{x})}{\boldsymbol{f}(\boldsymbol{x})}\right| \approx (\operatorname{cond} \boldsymbol{f})(\boldsymbol{x})\ \frac{\|\boldsymbol{x}_A - \boldsymbol{x}\|}{\|\boldsymbol{x}\|}\ \leq\ (\operatorname{cond} \boldsymbol{f})(\boldsymbol{x})\,\mathrm{eps}.$$

The inaccuracy, however, must be blamed on $\boldsymbol{f}$, not on the algorithm A.

42. (a) The condition of f is given by

$$(\operatorname{cond} f)(x) = \left|\frac{x\left(\frac{1}{2}\ \frac{-1}{\sqrt{1-x}}\right)}{\sqrt{1-x}-1}\right| \ =\ \frac{1}{2}\ \frac{|x|}{\sqrt{1-x}}\ \frac{1}{|\sqrt{1-x}-1|}$$
$$=\ \frac{1}{2}\left(1 + \frac{1}{\sqrt{1-x}}\right).$$

(b) We have

$$\begin{aligned} f_A(x) &= \left(\sqrt{(1-x)(1+\varepsilon_1)}(1+\varepsilon_2) - 1\right)(1+\varepsilon_3) \\ &\approx \left(\sqrt{1-x}\left(1+\tfrac{1}{2}\,\varepsilon_1\right)(1+\varepsilon_2) - 1\right)(1+\varepsilon_3) \\ &\approx \sqrt{1-x} - 1 + \sqrt{1-x}\left(\tfrac{1}{2}\,\varepsilon_1 + \varepsilon_2\right) + \left(\sqrt{1-x} - 1\right)\varepsilon_3. \end{aligned}$$

Setting this equal to $f(x_A)$ gives

$$\sqrt{1-x} - 1 + \sqrt{1-x}\left(\tfrac{1}{2}\,\varepsilon_1 + \varepsilon_2\right) + \left(\sqrt{1-x} - 1\right)\varepsilon_3 = \sqrt{1-x_A} - 1,$$

thus

$$\sqrt{1-x}\left(1 + \tfrac{1}{2}\,\varepsilon_1 + \varepsilon_2\right) + \varepsilon_3\left(\sqrt{1-x} - 1\right) = \sqrt{1-x_A}.$$

Squaring both sides, and neglecting higher-order terms in the ε_i yields

$$(1-x)(1+\varepsilon_1+2\varepsilon_2) + 2\sqrt{1-x}\left(\sqrt{1-x} - 1\right)\varepsilon_3 = 1 - x_A,$$

hence

$$-x + (1-x)(\varepsilon_1 + 2\varepsilon_2) + 2\sqrt{1-x}\left(\sqrt{1-x} - 1\right)\varepsilon_3 = -x_A,$$

$$x_A = x\left\{1 - \frac{1-x}{x}\ (\varepsilon_1 + 2\varepsilon_2) - 2\ \frac{\sqrt{1-x}}{x}\left(\sqrt{1-x} - 1\right)\varepsilon_3\right\},$$

that is,

$$\varepsilon_A = -\frac{1-x}{x}\,(\varepsilon_1 + 2\varepsilon_2) - 2\,\frac{\sqrt{1-x}}{x}\left(\sqrt{1-x}-1\right)\varepsilon_3.$$

There follows

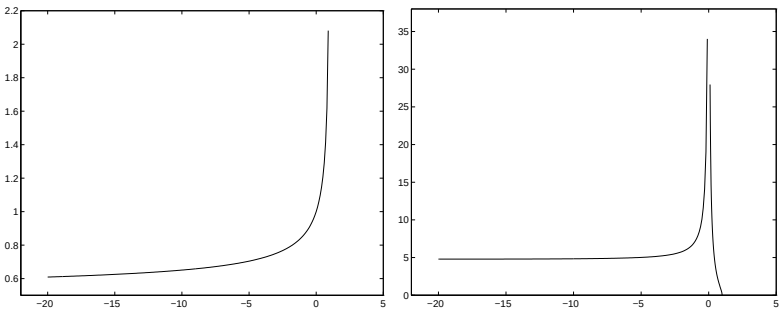
$$\begin{aligned}|\varepsilon_A| &\le \frac{1-x}{|x|}\ 3\ \text{eps}\ + 2\,\frac{\sqrt{1-x}}{|x|}\left|\sqrt{1-x}-1\right|\ \text{eps}\\ &= \left(3\,\frac{1-x}{|x|}\ +\ 2\,\frac{\sqrt{1-x}}{1+\sqrt{1-x}}\right)\ \text{eps}.\end{aligned}$$

Hence,

$$(\operatorname{cond} A)(x) \le 3\,\frac{1-x}{|x|}\ +\ 2\,\frac{\sqrt{1-x}}{1+\sqrt{1-x}}\,.$$

(c)

$(\operatorname{cond} f)(x)$ $(\operatorname{cond} A)(x)$



The ill-conditioning of A at $x = 0$ is a reflection of the cancellation phenomenon. Interestingly, the algorithm A is perfectly conditioned at $x = 1$ where f is ill conditioned.

43. (a) Given $f(x) = 1 - \mathrm{e}^{-x}$, we have, for $0 \le x \le 1$,

$$\begin{aligned}(\operatorname{cond} f)(x) &= \left|\frac{xf'(x)}{f(x)}\right| = \frac{x\mathrm{e}^{-x}}{1-\mathrm{e}^{-x}} = \frac{x}{\mathrm{e}^x - 1}\\ &= \frac{x}{x + \dfrac{x^2}{2!} + \dfrac{x^3}{3!} + \cdots} \le \frac{x}{x} = 1.\end{aligned}$$

(b) We have

$$f_A(x) = [1 - \mathrm{e}^{-x}(1+\varepsilon_1)](1+\varepsilon_2),\ |\varepsilon_i| \le\ \text{eps},\quad i = 1, 2.$$

Neglecting terms of $O(\text{eps}^2)$, we get

$$f_A(x) = 1 - \mathrm{e}^{-x} - \varepsilon_1 \mathrm{e}^{-x} + \varepsilon_2(1 - \mathrm{e}^{-x}).$$

We must set $f_A(x) = f(x_A)$ and estimate the relative vicinity of x_A to x:

$$1 - \mathrm{e}^{-x} - \varepsilon_1 \mathrm{e}^{-x} + \varepsilon_2(1 - \mathrm{e}^{-x}) = 1 - \mathrm{e}^{-x_A}$$

gives

$$\begin{aligned} \mathrm{e}^{-x_A} &= \mathrm{e}^{-x} + \varepsilon_1 \mathrm{e}^{-x} - \varepsilon_2(1 - \mathrm{e}^{-x}), \\ -x_A &= \ln\{\mathrm{e}^{-x}[1 + \varepsilon_1 - \varepsilon_2(\mathrm{e}^x - 1)]\} \\ &= -x + \ln(1 + \varepsilon_1 - \varepsilon_2(\mathrm{e}^x - 1)) \\ &= -x + \varepsilon_1 - \varepsilon_2(\mathrm{e}^x - 1) + O(\text{eps}^2). \end{aligned}$$

Therefore, up to terms of order eps^2,

$$|x - x_A| = |\varepsilon_1 - \varepsilon_2(\mathrm{e}^x - 1)| \le \text{eps} + (\mathrm{e}^x - 1)\text{eps} = \mathrm{e}^x \text{eps},$$

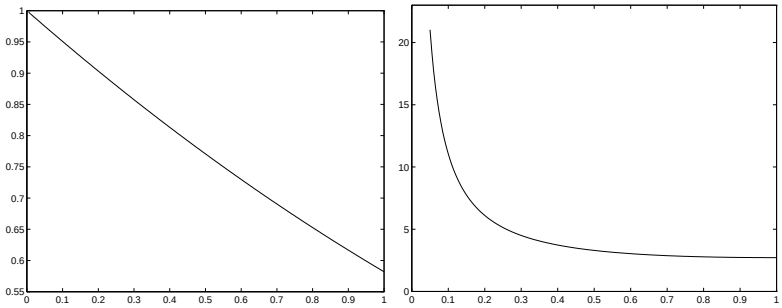
$$\frac{|x - x_A|}{|x|} \le \frac{\mathrm{e}^x}{x}\ \text{eps},$$

that is,

$$(\operatorname{cond} A)(x) \le \frac{\mathrm{e}^x}{x}\,.$$

(c)

$(\operatorname{cond} f)(x)$ $(\operatorname{cond} A)(x)$



While f is uniformly well conditioned on [0,1], the algorithm A is severely ill conditioned near $x = 0$, which is just a reflection of the large cancellation error incurred by the algorithm for small x.

44. See the text.

ANSWERS TO MACHINE ASSIGNMENTS

1. PROGRAM

```
%MAI_1
%
f0='%8.0f %13.5e %12.5e %12.5e %8.4f\n';
disp(['    n          p            dp          err' ...
'     err/(n*eps0)'])
eps0=eps('single');
x=single(1+pi/10^6); dx=1+pi/10^6;
for n=100000:100000:1000000
  pn=x^n; dpn=dx^n;
  err=abs((pn-dpn)/dpn);
  fprintf(f0,n,pn,dpn,err,err/(n*eps0))
end
```

OUTPUT

```
>> MAI_1
   n          p            dp          err     err/(n*eps0)
 100000   1.36335e+00  1.36911e+00  4.20619e-03   0.3528
 200000   1.85872e+00  1.87445e+00  8.39477e-03   0.3521
 300000   2.53408e+00  2.56633e+00  1.25657e-02   0.3514
 400000   3.45483e+00  3.51358e+00  1.67191e-02   0.3506
 500000   4.71014e+00  4.81047e+00  2.08549e-02   0.3499
 600000   6.42157e+00  6.58604e+00  2.49735e-02   0.3492
 700000   8.75483e+00  9.01700e+00  2.90747e-02   0.3484
 800000   1.19359e+01  1.23452e+01  3.31586e-02   0.3477
 900000   1.62728e+01  1.69019e+01  3.72254e-02   0.3470
1000000   2.21855e+01  2.31406e+01  4.12750e-02   0.3462
>>
```

For $n = 10^6$, the result should be $x^n \approx e^\pi = 23.1406926\ldots$. This is confirmed in the double-precision result. The single-precision results, even though we are in a case of benign error propagation, are much less accurate. The reason for this is the large number n of multiplies. The relative errors indeed are approximately proportional to $n \cdot \text{eps0}$, as can be seen from the last column.

2. PROGRAM

```
%MAI_2
%
f0='%5.0f %12.4e %12.4e %14.10f %14.10f\n';
disp(['    i      h1           h2             d1' ...
'             d2'])
for i=5:5:50
  h1=2^(-i); h2=2.2^(-i);
  d1=(exp(h1)-1)/h1; d2=(exp(h2)-1)/h2;
  fprintf(f0,i,h1,h2,d1,d2)
end
```

OUTPUT

```
>> MAI_2
    i      h1           h2             d1             d2
    5   3.1250e-02   1.9404e-02   1.0157890400   1.0097649524
   10   9.7656e-04   3.7651e-04   1.0004884402   1.0001882772
   15   3.0518e-05   7.3057e-06   1.0000152589   1.0000036528
   20   9.5367e-07   1.4176e-07   1.0000004768   1.0000000703
   25   2.9802e-08   2.7506e-09   1.0000000149   0.9999999856
   30   9.3132e-10   5.3373e-11   1.0000000000   0.9999979434
   35   2.9104e-11   1.0356e-12   1.0000000000   0.9999829518
   40   9.0949e-13   2.0095e-14   1.0000000000   1.0055156877
   45   2.8422e-14   3.8992e-16   1.0000000000   1.1389138076
   50   8.8818e-16   7.5660e-18   1.0000000000   0.0000000000
>>
```

When $h = 2^{-i}$, for large i one has $e^h - 1 \approx (1 + h) - 1$ and $1 + h$ is exactly machine-representable. Hence, the subtraction of 1 from it entails *no* rounding error, and the exact result $\frac{\mathrm{d}}{\mathrm{d}x}e^x\,|_{x=0} = 1$ is obtained. Not so if $h = (2.2)^{-i}$, where $.2 = 1/5$ has an *infinite* binary representation (cf. Sect. 1.1.1, Example (3)). Hence, serious cancellation errors ensue, which become evident in the last column beginning at about $i = 30$. From $i = 47$ onward, e^{h_2} becomes exactly 1, yielding $d(h_2) = 0$.

3. (a)

PROGRAM

```
%MAI_3
%
d0=0; d1=exp(1)-1; h=1;
n=0;
while d1~=d0
```

```
  n=n+1;
  h=h/2;
  d0=d1; d1=(exp(h)-1)/h;
end
fprintf('limit = %1.0f for n = %2.0f\n',d1,n)

 OUTPUT

>> MAI_3
limit = 1 for n = 27
>>
```

In Matlab, which uses the ANSI/IEEE Standard for double-precision floating-point arithmetic (cf. the Notes to Sect. 1.1.2(1)), one has $t = 52$, hence indeed $n_0 = 27$ (cf. part (b)).

(b) The value $n = n_0$ in question is the smallest n for which

$$\mathrm{fl}(\mathrm{e}^{2^{-(n-1)}}) = 1 + 2^{-(n-1)} \quad \text{and} \quad \mathrm{fl}(\mathrm{e}^{2^{-n}}) = 1 + 2^{-n}.$$

Thus, in view of the expansion of e^h for small h, and the rounding procedure used (chopping), n_0 is the smallest n for which

$$\frac{1}{2}\,(2^{-(n-1)})^2 \;\le\; 2^{-(t+1)}, \quad \text{i.e.,} \quad n \;\ge\; \frac{t+2}{2}\,.$$

Therefore,

$$n_0 = \left\lceil \frac{t+2}{2} \right\rceil.$$

(c) The smallest positive floating-point number is (cf. Sect. 1.1.2, Eq. (1.6))

$$\frac{1}{2}\; 2^{-(1+2+\cdots+2^{s-1})} = \frac{1}{2}\; 2^{-(2^s-1)} = 2^{-2^s}.$$

Therefore, underflow occurs (for the first time) when $n = n_1$, where n_1 is the smallest n for which

$$2^{-n} < 2^{-2^s}, \quad \text{i.e.,} \quad n > 2^s.$$

Thus, $n_1 = 2^s + 1$, and we have $n_1 < n_0$ precisely if

$$2^s + 1 < \left\lceil \frac{t+2}{2} \right\rceil.$$

4. PROGRAM

```
%MAI_4
%
```

```
f0='%8.0f %14.8f\n';
disp('     k            c')
gamma=.57721566490153286;
sum=0;
for k=1:10000
  sum=sum+1/k;
  if 1000*fix(k/1000)==k
    gam=sum-log(k);
    c=k*(gamma-gam);
    fprintf(f0,k,c)
  end
end
```

```
 OUTPUT (assuming d=1)

>> MAI_4
     k           c
    1000    -0.49991667
    2000    -0.49995833
    3000    -0.49997222
    4000    -0.49997917
    5000    -0.49998333
    6000    -0.49998611
    7000    -0.49998809
    8000    -0.49998958
    9000    -0.49999074
   10000    -0.49999167
>>
```

The results strongly suggest that $d = 1$ and $c = -\frac{1}{2}$. This can be verified from Eqs. 6.3.2 and 6.3.18 of Abramowitz and Stegun [1964] or Eqs. 5.4.14 and 5.11.2 of Olver et al. [2010].

5. PROGRAM

```
%MAI_5
%
f0='%8.0f %12.4e %12.4e\n';
disp('     N       err1         err2')
for N=1000:1000:10000
  s1=0; s2=0;
  for n=1:N
    s1=s1+single(log(n+2))-single(log(n+1));
    s2=s2+single(log(n+1));
  end
  ds1=log(N+2)-log(2);
```

```
  ds2=0;
  for n=1:N
    ds2=ds2+log(n+1);
  end
  err1=abs((s1-ds1)/ds1);
  err2=abs((s2-ds2)/ds2);
  fprintf(f0,N,err1,err2)
end

 OUTPUT

>> MAI_5
     N       err1         err2
    1000   2.3639e-07   2.7723e-07
    2000   1.3181e-06   9.0833e-08
    3000   1.2492e-05   1.4449e-08
    4000   1.2037e-05   3.2598e-07
    5000   1.4112e-05   1.9835e-08
    6000   1.9850e-05   2.9067e-07
    7000   2.0318e-05   3.1665e-07
    8000   2.1903e-05   4.6113e-07
    9000   2.0537e-05   2.7735e-07
   10000   2.3618e-05   3.1400e-07
>>
```

The first summation is subject to increasing amounts of cancellation, since for large n the general term $\ln(2+n) - \ln(1+n)$ involves the subtraction of almost equal numbers, both being rounded values of the logarithm function. Thus, e.g., if $n = 10\,000$, then $\ln(n+2) = 9.2105408$ in single precision, while $\ln(n+1) = 9.2104406$, the difference being 1.001×10^{-4}, implying a cancellation error of about 10^{-4}. However, considering that early on in the summation, cancellation is much less severe, and that in accumulating errors of mixed signs some error compensation is likely to occur, one expects the total error to be relatively small. Our computations suggest a total relative error in the range 2×10^{-7} (for $N = 1\,000$) to 2×10^{-5} (for $N = 10\,000$). In contrast, for the second summation (where there are no cancellation effects), the total relative error is seen to be much smaller, in the range 10^{-8} (for $N = 3\,000$) to 5×10^{-7} (for $N = 8\,000$).

6. (a) The exact answer is $S_N = 1 - \frac{1}{N+1}$.

```
 PROGRAM

%MAI_6A
%
f0='%10.0f %12.4e %12.4e\n';
```

```
disp('      N         err1         err2')
N=1;
for k=1:7
  N=10*N;
  s1=0; s2=0;
  for n=1:N
    s1=s1+1/n-1/(n+1);
    s2=s2+1/(n*(n+1));
  end
  err1=abs(s1-1+1/(N+1));
  err2=abs(s2-1+1/(N+1));
  fprintf(f0,N,err1,err2)
end
```

OUTPUT

```
>> MAI_6A
      N         err1         err2
        10   2.7756e-17   2.7756e-17
       100   8.6736e-18   3.4174e-16
      1000   3.3827e-17   6.9996e-16
     10000   1.6602e-17   6.4953e-16
    100000   5.4574e-17   1.3155e-14
   1000000   4.9877e-17   4.7579e-14
  10000000   4.4644e-17   1.9469e-13
>>
```

Paradoxically, the first summation is more accurate than the second. One would have expected otherwise, since the computation of each term is subject to a cancellation error which is growing as the summation progresses. In reality, however, the "telescoping" effect present mathematically is also at work computationally, even if partial terms have to be rounded: the same rounded quantity, once subtracted, is added again in the next term! In the second summation, the rounding errors committed in computing the terms of the series keep accumulating.

(b) The exact answer is $p_N = \frac{1}{N+1}$.

PROGRAM

```
%MAI_6B
%
f0='%10.0f %12.4e\n';
disp('      N        err')
N=1;
for k=1:7
  N=10*N;
```

```
  p=1;
  for n=1:N
    p=n*p/(n+1);
  end
  err=abs((N+1)*p-1);
  fprintf(f0,N,err)
end

 OUTPUT

>> MAI_6B
      N        err
        10   0.0000e+00
       100   1.1102e-16
      1000   1.5543e-15
     10000   5.3291e-15
    100000   1.0991e-14
   1000000   1.2612e-13
  10000000   4.6718e-13
>>
```

Although, mathematically, the telescoping effect is still present, there is no longer a numerical analogue thereof. Each term is now subject to rounding. The behavior of the error therefore is similar to the one in the second summation of (a).

7. See the text.

8. (a) The exact value is

$$s = 10^{11}n + \ln n!,$$

which by Stirling's formula is asymptotically equivalent to

$$10^{11}n + (n + \tfrac{1}{2})\ln(n+1) - (n+1) + \tfrac{1}{2}\ln 2\pi \quad \text{as } n \to \infty.$$

For $n = 10^6$ this is $1.000000000128155 \times 10^{17}$.

(b) The Matlab program yields the results

```
1.000000000159523e+17
1.000000000128155e+17
1.000000000128155e+17.
```

The last two are accurate to all 16 digits shown, the first only in the first eleven digits. One might argue that many of the initial terms of the series in the first summation are shifted out of range when added to $10^{11}n = 10^{17}$ and only those which are large enough will contribute to the sum. In the second and third summations, all terms in the series are fully accounted for. However, if this were the only reason for the

discrepancy, the first result would have to be smaller than the other two. But it is larger! Therefore, there must be other effects at work, most likely rounding effects. The summation producing the first result consistently adds large numbers to considerably smaller ones, which is not optimal in the presence of rounding errors (cf. MA 7(a)). A summation, like the other two, that adds numbers in increasing order, is preferable.

9. (a) Following the *Hint*, one finds

$$\left(\boldsymbol{H}_n^{-1}\right) = (-1)^{i+j} f_i f_j / ((i+j-1)(i-1)!^2 (j-1)!^2),$$

where

$$f_k = (n-k+1)(n-k+2)\cdots(n+k-1).$$

(b)

```
 PROGRAM

%
%MAI_9
f0='%8.0f %12.4e\n';
for n=[10 20 40]
  H=hilb(n); Hinv=zeros(n);
  for i=1:n
    for j=1:n
      ki=-i+1:i-1; kj=-j+1:j-1;
      Hinv(i,j)=((-1)^(i+j))*prod(n+ki)*prod(n+kj) ...
        /((i+j-1)*(factorial(i-1))^2*(factorial(j-1))^2);
    end
  end
  condH=max(eig(H))*max(eig(Hinv));
  fprintf(f0,n,condH)
end

 OUTPUT

>> MAI_9
      10   1.6026e+13
      20   2.4522e+28
      40   7.6529e+58
>>
```

10. (a) Use

$$\operatorname{sinc}\left(\frac{x-kh}{h}\right) = \operatorname{sinc}\left(\frac{x}{h}-k\right) = \frac{\sin\pi\left(\frac{x}{h}-k\right)}{\pi\left(\frac{x}{h}-k\right)} = \frac{h}{\pi}\,\frac{(-1)^k \sin\left(\pi\frac{x}{h}\right)}{x-kh}$$

in the definition of $C_N(f,h)(x)$.

(b) If x is very close to one of the abscissae kh, one term in the summation is extremely large in absolute value and overshadows all of the other terms. Cf. MA 8.

(c) Following the *Hint* and using

$$\sin\frac{\pi x}{h} = \sin(\pi(k_0+t)) = (-1)^{k_0}\sin\pi t,$$

one obtains

$$C_N(f,h)(x) = (-1)^{k_0}\frac{\sin\pi t}{\pi t}\sum_{|k|\le N}\frac{(-1)^k t}{t+k_0-k}\,f(kh).$$

Introducing the new index of summation $\kappa = k - k_0$, and separating out the term with $\kappa = 0$ (which occurs if we assume $|k_0| \le N$), one finds

$$C_N(f,h)(x) = \frac{\sin\pi t}{\pi t}\left\{f(k_0h) + \sum_{\kappa}\frac{(-1)^\kappa t}{t-\kappa}\,f((k_0+\kappa)h)\right\},$$

where the summation is from $-N-k_0$ to $N-k_0$ with $\kappa = 0$ omitted.

(d) The following Matlab program computes the cardinal series according to the formula in (a) and in (c) for the values of N, h, and x specified.

PROGRAM

```
%MAI_10D
%
f0='%19.15f %19.15f %12.4e\n';
f1='%19.15f %32.4e\n';
disp('        card                  f                err')
h=.1; N=100;
for x=[.55 .5+1e-8 .5+1e-15]
  fexact=x*exp(-x^2);
%
%  computing the cardinal series by the
%  formula in (a)
%
  k=-N:N;
  card0=(h/pi)*sin(pi*x/h)*sum((-1).^k.*(h*k.* ...
    exp(-(h*k).^2))./(x-h*k));
  err=abs(card0-fexact);
  fprintf(f0,card0,fexact,err)
%
% computing the cardinal series by the
```

```
% formula in (c)
  k0=round(x/h); t=x/h-k0;
  s=1;
  if t~=0, s=sin(pi*t)/(pi*t); end
  k=nonzeros(-N-k0:N-k0);
  card=s*(h*k0*exp(-(h*k0)^2)+sum((-1).^k.*(t* ...
    h*(k0+k).*exp(-(h*(k0+k)).^2))./(t-k)));
  err=abs(card-fexact);
  fprintf(f1,card,err)
  fprintf('\n')
end
```

OUTPUT

```
>> MAI_10D
        card                 f               err
  0.406432668542419   0.406432668542419   2.7756e-16
  0.406432668542419                       0.0000e+00

  0.389400395354676   0.389400395429706   7.5030e-11
  0.389400395429706                       5.5511e-17

  0.367008607887599   0.389400391535703   2.2392e-02
  0.389400391535703                       0.0000e+00
>>
```

The deterioration in accuracy in the first lines as x approaches .5 is clearly evident, whereas in the second lines the accuracy remains perfect.

11. See the text.

12. The series in (a) is $\sum_{n=0}^{\infty} t_n$, where the terms t_n can be generated recursively by

$$t_0 = 1; \quad t_n = -t_{n-1}/n^2, \quad n = 1, 2, 3, \ldots \ .$$

Likewise, for the series $\sum_{n=0}^{\infty} u_n$ in (b), we have

$$u_0 = 1; \quad u_n = u_{n-1}/n^2, \quad n = 1, 2, 3, \ldots \ .$$

PROGRAM

```
%MAI_12
%
% series in (a)
%
s1=1; s0=0; term=1; n=0;
```

```
while s1~=s0
  n=n+1;
  s0=s1;
  term=-term/n^2;
  s1=s0+term;
end
J0=besselj(0,2);
fprintf(['n=%3.0f, series in (a) =%19.16f,' ...
' J_0(2) =%19.16f\n'],n,s1,J0)
%
% series in (b)
%
s1=1; s0=0; term=1; n=0;
while s1~=s0
  n=n+1;
  s0=s1;
  term=term/n^2;
  s1=s0+term;
end
I0=besseli(0,2);
fprintf(['n=%3.0f, series in (b) =%18.15f,' ...
'  I_0(2) =%18.15f\n'],n,s1,I0)
```

OUTPUT

```
>> MAI_12
n= 12, series in (a) = 0.2238907791412356, J_0(2) = 0.2238907791412356
n= 12, series in (b) = 2.279585302336067,  I_0(2) = 2.279585302336067
>>
```

The accuracy is perfect to 16 significant digits, there being no cancellation of terms, not even in the alternating series, since the terms decrease rapidly in absolute value.

13. PROGRAM

```
%MAI_13
%
f0='%10.0f %11.8f\n';
disp('      n        s_k')
n=10000000; s=0;
for k=1:n
  s=s+1/(k^(1/2)*(k+1));
  if 1000000*fix(k/1000000)==k
    fprintf(f0,k,s)
```

```
  end
end

 OUTPUT

>> MAI_13
      n        s_k
   1000000  1.85802508
   2000000  1.85861087
   3000000  1.85887038
   4000000  1.85902508
   5000000  1.85913065
   6000000  1.85920858
   7000000  1.85926915
   8000000  1.85931797
   9000000  1.85935841
  10000000  1.85939262
>>
```

The sum, known as the Theodorus constant, has the exact value

$$s = 1.86002507922\ldots$$

(cf. Gautschi [2010, Table 1]). Evidently, the series converges extremely slowly.

14. For n sufficiently large, $1 + 1/n$ in finite-precision arithmetic, will become equal to 1. From this point on, the "machine limit" will be 1 instead of e. This is borne out by the following computer run.

```
 PROGRAM

%MAI_14
%
f0='%12.2e %19.15f\n';
disp('       n           (1+1/n)^n')
for in=2:2:20
  n=10^in;
  e=(1+1/n)^n;
  fprintf(f0,n,e)
end

 OUTPUT

>> MAI_14
     n           (1+1/n)^n
```

```
  1.00e+02   2.704813829421528
  1.00e+04   2.718145926824926
  1.00e+06   2.718280469095753
  1.00e+08   2.718281798347358
  1.00e+10   2.718282053234788
  1.00e+12   2.718523496037238
  1.00e+14   2.716110034087023
  1.00e+16   1.000000000000000
  1.00e+18   1.000000000000000
  1.00e+20   1.000000000000000
>>
```

15. PROGRAM

```
%MAI_15
%
f0='%6.0f %22.15e  n=%1.0f\n';
f1='%6.0f %22.15e\n';
%f2='%9.2e %9.2e %9.2e %9.2e %9.2e\n';
disp('       n            x')
%R=zeros(5,5);
for n=1:5
  for K=10:10:50
    x=1/n;
    for k=1:K
      x=(n+1)*x-1;
    end
%   R(K/10,n)=n*x/((n+1)^K);
    if K==10
      fprintf(f0,K,x,n)
    else
      fprintf(f1,K,x)
    end
  end
  fprintf('\n')
end
%fprintf(f2,R')I_14
```

OUTPUT

```
>> MAI_15
    K            x
   10  1.000000000000000e+00  n=1
   20  1.000000000000000e+00
   30  1.000000000000000e+00
```

```
40  1.000000000000000e+00
50  1.000000000000000e+00

10  5.000000000000000e-01  n=2
20  5.000000000000000e-01
30  5.000000000000000e-01
40  5.000000000000000e-01
50  5.000000000000000e-01

10  3.333333333139308e-01  n=3
20  3.333129882812500e-01
30 -2.100000000000000e+01
40 -2.236962100000000e+07
50 -2.345624805922100e+13

10  2.500000000000000e-01  n=4
20  2.500000000000000e-01
30  2.500000000000000e-01
40  2.500000000000000e-01
50  2.500000000000000e-01

10  2.000000017901584e-01  n=5
20  3.082440341822803e-01
30  6.545103021815777e+06
40  3.957573391620096e+14
50  2.392993292306176e+22
>>
```

With $x_0 = \frac{1}{n}$ the iteration, in exact arithmetic, converges in one step since $f(\frac{1}{n}) = \frac{n+1}{n} - 1 = \frac{1}{n}$. In floating-point arithmetic, assume that the only error committed is the error in computing the starting value,

$$x_0^* = \frac{1}{n}(1+\varepsilon).$$

Then by induction on k, one finds that

$$x_k^* = \frac{1}{n}(1+\varepsilon_k), \quad \text{where} \;\; \varepsilon_k = (n+1)^k \varepsilon.$$

Below is a table of $(n+1)^k$.

	n				
k	1	2	3	4	5
10	1.02e+03	5.90e+04	1.05e+06	9.77e+06	6.05e+07
20	1.05e+06	3.49e+09	1.10e+12	9.54e+13	3.66e+15
30	1.07e+09	2.06e+14	1.15e+18	9.31e+20	2.21e+23
40	1.10e+12	1.22e+19	1.21e+24	9.09e+27	1.34e+31
50	1.13e+15	7.18e+23	1.27e+30	8.88e+34	8.08e+38

It can be seen that the error magnification can be very substantial, as is already evident in the OUTPUT. There is no error magnification whatsoever when $n = 1, 2, 4$, since in these cases $\varepsilon = 0$.

16. PROGRAM

```
%MAI_16
%
f0='%8.0f %19.15f %11.4e\n';
disp('     n           Riemann          err')
for n=5000:5000:100000
  k=1:n;
  riem=sum(exp((k-1/2)/n))/n;
  int=exp(1)-1;
  err=abs(riem-int);
  fprintf(f0,n,riem,err)
end
```

OUTPUT

```
>> MAI_16
     n          Riemann          err
    5000   1.718281825595241  2.8638e-09
   10000   1.718281827743095  7.1595e-10
   15000   1.718281828140854  3.1819e-10
   20000   1.718281828280041  1.7900e-10
   25000   1.718281828344491  1.1455e-10
   30000   1.718281828379489  7.9557e-11
   35000   1.718281828400596  5.8449e-11
   40000   1.718281828414301  4.4745e-11
   45000   1.718281828423698  3.5348e-11
   50000   1.718281828430396  2.8649e-11
   55000   1.718281828435387  2.3658e-11
   60000   1.718281828439153  1.9893e-11
   65000   1.718281828442101  1.6945e-11
   70000   1.718281828444444  1.4602e-11
   75000   1.718281828446330  1.2716e-11
   80000   1.718281828447866  1.1180e-11
   85000   1.718281828449122  9.9238e-12
   90000   1.718281828450222  8.8238e-12
   95000   1.718281828451119  7.9263e-12
  100000   1.718281828451882  7.1634e-12
>>
```

Since only benign arithmetic operations are involved (summation of positive terms) the results are as accurate as one can expect. However, it takes a great

many subdivisions to achieve a relatively high degree of accuracy. Chapter 3, Sect. 3.2.2, will show how to get the same accuracy with only five, albeit nonuniform, subdivisions.

17. (a) The starting value is $y_0 = \int_0^1 \mathrm{e}^{-t}\mathrm{d}t = 1 - \mathrm{e}^{-1}$. For $k > 0$, integration by parts gives

$$y_k = \int_0^1 t^k \mathrm{e}^{-t}\mathrm{d}t = -t^k \mathrm{e}^{-t}\Big|_0^1 + k\int_0^1 t^{k-1}\mathrm{e}^{-t}\mathrm{d}t$$
$$= -\mathrm{e}^{-1} + k\int_0^1 t^{k-1}\mathrm{e}^{-t}\mathrm{d}t,$$

so that

$$y_0 = 1 - \mathrm{e}^{-1},$$
$$y_k = ky_{k-1} - \mathrm{e}^{-1}, \quad k = 1, 2, \ldots \ .$$

(b)

PROGRAM

```
%MAI_17B
%
f0='%8.0f %19.15f\n';
disp('       k        fl(y_k)')
em1=exp(-1);
k=0; y=1-em1;
fprintf(f0,k,y)
for k=1:20
  y=k*y-em1;
  fprintf(f0,k,y)
end
```

OUTPUT

```
>> MAI_17B
   k        fl(y_k)
   0   0.632120558828558
   1   0.264241117657115
   2   0.160602794142788
   3   0.113928941256923
   4   0.087836323856248
   5   0.071302178109799
   6   0.059933627487352
   7   0.051655951240024
```

```
   8   0.045368168748749
   9   0.040434077567295
  10   0.036461334501509
  11   0.033195238345154
  12   0.030463418970410
  13   0.028145005443888
  14   0.026150635042994
  15   0.024380084473469
  16   0.022201910404061
  17   0.009553035697594
  18  -0.195924798614756
  19  -4.090450614851804
  20 -82.176891738207516
>>
```

From the way y_n was defined, it is clear that y_n decreases monotonically with n and tends to zero as $n \to \infty$. The computed y_n are seen to also decrease monotonically, but they tend to $-\infty$ instead of zero! The recursion of part (a) gives a clue as to what is going on: The first term on the right-hand side of the recursion tends to become bigger and bigger (owing to the multiplication by k), yet the right-hand side as a whole should get smaller and smaller. This means that the subtraction involved will be subject to increasing cancellation errors (more and more digits of the constant e^{-1} must be eliminated), which eventually cause the results to become negative. Once this has happened (for $n = 18$ in our `OUTPUT`), the computed y_n become rapidly more negative and tend to $-\infty$.

The true underlying cause is ill-conditioning: The function $f : \ y_0 \mapsto y_n$ is ill-conditioned at $y_0 = 1 - \mathrm{e}^{-1}$. In fact, we have $y_n = n!y_0 + \text{const}$, where the constant depends on n but not on y_0. There follows

$$(\operatorname{cond} y_n)(y_0) = \frac{y_0 n!}{y_n} > n! \quad (\text{since } y_n < y_0).$$

From this, and the `OUTPUT` of part (c) (column headed $Y(n, 5)$), one gets the following table.

n	cond $y_n =$	cond $y_n >$	n	cond $y_n =$	cond $y_n >$
1	2.4	1.0	11	7.6×10^8	4.0×10^7
2	7.9	2.0	12	9.9×10^9	4.8×10^8
3	3.3×10^1	6.0	13	1.4×10^{11}	6.2×10^9
4	1.7×10^2	2.4×10^1	14	2.1×10^{12}	8.7×10^{10}
5	1.1×10^3	1.2×10^2	15	3.4×10^{13}	1.3×10^{12}

6	7.6×10^3	7.2×10^2	16	5.8×10^{14}	2.1×10^{13}
7	6.2×10^4	5.0×10^3	17	1.0×10^{16}	3.6×10^{14}
8	5.6×10^5	4.0×10^4	18	2.0×10^{17}	6.4×10^{15}
9	5.7×10^6	3.6×10^5	19	4.0×10^{18}	1.2×10^{17}
10	6.3×10^7	3.6×10^6	20	8.4×10^{19}	2.4×10^{18}

As can be seen, for $n = 10$, seven significant digits are "wiped out", solely by a rounding error in y_0, and for $n = 20$ almost twenty. Note also that when $n = 18$ the product $\texttt{eps} \cdot \operatorname{cond} y_n$ for the first time becomes (substantially) larger than 1, which is precisely the instance when the computed y_n turns negative.

(c)

```
 PROGRAM

%MAI_17C
%
f0='%8.0f %12.4e\n';
f1='%8.0f %20.16f\n';
i=0; em1=exp(-1);
for N=22:2:30
  i=i+1;
  y=0;
  for k=N:-1:1
    y=(y+em1)/k;
    if k<=21, Y(k,i)=y; end
  end
end
for i=1:4
  e(i)=max(abs((Y(:,i+1)-Y(:,i))./Y(:,i+1)));
end
disp('            e')
fprintf('%18.4e\n',e)
fprintf('\n')
disp('    n       Y(n,5)')
for k=1:21
  fprintf('%5.0f %16.12f\n',k-1,Y(k,5))
end

 OUTPUT

>> MAI_17C
            e
        1.9653e-03
        3.2653e-06
```

```
        4.6392e-09
        5.7001e-12

    n       Y(n,5)
    0   0.632120558829
    1   0.264241117657
    2   0.160602794143
    3   0.113928941257
    4   0.087836323856
    5   0.071302178110
    6   0.059933627487
    7   0.051655951240
    8   0.045368168750
    9   0.040434077580
   10   0.036461334624
   11   0.033195239694
   12   0.030463435153
   13   0.028145215823
   14   0.026153580349
   15   0.024424264063
   16   0.022908783832
   17   0.021569883973
   18   0.020378470348
   19   0.019311495443
   20   0.018350467697
>>
```

The recursion in backward direction, with $N > 20$, reads

$$y_{k-1} = \frac{1}{k}\,(y_k + \mathrm{e}^{-1}), \quad k = N, N-1, \ldots, 21, \ldots, 1.$$

For any n with $0 \le n \le 20$, we have

$$y_n = \frac{1}{(n+1)(n+2)\cdots N}\,y_N + \text{const} = \frac{n!}{N!}\,y_N + \text{const},$$

where the constant depends on n and N, but not on y_N. Hence,

$$(\operatorname{cond} y_n)(y_N) = \frac{y_N n!}{y_n N!} < \frac{n!}{N!} \quad (\text{since } y_n > y_N).$$

This holds for any y_N. In particular, for $y_n^{(N)}$ (defined by $y_N^{(N)} = 0$), we get

$$\left|\frac{y_n^{(N)} - y_n}{y_n}\right| \approx (\operatorname{cond} y_n)(y_N^{(N)}) \left|\frac{y_N^{(N)} - y_N}{y_N}\right| < \frac{n!}{N!} \left|\frac{y_N^{(N)} - y_N}{y_N}\right| = \frac{n!}{N!} \le \frac{20!}{N!}\,.$$

Numerical values of the upper bound on the far right are shown below.

N	$20!/N!$	N	$20!/N!$
21	4.8×10^{-2}	26	6.0×10^{-9}
22	2.2×10^{-3}	27	2.2×10^{-10}
23	9.4×10^{-5}	28	8.0×10^{-12}
24	3.9×10^{-6}	29	2.8×10^{-13}
25	1.6×10^{-7}	30	9.2×10^{-15}

We see that for $N = 29$ we should get 12 correct significant digits for y_n, $0 \leq n \leq 20$, or a little less, owing to rounding errors committed during backward recursion (which, however, are consistently attenuated). This is confirmed by the `OUTPUT` (last entry in the column headed "`e`").

EXERCISES AND MACHINE ASSIGNMENTS TO CHAPTER 2

EXERCISES

1. Suppose you want to approximate the function

$$f(t) = \begin{cases} -1 & \text{if} \quad -1 \le t < 0, \\ 0 & \text{if} \quad t = 0, \\ 1 & \text{if} \quad 0 < t \le 1, \end{cases}$$

by a constant function $\varphi(x) = c$:

(a) on [–1,1] in the continuous L_1 norm,

(b) on $\{t_1, t_2, \ldots, t_N\}$ in the discrete L_1 norm,

(c) on [–1,1] in the continuous L_2 norm,

(d) on $\{t_1, t_2, \ldots, t_N\}$ in the discrete L_2 norm,

(e) on [–1,1] in the ∞-norm,

(f) on $\{t_1, t_2, \ldots, t_N\}$ in the discrete ∞-norm.

The weighting in all norms is uniform (i.e., $w(t) \equiv 1$, $w_i = 1$) and $t_i = -1 + \frac{2(i-1)}{N-1}$, $i = 1, 2, \ldots, N$. Determine the best constant c (or constants c, if there is nonuniqueness) and the minimum error.

2. Consider the data

$$f(t_i) = 1, \quad i = 1, 2, \ldots, N-1; \qquad f(t_N) = y \gg 1.$$

(a) Determine the discrete L_∞ approximant to f by means of a constant c (polynomial of degree zero).

(b) Do the same for discrete (equally weighted) least square approximation.

(c) Compare and discuss the results, especially as $N \to \infty$.

3. Let $x_0, x_1, \ldots, x_n$ be pairwise distinct points in $[a, b]$, $-\infty < a < b < \infty$, and $f \in C^1[a, b]$. Show that, given any $\varepsilon > 0$, there exists a polynomial p such that $\|f - p\|_\infty < \varepsilon$ and, at the same time, $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$. Here $\|u\|_\infty = \max_{a \le x \le b} |u(x)|$. {*Hint*: write $p = p_n(f; \cdot) + \omega_n q$, where $p_n(f; \cdot)$ is the interpolation polynomial of degree n (cf. Sect. 2.2.1, (2.51)), $\omega_n(x) = \prod_{i=0}^{n}(x - x_i)$, $q \in \mathbb{P}$, and apply Weierstrass's approximation theorem.}

4. Consider the function $f(t) = t^\alpha$ on $0 \le t \le 1$, where $\alpha > 0$. Suppose we want to approximate f best in the L_p norm by a constant c, $0 < c < 1$, that is, minimize the L_p error

$$E_p(c) = \|t^\alpha - c\|_p = \left(\int_0^1 |t^\alpha - c|^p \mathrm{d}t \right)^{1/p}$$

as a function of c. Find the optimal $c = c_p$ for $p = \infty$, $p = 2$, and $p = 1$, and determine $E_p(c_p)$ for each of these p-values.

5. Taylor expansion yields the simple approximation $\mathrm{e}^x \approx 1 + x$, $0 \le x \le 1$. Suppose you want to improve this by seeking an approximation of the form $\mathrm{e}^x \approx 1 + cx$, $0 \le x \le 1$, for some suitable c.

   (a) How must c be chosen if the approximation is to be optimal in the (continuous, equally weighted) least squares sense?

   (b) Sketch the error curves $e_1(x) := \mathrm{e}^x - (1 + x)$ and $e_2(x) := \mathrm{e}^x - (1 + cx)$ with c as obtained in (a) and determine $\max_{0\le x\le 1} |e_1(x)|$ and $\max_{0\le x\le 1} |e_2(x)|$.

   (c) Solve the analogous problem with three instead of two terms in the modified Taylor expansion: $\mathrm{e}^x \approx 1 + c_1 x + c_2 x^2$, and provide error curves for $e_1(x) = \mathrm{e}^x - 1 - x - \frac{1}{2}x^2$ and $e_2(x) = \mathrm{e}^x - 1 - c_1 x - c_2 x^2$.

6. Prove Schwarz's inequality

$$|(u, v)| \le \|u\| \cdot \|v\|$$

   for the inner product (2.10). {*Hint*: use the nonnegativity of $\|u + tv\|^2$, $t \in \mathbb{R}$.}

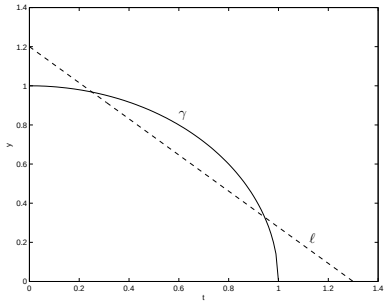
7. Discuss uniqueness and nonuniqueness of the least squares approximant to a function f in the case of a discrete set $T = \{t_1, t_2\}$ (i.e., $N = 2$) and $\Phi_n = \mathbb{P}_{n-1}$ (polynomials of degree $\le n - 1$). In case of nonuniqueness, determine *all* solutions.

8. Determine the least squares approximation

$$\varphi(t) = \frac{c_1}{1 + t} + \frac{c_2}{(1 + t)^2}\ , \quad 0 \le t \le 1,$$

   to the exponential function $f(t) = \mathrm{e}^{-t}$, assuming $\mathrm{d}\lambda(t) = \mathrm{d}t$ on [0,1]. Determine the condition number $\mathrm{cond}_\infty \boldsymbol{A} = \|\boldsymbol{A}\|_\infty \|\boldsymbol{A}^{-1}\|_\infty$ of the coefficient matrix $\boldsymbol{A}$ of the normal equations. Calculate the error $f(t) - \varphi(t)$ at $t = 0$, $t = 1/2$, and $t = 1$. {*Point of information*: the integral $\int_1^\infty t^{-m}\mathrm{e}^{-xt}\mathrm{d}t = E_m(x)$ is known as the "mth exponential integral"; cf. Abramowitz and Stegun [1964, (5.1.4)] or Olver et al. [2010, (8.19.3)].}

9. Approximate the circular quarter arc γ given by the equation $y(t) = \sqrt{1 - t^2}$, $0 \le t \le 1$ (see figure) by a straight line ℓ in the least squares sense, using either the weight function $w(t) = (1 - t^2)^{-1/2}$, $0 \le t \le 1$, or $w(t) = 1$, $0 \le t \le 1$. Where does ℓ intersect the coordinate axes in these two cases? {*Points of information*: $\int_0^{\pi/2} \cos^2\theta \mathrm{d}\theta = \frac{\pi}{4}$, $\int_0^{\pi/2} \cos^3\theta \mathrm{d}\theta = \frac{2}{3}$.}



10. (a) Let the class Φ_n of approximating functions have the following properties. Each $\varphi \in \Phi_n$ is defined on an interval $[a, b]$ symmetric with respect to the origin (i.e., $a = -b$), and $\varphi(t) \in \Phi_n$ implies $\varphi(-t) \in \Phi_n$. Let $\mathrm{d}\lambda(t) = \omega(t)\mathrm{d}t$, with $\omega(t)$ an even function on $[a, b]$ (i.e., $\omega(-t) = \omega(t)$). Show: if f is an even function on $[a, b]$, then so is its least squares approximant, $\hat{\varphi}_n$, on $[a, b]$ from Φ_n.

    (b) Consider the "hat function" $f(t) = \begin{cases} 1-t & \text{if} \quad 0 \le t \le 1, \\ 1+t & \text{if} \quad -1 \le t \le 0. \end{cases}$

    Determine its least squares approximation on [–1,1] by a polynomial of degree ≤ 2. (Use $\mathrm{d}\lambda(t) = \mathrm{d}t$.) Simplify your calculation by using part (a). Determine where the error vanishes.

11. Suppose you want to approximate the step function

$$f(t) = \begin{cases} 1 & \text{if } 0 \le t \le 1, \\ 0 & \text{if } t > 1 \end{cases}$$

on the positive line $\mathbb{R}_+$ by a linear combination of exponentials $\pi_j(t) = \mathrm{e}^{-jt}$, $j = 1, 2, \ldots, n$, in the (continuous, equally weighted) least squares sense.

    (a) Derive the normal equations. How is the matrix related to the Hilbert matrix?

    (b) Use Matlab to solve the normal equations for $n = 1, 2, \ldots, 8$. Print n, the Euclidean condition number of the matrix (supplied by the Matlab function `cond.m`), along with the solution. Plot the approximations vs. the exact function for $1 \le n \le 4$.

12. Let $\pi_j(t) = (t - a_j)^{-1}$, $j = 1, 2, \ldots, n$, where a_j are distinct real numbers with $|a_j| > 1$, $j = 1, 2, \ldots, n$. For $\mathrm{d}\lambda(t) = \mathrm{d}t$ on $-1 \le t \le 1$ and $\mathrm{d}\lambda(t) = 0$, $t \notin [-1, 1]$, determine the matrix of the normal equations for the least squares problem $\int_{\mathbb{R}} (f - \varphi)^2 \mathrm{d}\lambda(t) = \min$, $\varphi = \sum_{j=1}^{n} c_j \pi_j$. Can the sytem $\{\pi_j\}_{j=1}^{n}$, $n > 1$, be an orthogonal system for suitable choices of the constants a_j? Explain.

13. Given an integer $n \geq 1$, consider the subdivision Δ_n of the interval $[0,1]$ into n equal subintervals of length $1/n$. Let $\pi_j(t)$, $j = 0, 1, \ldots, n$, be the function having the value 1 at $t = j/n$, decreasing on either side linearly to zero at the neighboring subdivision points (if any), and being zero elsewhere.

    (a) Draw a picture of these functions. Describe in words the meaning of a linear combination $\pi(t) = \sum_{j=0}^{n} c_j \pi_j(t)$.

    (b) Determine $\pi_j(k/n)$ for j, $k = 0, 1, \ldots, n$.

    (c) Show that the system $\{\pi_j(t)\}_{j=0}^{n}$ is linearly independent on the interval $0 \leq t \leq 1$. Is it also linearly independent on the set of subdivision points 0, $\frac{1}{n}$, $\frac{2}{n}, \ldots, \frac{n-1}{n}$, 1 of Δ_n? Explain.

    (d) Compute the matrix of the normal equations for $\{\pi_j\}$, assuming $\mathrm{d}\lambda(t) = \mathrm{d}t$ on [0,1]. That is, compute the $(n+1) \times (n+1)$ matrix $\boldsymbol{A} = [a_{ij}]$, where $a_{ij} = \int_0^1 \pi_i(t)\pi_j(t)\mathrm{d}t$.

14. Even though the function $f(t) = \ln(1/t)$ becomes infinite as $t \to 0$, it can be approximated on [0,1] arbitrarily well by polynomials of sufficiently high degree in the (continuous, equally weighted) least squares sense. Show this by proving

$$e_{n,2} := \min_{p \in \mathbb{P}_n} \|f - p\|_2 \;=\; \frac{1}{n+1}\,.$$

    {*Hint*: use the following known facts about the "shifted" Legendre polynomial $\pi_j(t)$ of degree j (orthogonal on [0,1] with respect to the weight function $w \equiv 1$ and normalized to satisfy $\pi_j(1) = 1$):

$$\int_0^1 \pi_j^2(t)\mathrm{d}t = \frac{1}{2j+1}, \quad j \geq 0; \qquad \int_0^1 \pi_j(t)\ln(1/t)\mathrm{d}t = \begin{cases} 1 & \text{if } \; j = 0, \\ \frac{(-1)^j}{j(j+1)} & \text{if } \; j > 0. \end{cases}$$

    The first relation is well known from the theory of orthogonal polynomials (see, e.g., Sect. 1.5.1, p. 27 of Gautschi [2004]); the second is due to Blue [1979].}

15. Let $\mathrm{d}\lambda$ be a continuous (positive) measure on $[a,b]$ and $n \geq 1$ a given integer. Assume f continuous on $[a,b]$ and not a polynomial of degree $\leq n-1$. Let $\hat{p}_{n-1} \in \mathbb{P}_{n-1}$ be the least squares approximant to f on $[a,b]$ from polynomials of degree $\leq n-1$:

$$\int_a^b [\hat{p}_{n-1}(t) - f(t)]^2 \mathrm{d}\lambda(t) \leq \int_a^b [p(t) - f(t)]^2 \mathrm{d}\lambda(t), \qquad \text{all } p \in \mathbb{P}_{n-1}.$$

    Prove: the error $e_n(t) = \hat{p}_{n-1}(t) - f(t)$ changes sign at least n times in $[a,b]$. {*Hint*: assume the contrary and develop a contradiction.}

16. Let f be a given function on [0,1] satisfying $f(0) = 0,\ f(1) = 1$.

(a) Reduce the problem of approximating f on [0,1] in the (continuous, equally weighted) least squares sense by a quadratic polynomial p satisfying $p(0) = 0,\ p(1) = 1$ to an unconstrained least squares problem (for a different function).

(b) Apply the result of (a) to $f(t) = t^r,\ r > 2$. Plot the approximation against the exact function for $r = 3$.

17. Suppose you want to approximate $f(t)$ on $[a, b]$ by a function of the form $r(t) = \pi(t)/q(t)$ in the least squares sense with weight function w, where $\pi \in \mathbb{P}_n$ and q is a *given* function (e.g., a polynomial) such that $q(t) > 0$ on $[a, b]$. Formulate this problem as an ordinary polynomial least squares problem for an appropriate new function $\overline{f}$ and new weight function $\overline{w}$.

18. The Bernstein polynomials of degree n are defined by

$$B_j^n(t) = \binom{n}{j} t^j (1-t)^{n-j}, \quad j = 0, 1, \ldots, n,$$

and are usually employed on the interval $0 \le t \le 1$.

(a) Show that $B_0^n(0) = 1$, and for $j = 1, 2, \ldots, n$

$$\frac{\mathrm{d}^r}{\mathrm{d}t^r}\ B_j^n(t)\bigg|_{t=0} = 0, \quad r = 0, 1, \ldots, j-1; \quad \frac{\mathrm{d}^j}{\mathrm{d}t^j}\ B_j^n(t)\bigg|_{t=0} \neq 0.$$

(b) What are the analogous properties at $t = 1$, and how are they most easily derived?

(c) Prepare a plot of the fourth-degree polynomials $B_j^4(t),\ j = 0, 1, \ldots, 4$, $0 \le t \le 1$.

(d) Use (a) to show that the system $\{B_j^n(t)\}_{j=0}^n$ is linearly independent on [0,1] and spans the space $\mathbb{P}_n$.

(e) Show that $\sum_{j=0}^n B_j^n(t) \equiv 1$. {*Hint*: use the binomial theorem.}

19. Prove that, if $\{\pi_j\}_{j=1}^n$ is linearly dependent on the support of $\mathrm{d}\lambda$, then the matrix $\boldsymbol{A} = [a_{ij}]$, where $a_{ij} = (\pi_i, \pi_j)_{\mathrm{d}\lambda} = \int_{\mathbb{R}} (\pi_i(t), \pi_j(t) \mathrm{d}\lambda(t)$, is singular.

20. Given the recursion relation $\pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t)$, $k = 0, 1, 2, \ldots$, for the (monic) orthogonal polynomials $\{\pi_j(\,\cdot\,; \mathrm{d}\lambda)\}$, and defining $\beta_0 = \int_{\mathbb{R}} \mathrm{d}\lambda(t)$, show that $\|\pi_k\|^2 = \beta_0 \beta_1 \cdots \beta_k$, $k = 0, 1, 2, \ldots$.

21. (a) Derive the three-term recurrence relation

$$\sqrt{\beta_{k+1}}\,\tilde{\pi}_{k+1}(t) = (t - \alpha_k)\tilde{\pi}_k(t) - \sqrt{\beta_k}\,\tilde{\pi}_{k-1}, \quad k = 0, 1, 2, \ldots,$$
$$\tilde{\pi}_{-1}(t) = 0, \quad \tilde{\pi}_0 = 1/\sqrt{\beta_0}$$

for the orthonormal polynomials $\tilde{\pi}_k = \pi_k / \|\pi_k\|$, $k = 0, 1, 2, \ldots$.

(b) Use the result of (a) to derive the *Christoffel–Darboux* formula

$$\sum_{k=0}^{n} \tilde{\pi}_k(x)\tilde{\pi}_k(t) = \sqrt{\beta_{n+1}}\,\frac{\tilde{\pi}_{n+1}(x)\tilde{\pi}_n(t) - \tilde{\pi}_n(x)\tilde{\pi}_{n+1}(t)}{x-t}\,.$$

22. (a) Let $\pi_n(\,\cdot\,) = \pi_n(\,\cdot\,;\mathrm{d}\lambda)$ be the (monic) orthogonal polynomial of degree n relative to the positive measure $\mathrm{d}\lambda$ on $\mathbb{R}$. Show:

$$\int_{\mathbb{R}} \pi_n^2(t;\mathrm{d}\lambda)\mathrm{d}\lambda(t) \le \int_{\mathbb{R}} p^2(t)d\lambda(t), \quad \text{all } \; p \in \overset{\circ}{P}_n,$$

where $\overset{\circ}{P}_n$ is the class of monic polynomials of degree n. Discuss the case of equality. {*Hint*: represent p in terms of $\pi_j(\,\cdot\,;\mathrm{d}\lambda), \;\; j = 0, 1, \ldots, n.$}

(b) If $\mathrm{d}\lambda(t) = \mathrm{d}\lambda_N(t)$ is a discrete measure with exactly N support points $t_1, t_2, \ldots, t_N$, and $\pi_j(t) = \pi_j(\,\cdot\,;\mathrm{d}\lambda_N),\ j = 0, 1, \ldots, N-1$, are the corresponding (monic) orthogonal polynomials, let $\pi_N(t) = (t-\alpha_{N-1})\pi_{N-1}(t) - \beta_{N-1}\pi_{N-2}(t)$, with $\alpha_{N-1},\ \beta_{N-1}$ defined as in Sect. 2.1.4(2). Show that $\pi_N(t_j) = 0$ for $j = 1, 2, \ldots, N$.

23. Let $\{\pi_j\}_{j=0}^n$ be a system of orthogonal polynomials, not necessarily monic, relative to the (positive) measure $\mathrm{d}\lambda$. For some a_{ij}, define

$$p_i(t) = \sum_{j=0}^{n} a_{ij}\pi_j(t), \quad i = 1, 2, \ldots, n,$$

(a) Derive conditions on the matrix $\boldsymbol{A} = [a_{ij}]$ which ensure that the system $\{p_i\}_{i=0}^n$ is also a system of orthogonal polynomials.

(b) Assuming all π_j monic and $\{p_i\}_{i=0}^n$ an orthogonal system, show that each p_i is monic if and only if $\boldsymbol{A} = \boldsymbol{I}$ is the identity matrix.

(c) Prove the same as in (b), with "monic" replaced by "orthonormal" throughout.

24. Let $(u, v) = \sum_{k=1}^{N} w_k u(t_k)v(t_k)$ be a discrete inner product on the interval [–1,1] with $-1 \le t_1 < t_2 < \cdots < t_N \le 1$, and let $\alpha_k,\ \beta_k$ be the recursion coefficients for the (monic) orthogonal polynomials $\{\pi_k(t)\}_{k=0}^{N-1}$ associated with $(u,\, v)$:

$$\begin{cases} \pi_{k+1}(t) = (t-\alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \\ \qquad\qquad\qquad\qquad k = 0, 1, 2, \ldots, N-2, \\ \pi_0(t) = 1, \;\; \pi_{-1}(t) = 0. \end{cases}$$

Let $x = \frac{b-a}{2}t + \frac{a+b}{2}$ map the interval [–1,1] to $[a, b]$, and the points $t_k \in [-1, 1]$ to $x_k \in [a, b]$. Define $(u, v)^* = \sum_{k=1}^{N} w_k u(x_k)v(x_k)$, and let $\{\pi_k^*(x)\}_{k=0}^{N-1}$ be

the (monic) orthogonal polynomials associated with $(u, v)^*$. Express the recursion coefficients α_k^*, β_k^* for the $\{\pi_k^*\}$ in terms of those for $\{\pi_k\}$. {*Hint*: first show that $\pi_k^*(x) = (\frac{b-a}{2})^k \pi_k(\frac{2}{b-a}(x - \frac{a+b}{2}))$.}

25. Let

$$(\star) \qquad \begin{cases} \pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \\ \qquad\qquad\qquad\qquad k = 0, 1, 2, \ldots, n-1, \\ \pi_0(t) = 1, \quad \pi_{-1}(t) = 0 \end{cases}$$

and consider

$$p_n(t) = \sum_{j=0}^{n} c_j \pi_j(t).$$

Show that p_n can be computed by the following algorithm (*Clenshaw's algorithm*):

$$(\star\star) \qquad \begin{cases} u_n = c_n, \quad u_{n+1} = 0, \\ u_k = (t - \alpha_k)u_{k+1} - \beta_{k+1}u_{k+2} + c_k, \\ \qquad\qquad\qquad\qquad k = n-1, n-2, \ldots, 0, \\ p_n = u_0. \end{cases}$$

{*Hint*: write $(\star)$ in matrix form in terms of the vector $\boldsymbol{\pi}^T = [\pi_0, \pi_1, \ldots, \pi_n]$ and a unit triangular matrix. Do likewise for $(\star\star)$.}

26. Show that the elementary Lagrange interpolation polynomials $\ell_i(x)$ are invariant with respect to any linear transformation of the independent variable.

27. Use Matlab to prepare plots of the Lebesgue function for interpolation, $\lambda_n(x)$, $-1 \le x \le 1$, for $n = 5, 10, 20$, with the interpolation nodes x_i being given by

    (a) $x_i = -1 + \frac{2i}{n}, \quad i = 0, 1, 2, \ldots, n;$

    (b) $x_i = \cos \frac{2i+1}{2n+2}\pi, \quad i = 0, 1, 2, \ldots, n.$

    Compute $\lambda_n(x)$ on a grid obtained by dividing each interval $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$, into 20 equal subintervals. Plot $\log_{10} \lambda_n(x)$ in case (a), and $\lambda_n(x)$ in case (b). Comment on the results.

28. Let $\omega_n(x) = \prod_{k=0}^{n} (x - k)$ and denote by x_n the location of the extremum of ω_n on [0,1], that is, the unique x in [0,1], where $\omega_n'(x) = 0$.

    (a) Prove or disprove that $x_n \to 0$ as $n \to \infty$.

    (b) Investigate the monotonicity of x_n as n increases.

29. Consider equidistant sampling points $x_k = k$ $(k = 0, 1, \ldots, n)$ and $\omega_n(x) = \prod_{k=0}^{n} (x - k)$, $0 \le x \le n$.

    (a) Show that $\omega_n(x) = (-1)^{n+1}\omega_n(n - x)$. What kind of symmetry does this imply?

    (b) Show that $|\,\omega_n(x)| < |\,\omega_n(x + 1)|$ for nonintegral $x > (n - 1)/2$.

    (c) Show that the relative maxima of $|\,\omega_n(x)|$ increase monotonically (from the center of $[0, n]$ outward).

30. Let

$$\lambda_n(x) = \sum_{i=0}^{n} |\ell_i(x)|$$

be the Lebesgue function for polynomial interpolation at the distinct points $x_i \in [a, b]$, $i = 0, 1, \ldots, n$, and $\Lambda_n = \|\lambda_n\|_\infty = \max_{a \le x \le b} |\lambda_n(x)|$ the Lebesgue constant. Let $p_n(f; \cdot)$ be the polynomial of degree $\le n$ interpolating f at the nodes x_i. Show that in the inequality

$$\|p_n(f; \cdot)\|_\infty \le \Lambda_n \|f\|_\infty, \quad f \in C[a, b],$$

equality can be attained for some $f = \varphi \in C[a, b]$. {*Hint*: let $\|\lambda_n\|_\infty = \lambda_n(x_\infty)$; take $\varphi \in C[a, b]$ piecewise linear and such that $\varphi(x_i) = \operatorname{sgn} \ell_i(x_\infty)$, $i = 0, 1, \ldots, n$.}

31. (a) Let $x_0, x_1, \ldots, x_n$ be $n + 1$ distinct points in $[a, b]$ and $f_i = f(x_i)$, $i = 0, 1, \ldots, n$, for some function f. Let $f_i^* = f_i + \varepsilon_i$, where $|\varepsilon_i| \le \varepsilon$. Use the Lagrange interpolation formula to show that $|p_n(f^*; x) - p_n(f; x)| \le \varepsilon\lambda_n(x)$, $a \le x \le b$, where $\lambda_n(x)$ is the Lebesgue function (cf. Ex. 30).

    (b) Show: $\lambda_n(x_j) = 1$ for $j = 0, 1, \ldots, n$.

    (c) For quadratic interpolation at three equally spaced points, show that $\lambda_2(x) \le 1.25$ for any x between the three points.

    (d) Obtain $\lambda_2(x)$ for $x_0 = 0$, $x_1 = 1$, $x_2 = p$, where $p \gg 1$, and determine $\max_{1 \le x \le p} \lambda_2(x)$. How fast does this maximum grow with p? {*Hint*: to simplify the algebra, note from (b) that $\lambda_2(x)$ on $1 \le x \le p$ must be of the form $\lambda_2(x) = 1 + c(x - 1)(p - x)$ for some constant c.}

32. In a table of the Bessel function $J_0(x) = \frac{1}{\pi}\int_0^\pi \cos(x \sin\theta)\mathrm{d}\theta$, where x is incremented in steps of size h, how small must h be chosen if the table is to be "linearly interpolable" with error less that 10^{-6} in absolute value? {*Point of information*: $\int_0^{\pi/2} \sin^2\theta \mathrm{d}\theta = \frac{\pi}{4}$.}

33. Suppose you have a table of the logarithm function $\ln x$ for positive integer values of x, and you compute $\ln 11.1$ by quadratic interpolation at $x_0 = 10$, $x_1 = 11$, $x_2 = 12$. Estimate the relative error incurred.

34. The "Airy function" $y(x) = \mathrm{Ai}(x)$ is a solution of the differential equation $y'' = xy$ satisfying appropriate initial conditions. It is known that $\mathrm{Ai}(x)$ on $[0, \infty)$ is monotonically decreasing to zero and $\mathrm{Ai}'(x)$ monotonically increasing to zero. Suppose you have a table of Ai and Ai′ (with tabular step h) and you want to interpolate

    (a) linearly between x_0 and x_1,

    (b) quadratically between x_0, x_1, and x_2,

    where $x_0,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h$ are (positive) tabular arguments. Determine close upper bounds for the respective errors in terms of quantities $y_k = y(x_k),\ y'_k = y'(x_k),\ k = 0, 1, 2$, contained in the table.

35. The error in linear interpolation of f at x_0, x_1 is known to be

$$f(x) - p_1(f; x) = (x - x_0)(x - x_1)\ \frac{f''(\xi(x))}{2}\ , \quad x_0 < x < x_1,$$

    if $f \in C^2[x_0, x_1]$. Determine $\xi(x)$ explicitly in the case $f(x) = \frac{1}{x}$, $x_0 = 1$, $x_1 = 2$, and find $\max_{1\le x\le 2} \xi(x)$ and $\min_{1\le x\le 2} \xi(x)$.

36. (a) Let $p_n(f; x)$ be the interpolation polynomial of degree $\le n$ interpolating $f(x) = \mathrm{e}^x$ at the points $x_i = i/n,\ i = 0, 1, 2, \ldots, n$. Derive an upper bound for

$$\max_{0\le x\le 1} |\mathrm{e}^x - p_n(f; x)|,$$

    and determine the smallest n guaranteeing an error less than 10^{-6} on [0, 1]. {*Hint*: first show that for any integer i with $0 \le i \le n$ one has $\max_{0\le x\le 1} |(x - \frac{i}{n})(x - \frac{n-i}{n})| \le \frac{1}{4}$.}

    (b) Solve the analogous problem for the nth-degree Taylor polynomial $t_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$, and compare the result with the one in (a).

37. Let $x_0 < x_1 < x_2 < \cdots < x_n$ and $H = \max_{0\le i\le n-1}(x_{i+1} - x_i)$. Defining $\omega_n(x) = \prod_{i=0}^{n}(x - x_i)$, find an upper bound for $\|\omega_n\|_\infty = \max_{x_0\le x\le x_n} |\,\omega_n(x)|$ in terms of H and n. {*Hint*: assume $x_j \le x \le x_{j+1}$ for some $0 \le j < n$ and estimate $(x - x_j)(x - x_{j+1})$ and $\prod_{\substack{i\ne j\\ i\ne j+1}} (x - x_i)$ separately.}

38. Show that the power x^n on the interval $-1 \le x \le 1$ can be uniformly approximated by a linear combination of powers $1, x, x^2, \ldots, x^{n-1}$ with error $\le 2^{-(n-1)}$. In this sense, the powers of x become "less and less linearly independent" on [–1,1] with growing exponent n.

39. Determine

$$\min \max_{a\le x\le b} |a_0x^n + a_1x^{n-1} + \cdots + a_n|, \quad n \ge 1,$$

    where the minimum is taken over all real $a_0, a_1, \ldots, a_n$ with $a_0 \ne 0$. {*Hint*: use Theorem 2.2.1.}

40. Let $a > 1$ and $\mathbb{P}_n^a = \{p \in \mathbb{P}_n : p(a) = 1\}$. Define $\hat{p}_n \in \mathbb{P}_n^a$ by $\hat{p}_n(x) = \frac{T_n(x)}{T_n(a)}$, where T_n is the Chebyshev polynomial of degree n, and let $\|\cdot\|_\infty$ denote the maximum norm on the interval $[-1, 1]$. Prove:

$$\|\hat{p}_n\|_\infty \leq \|p\|_\infty \quad \text{for all } \; p \in \mathbb{P}_n^a.$$

{*Hint*: imitate the proof of Theorem 2.2.1.}

41. Let

$$f(x) = \int_5^\infty \frac{e^{-t}}{t - x} \mathrm{d}t, \qquad -1 \leq x \leq 1,$$

and let $p_{n-1}(f; \cdot)$ be the polynomial of degree $\leq n - 1$ interpolating f at the n Chebyshev points $x_\nu = \cos(\frac{2\nu-1}{2n}\pi)$, $\nu = 1, 2, \ldots, n$. Derive an upper bound for $\max_{-1\leq x\leq 1} |f(x) - p_{n-1}(f, x)|$.

42. Let f be a positive function defined on $[a, b]$ and assume

$$\min_{a\leq x\leq b} |f(x)| = m_0, \quad \max_{a\leq x\leq b} |f^{(k)}(x)| = M_k, \quad k = 0, 1, 2, \ldots \; .$$

(a) Denote by $p_{n-1}(f; \cdot)$ the polynomial of degree $\leq n - 1$ interpolating f at the n Chebyshev points (relative to the interval $[a, b]$). Estimate the maximum relative error $r_n = \max_{a\leq x\leq b} |(f(x) - p_{n-1}(f; x))/f(x)|$.

(b) Apply the result of (a) to $f(x) = \ln x$ on $I_r = \{\mathrm{e}^r \leq x \leq \mathrm{e}^{r+1}\}$, $r \geq 1$ an integer. In particular, show that $r_n \leq \alpha(r, n)c^n$, where $0 < c < 1$ and α is slowly varying. Exhibit c.

(c) (This relates to the function $f(x) = \ln x$ of part (b).) How does one compute $f(\overline{x})$, $\overline{x} \in I_s$, from $f(x)$, $x \in I_r$?

43. (a) For quadratic interpolation on equally spaced points x_0, $x_1 = x_0 + h$, $x_2 = x_0 + 2h$, derive an upper bound for $\|f - p_2(f; \cdot)\|_\infty$ involving $\|f'''\|_\infty$ and h. (Here $\|u\|_\infty = \max_{x_0\leq x\leq x_2} |u(x)|$.)

(b) Compare the bound obtained in (a) with the analogous one for interpolation at the three Chebyshev points on $[x_0, x_2]$.

44. (a) Suppose the function $f(x) = \ln(2 + x)$, $-1 \leq x \leq 1$, is interpolated by a polynomial p_n of degree $\leq n$ at the Chebyshev points $x_k = \cos\left(\frac{2k+1}{2n+2}\pi\right)$, $k = 0, 1, \ldots, n$. Derive a bound for the maximum error $\|f - p_n\|_\infty = \max_{-1\leq x\leq 1} |f(x) - p_n(x)|$.

(b) Compare the result of (a) with bounds for $\|f - t_n\|_\infty$, where $t_n(x)$ is the nth-degree Taylor polynomial of f and where either Lagrange's form of the remainder is used or the full Taylor expansion of f.

45. Consider $f(t) = \cos^{-1} t$, $-1 \leq t \leq 1$. Obtain the least squares approximation $\hat{\varphi}_n \in \mathbb{P}_n$ of f relative to the weight function $w(t) = (1 - t)^{-\frac{1}{2}}$; that is, find

the solution $\varphi = \hat{\varphi}_n$ of

$$\text{minimize} \quad \left\{ \int_{-1}^{1} [f(t) - \varphi(t)]^2 \frac{\mathrm{d}t}{\sqrt{1-t^2}} : \ \varphi \in \mathbb{P}_n \right\}.$$

Express $\hat{\varphi_n}$ in terms of Chebyshev polynomials $\pi_j(t) = T_j(t)$.

46. Compute $T_n'(0)$, where T_n is the Chebyshev polynomial of degree n.

47. Prove that the system of Chebyshev polynomials $\{T_k : 0 \le k < n\}$ is orthogonal with respect to the discrete inner product $(u, v) = \sum_{\nu=1}^{n} u(x_\nu)v(x_\nu)$, where x_ν are the Chebyshev points $x_\nu = \cos \frac{2\nu-1}{2n}\pi$.

48. Let $T_k(x)$ denote the Chebyshev polynomial of degree k. Clearly, $T_n(T_m(x))$ is a polynomial of degree $n \cdot m$. Identify it.

49. Let T_n denote the Chebyshev polynomial of degree $n \ge 2$. The equation

$$x = T_n(x)$$

is an algebraic equation of degree n and hence has exactly n roots. Identify them.

50. For any x with $0 \le x \le 1$ show that $T_n(2x-1) = T_{2n}(\sqrt{x})$.

51. Let $f(x)$ be defined for all $x \in \mathbb{R}$ and infinitely often differentiable on $\mathbb{R}$. Assume further that

$$|f^{(m)}(x)| \le 1, \quad \text{all} \ \ x \in \mathbb{R}, \ m = 1, 2, 3, \dots \ .$$

Let $h > 0$ and p_{2n-1} be the polynomial of degree $< 2n$ interpolating f at the $2n$ points $x = kh$, $k = \pm 1, \pm 2, \dots, \pm n$. For what values of h is it true that

$$\lim_{n\to\infty} p_{2n-1}(0) = f(0)\,?$$

(Note that $x = 0$ is *not* an interpolation node.) Explain why the convergence theory discussed in Sect. 2.2.3 does not apply here. {*Point of information*: $n! \sim \sqrt{2\pi n}(n/\mathrm{e})^n$ as $n \to \infty$ (Stirling's formula).}

52. (a) Let $x_i^C = \cos\left(\frac{2i+1}{2n+2}\pi\right)$, $i = 0, 1, \dots, n$, be Chebyshev points on $[-1, 1]$. Obtain the analogous Chebyshev points t_i^C on $[a, b]$ (where $a < b$) and find an upper bound of $\prod_{i=0}^{n}(t - t_i^C)$ for $a \le t \le b$.

    (b) Consider $f(t) = \ln t$ on $[a, b]$, $0 < a < b$, and let $p_n(t) = p_n(f; t_0^{(n)}, t_1^{(n)}, \dots, t_n^{(n)}; t)$. Given $a > 0$, how large can b be chosen such that $\lim_{n\to\infty} p_n(t) = f(t)$ for arbitrary nodes $t_i^{(n)} \in [a, b]$ and arbitrary $t \in [a, b]$?

(c) Repeat (b), but with $t_i^{(n)} = t_i^C$ (see (a)).

53. Let $\mathbb{P}_m^+$ be the set of all polynomials of degree $\le m$ that are nonnegative on the real line,

$$\mathbb{P}_m^+ = \{p : \quad p \in \mathbb{P}_m,\ p(x) \ge 0 \text{ for all } x \in \mathbb{R}\}.$$

Consider the following interpolation problem: find $p \in \mathbb{P}_m^+$ such that $p(x_i) = f_i,\ i = 0, 1, \ldots, n$, where $f_i \ge 0$ and x_i are distinct points on $\mathbb{R}$.

(a) Show that, if $m = 2n$, the problem admits a solution for arbitrary $f_i \ge 0$.

(b) Prove: if a solution is to exist for arbitrary $f_i \ge 0$, then, necessarily, $m \ge 2n$. {*Hint*: consider $f_0 = 1,\ f_1 = f_2 = \cdots = f_n = 0$.}

54. Defining forward differences by $\Delta f(x) = f(x+h) - f(x),\ \Delta^2 f(x) = \Delta(\Delta f(x)) = f(x+2h) - 2f(x+h) + f(x)$, and so on, show that

$$\Delta^k f(x) = k! h^k [x_0, x_1, \ldots, x_k] f,$$

where $x_j = x + jh,\ j = 0, 1, 2, \ldots$. Prove an analogous formula for backward differences.

55. Let $f(x) = x^7$. Compute the fifth divided difference [0,1,1,1,2,2]f of f. It is known that this divided difference is expressible in terms of the fifth derivative of f evaluated at some ξ, $0 < \xi < 2$ (cf. (2.117)). Determine ξ.

56. In this problem $f(x) = \mathrm{e}^x$ throughout.

(a) Prove: for any real number t, one has

$$[t, t+1, \ldots, t+n] f = \frac{(\mathrm{e}-1)^n}{n!}\, \mathrm{e}^t.$$

{*Hint*: use induction on n.}

(b) From (2.117) we know that

$$[0, 1, \ldots, n] f = \frac{f^{(n)}(\xi)}{n!}, \quad 0 < \xi < n.$$

Use the result in (a) to determine ξ. Is ξ located to the left or to the right of the midpoint $n/2$?

57. (Euler, 1734) Let $x_k = 10^k$, $k = 0, 1, 2, 3, \ldots$, and $f(x) = \log_{10} x$.

(a) Show that

$$[x_0, x_1, \ldots, x_n] f = \frac{(-1)^{n-1}}{10^{n(n-1)/2}(10^n - 1)}, \quad n = 1, 2, 3, \ldots.$$

{*Hint*: prove more generally

$$[x_r, x_{r+1}, \ldots, x_{r+n}]f = \frac{(-1)^{n-1}}{10^{rn+n(n-1)/2}(10^n - 1)}, \quad r \geq 0,$$

by induction on n.}

(b) Use Newton's interpolation formula to determine $p_n(x) = p_n(f; x_0, x_1, \ldots, x_n; x)$. Show that $\lim_{n\to\infty} p_n(x)$ exists for $1 \leq x < 10$. Is the limit equal to $\log_{10} x$? (Check, e.g., for $x = 9$.)

58. Show that

$$\frac{\partial}{\partial x_0}[x_0, x_1, \ldots, x_n]f = [x_0, x_0, x_1, \ldots, x_n]f,$$

assuming f is differentiable at x_0. What about the partial derivative with respect to one of the other variables?

59. (a) For $n+1$ distinct nodes x_ν, show that

$$[x_0, x_1, \ldots, x_n]f = \sum_{\nu=0}^{n} \frac{f(x_\nu)}{\prod_{\mu\neq\nu}(x_\nu - x_\mu)}.$$

(b) Show that

$$[x_0, x_1, \ldots, x_n](fg_j) = [x_0, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]f,$$

where $g_j(x) = x - x_j$.

60. (Mikeladze, 1941) Assuming $x_0, x_1, \ldots, x_n$ mutually distinct, show that

$$[\underbrace{x_0, x_0, \ldots, x_0}_{m \text{ times}}, x_1, x_2, \ldots, x_n]f$$

$$= \frac{[\overbrace{x_0, \ldots, x_0}^{m \text{ times}}]f}{\prod_{\mu=1}^{n}(x_0 - x_\mu)} + \sum_{\nu=1}^{n} \frac{[\overbrace{x_0, \ldots, x_0}^{(m-1) \text{ times}}, x_\nu]f}{\prod_{\substack{\mu=0 \\ \mu\neq\nu}}^{n}(x_\nu - x_\mu)}.$$

{*Hint*: use induction on m.}

61. Determine the number of additions and the number of multiplications/divisions required

(a) to compute all divided differences for $n+1$ data points,

(b) to compute all auxiliary quantities $\lambda_i^{(n)}$ in (2.103), and

(c) to compute $p_n(f;\,\cdot\,)$ (efficiently) from Newton's formula (2.111), once the divided differences are available. Compare with the analogous count for the barycentric formula (2.105), assuming all auxiliary quantities available. Overall, which of the two formulae can be computed more economically?

62. Consider the data $f(0) = 5$, $f(1) = 3$, $f(3) = 5$, $f(4) = 12$.

(a) Obtain the appropriate interpolation polynomial $p_3(f;x)$ in Newton's form.

(b) The data suggest that f has a minimum between $x = 1$ and $x = 3$. Find an approximate value for the location $x_{\min}$ of the minimum.

63. Let $f(x) = (1+a)^x$, $|a| < 1$. Show that $p_n(f;0,1,\ldots,n;x)$ is the truncation of the binomial series for f to $n+1$ terms. {*Hint*: use Newton's form of the interpolation polynomial.}

64. Suppose f is a function on [0,3] for which one knows that

$$f(0) = 1, \quad f(1) = 2, \quad f'(1) = -1, \quad f(3) = f'(3) = 0.$$

(a) Estimate $f(2)$, using Hermite interpolation.

(b) Estimate the maximum possible error of the answer given in (a) if one knows, in addition, that $f \in C^5[0,3]$ and $|f^{(5)}(x)| \le M$ on [0,3]. Express the answer in terms of M.

65. (a) Use Hermite interpolation to find a polynomial of lowest degree satisfying $p(-1) = p'(-1) = 0$, $p(0) = 1$, $p(1) = p'(1) = 0$. Simplify your expression for p as much as possible.

(b) Suppose the polynomial p of (a) is used to approximate the function $f(x) = [\cos(\pi x/2)]^2$ on $-1 \le x \le 1$.

(b1) Express the error $e(x) = f(x) - p(x)$ (for some *fixed* x in $[-1,1]$) in terms of an appropriate derivative of f.

(b2) Find an upper bound for $|e(x)|$ (still for a *fixed* $x \in [-1,1]$).

(b3) Estimate $\max_{-1\le x\le 1} |e(x)|$.

66. Consider the problem of finding a polynomial $p \in \mathbb{P}_n$ such that

$$p(x_0) = f_0, \quad p'(x_i) = f_i', \quad i = 1,2,\ldots,n,$$

where x_i, $i = 1,2,\ldots,n$, are distinct nodes. (It is not excluded that $x_1 = x_0$.) This is neither a Lagrange nor a Hermite interpolation problem (why not?). Nevertheless, show that the problem has a unique solution and describe how it can be obtained.

67. Let

$$f(t) = \begin{cases} 0 & \text{if} \quad 0 \le t \le \frac{1}{2}, \\ 1 & \text{if} \quad \frac{1}{2} \le t \le 1. \end{cases}$$

(a) Find the linear least squares approximant $\hat{p}_1$ to f on $[0,1]$, that is, the polynomial $p_1 \in \mathbb{P}_1$ for which

$$\int_0^1 [p_1(t) - f(t)]^2 \mathrm{d}t = \min.$$

Use the normal equations with $\pi_0(t) = 1,\ \pi_1(t) = t$.

(b) Can you do better with continuous piecewise linear functions (relative to the partition $[0,1] = [0,\frac{1}{2}] \cup [\frac{1}{2},1]$)? Use the normal equations for the B-spline basis B_0, B_1, B_2 (cf. Sect. 2.3.2 and Ex. 13).

68. Show that $\mathbb{S}_m^m(\Delta) = \mathbb{P}_m$.

69. Let Δ be the subdivision

$$\Delta = [0,1] \cup [1,2] \cup [2,3]$$

of the interval [0,3]. Define the function s by

$$s(x) = \begin{cases} 2 - x(3 - 3x + x^2) & \text{if} \quad 0 \le x \le 1, \\ 1 & \text{if} \quad 1 \le x \le 2, \\ \frac{1}{4}\, x^2(3 - x) & \text{if} \quad 2 \le x \le 3. \end{cases}$$

To which class $\mathbb{S}_m^k(\Delta)$ does s belong?

70. In

$$s(x) = \begin{cases} p(x) & \text{if} \ \ 0 \le x \le 1, \\ (2 - x)^3 & \text{if} \ \ 1 \le x \le 2 \end{cases}$$

determine $p \in \mathbb{P}_3$ such that $s(0) = 0$ and s is a cubic spline in $\mathbb{S}_3^2(\Delta)$ on the subdivision $\Delta = [0,1] \cup [1,2]$ of the interval [0,2]. Do you get a natural spline?

71. Let Δ: $a = x_1 < x_2 < x_3 < \cdots < x_n = b$ be a subdivision of $[a,b]$ into $n-1$ subintervals. What is the dimension of the space $\mathbb{S}_m^k = \{s \in C^k[a,b]: s|_{[x_i,x_{i+1}]} \in \mathbb{P}_m,\ i = 1,2,\ldots,n-1\}$?

72. Given the subdivision $\Delta:\ a = x_1 < x_2 < \cdots < x_n = b$ of $[a,b]$, determine a basis of "hat functions" for the space $\boldsymbol{S} = \{s \in \mathbb{S}_1^0:\ s(a) = s(b) = 0\}$.

73. Let $\Delta : a = x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b$ be a subdivision of $[a, b]$ into $n-1$ subintervals. Suppose we are given values $f_i = f(x_i)$ of some function $f(x)$ at the points $x = x_i$, $i = 1, 2, \ldots, n$. In this problem $s \in \mathbb{S}_2^1$ is a quadratic spline in $C^1[a, b]$ that interpolates f on Δ, that is, $s(x_i) = f_i$, $i = 1, 2, \ldots, n$.

    (a) Explain why one expects an additional condition to be required in order to determine s uniquely.

    (b) Define $m_i = s'(x_i)$, $i = 1, 2, \ldots, n-1$. Determine $p_i := s\,\big|_{[x_i, x_{i+1}]}$, $i = 1, 2, \ldots, n-1$, in terms of f_i, f_{i+1}, and m_i.

    (c) Suppose one takes $m_1 = f'(a)$. (According to (a), this determines s uniquely.) Show how $m_2, m_3, \ldots, m_{n-1}$ can be computed.

74. Let the subdivision Δ of $[a, b]$ be given by

$$\Delta : \quad a = x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b, \quad n \geq 2,$$

    and let $f_i = f(x_i)$, $i = 1, 2, \ldots, n$, for some function f. Suppose you want to interpolate this data by a quintic spline $s_5(f; \cdot)$ (a piecewise fifth-degree polynomial of smoothness class $C^4[a, b]$). By counting the number of parameters at your disposal and the number of conditions imposed, state how many additional conditions (if any) you expect are needed to make $s_5(f; \cdot)$ unique.

75. Let

$$\Delta : \quad a = x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b.$$

    Consider the following problem: given $n-1$ numbers f_ν and $n-1$ points ξ_ν with $x_\nu < \xi_\nu < x_{\nu+1}$ $(\nu = 1, 2, \ldots, n-1)$, find a piecewise linear function $s \in \mathbb{S}_1^0(\Delta)$ such that

$$s(\xi_\nu) = f_\nu \quad (\nu = 1, 2, \ldots, n-1), \quad s(x_1) = s(x_n).$$

    Representing s in terms of the basis B_1, $B_2, \ldots, B_n$ of "hat functions," determine the structure of the linear system of equations that you obtain for the coefficients c_j in $s(x) = \sum_{j=1}^n c_j B_j(x)$. Describe how you would solve the system.

76. Let $s_1(x) = 1 + c(x+1)^3$, $-1 \leq x \leq 0$, where c is a (real) parameter. Determine $s_2(x)$ on $0 \leq x \leq 1$ so that

$$s(x) := \begin{cases} s_1(x) & \text{if} \quad -1 \leq x \leq 0, \\ s_2(x) & \text{if} \quad 0 \leq x \leq 1 \end{cases}$$

    is a natural cubic spline on $[-1, 1]$ with knots at -1, 0, 1. How must c be chosen if one wants $s(1) = -1$?

77. Derive (2.136).

78. Determine the quantities m_i in the variant of piecewise cubic Hermite interpolation mentioned at the end of Sect. 2.3.4(a).

79. (a) Derive the two extra equations for $m_1,\ m_2, \ldots, m_n$ that result from the "not-a-knot" condition (Sect. 2.3.4(b.4)) imposed on the cubic spline interpolant $s \in \mathbb{S}_3^2(\Delta)$ (with Δ as in Ex. 73).

    (b) Adjoin the first of these equations to the top and the second to the bottom of the system of $n-2$ equations derived in Sect. 2.3.4(b). Then apply elementary row operations to produce a tridiagonal system. Display the new matrix elements in the first and last equations, simplified as much as possible.

    (c) Is the tridiagonal system so obtained diagonally dominant?

80. Let $\mathbb{S}_1^0(\Delta)$ be the class of continuous piecewise linear functions relative to the subdivision $a = x_1 < x_2 < \cdots < x_n = b$. Let $\|g\|_\infty = \max_{a\le x\le b} |g(x)|$, and denote by $s_1(g;\,\cdot\,)$ the piecewise linear interpolant (from $\mathbb{S}_1^0(\Delta)$) to g.

    (a) Show: $\|s_1(g;\,\cdot\,)\|_\infty \le \|g\|_\infty$ for any $g \in C[a,b]$.

    (b) Show: $\|f - s_1(f;\,\cdot\,)\|_\infty \le 2\|f-s\|_\infty$ for any $s \in \mathbb{S}_1^0$, $f \in C[a,b]$. {*Hint*: use additivity of $s_1(f;\,\cdot\,)$ with respect to f.}

    (c) Interpret the result in (b) when s is the best uniform spline approximant to f.

81. Consider the interval $[a,b] = [-1,1]$ and its subdivision $\Delta = [-1,0] \cup [0,1]$, and let $f(x) = \cos \frac{\pi}{2}\, x,\ -1 \le x \le 1$.

    (a) Determine the natural cubic spline interpolant to f on Δ.

    (b) Illustrate Theorem 2.3.2 by taking in turn $g(x) = p_2(f; -1, 0, 1; x)$ and $g(x) = f(x)$.

    (c) Discuss analogously the complete cubic spline interpolant to f on Δ' (cf. (2.149)) and the choices $g(x) = p_3(f; -1, 0, 1, 1; x)$ and $g(x) = f(x)$.

MACHINE ASSIGNMENTS

1. (a) A simple-minded approach to best uniform approximation of a function $f(x)$ on [0,1] by a linear function $ax+b$ is to first discretize the problem and then, for various (appropriate) trial values of a, solve the problem of (discrete) uniform approximation of $f(x)-ax$ by a constant b (which admits an easy solution). Write a program to implement this idea.

    (b) Run your program for $f(x) = \mathrm{e}^x,\ f(x) = 1/(1+x),\ f(x) = \sin \frac{\pi}{2} x,\ f(x) = x^\alpha$ $(\alpha = 2, 3, 4, 5)$. Print the respective optimal values of a and b and the associated minimum error. What do you find particularly interesting in the results (if anything)?

(c) Give a heuristic explanation (and hence exact values) for the results, using the known fact that the error curve for the optimal linear approximation attains its maximum modulus at three consecutive points $0 \le x_0 < x_1 < x_2 \le 1$ with alternating signs (*Principle of Alternation*).

2. (a) Determine the $(n+1) \times (n+1)$ matrix $\boldsymbol{A} = [a_{ij}]$, $a_{ij} = (B_i^n, B_j^n)$, of the normal equations relative to the Bernstein basis

$$B_j^n(t) = \binom{n}{j} t^j (1-t)^{n-j}, \quad j = 0, 1, \ldots, n,$$

and weight function $w(t) \equiv 1$ on [0,1]. {*Point of information*: $\int_0^1 t^k (1-t)^\ell \mathrm{d}t = k!\ell!/(k+\ell+1)!$}.

(b) Use Matlab to solve the normal equations of (a) for $n = 5:5:25$, when the function to be approximated is $f(t) \equiv 1$. What should the exact answer be? For each n, print the infinity norm of the error vector and an estimate of the condition number of $\boldsymbol{A}$. Comment on your results.

3. Compute discrete least squares approximations to the function $f(t) = \sin\left(\frac{\pi}{2}t\right)$ on $0 \le t \le 1$ by polynomials of the form

$$\varphi_n(t) = t + t(1-t)\sum_{j=1}^{n} c_j t^{j-1}, \ n = 1(1)5,$$

using N abscissae $t_k = k/(N+1)$, $k = 1, 2, \ldots, N$, and equal weights 1. Note that $\varphi_n(0) = 0$, $\varphi_n(1) = 1$ are the exact values of f at $t = 0$ and $t = 1$, respectively. {*Hint*: approximate $f(t) - t$ by a linear combination of $\pi_j(t) = t^j(1-t)$; $j = 1, 2, \ldots, n$.} Write a Matlab program for solving the normal equations $\boldsymbol{Ac} = \boldsymbol{b}$, $\boldsymbol{A} = [(\pi_i, \pi_j)]$, $\boldsymbol{b} = [(\pi_i, f-t)]$, $\boldsymbol{c} = [c_j]$, that does the computation in both single and double precision. For each $n = 1, 2, \ldots, 5$ output the following:

- the condition number of the system (computed in double precision);
- the maximum relative error in the coefficients, $\max_{1 \le j \le n} |(c_j^s - c_j^d)/c_j^d|$, where c_j^s are the single-precision values of c_j and c_j^d the double-precision values;
- the minimum and maximum error (computed in double precision),

$$e_{\min} = \min_{1 \le k \le N} |\varphi_n(t_k) - f(t_k)|, \quad e_{\max} = \max_{1 \le k \le N} |\varphi_n(t_k) - f(t_k)|.$$

Make two runs:

(a) $N = 5, 10, 20$; (b) $N = 4$.

Comment on the results.

4. Write a program for discrete polynomial least squares approximation of a function f defined on [–1,1], using the inner product

$$(u, v) = \frac{2}{N+1} \sum_{i=0}^{N} u(t_i)v(t_i), \quad t_i = -1 + \frac{2i}{N}.$$

Follow these steps.

(a) The recurrence coefficients for the appropriate (monic) orthogonal polynomials $\{\pi_k(t)\}$ are known explicitly:

$$\alpha_k = 0, \quad k = 0, 1, \ldots, N; \qquad \beta_0 = 2,$$

$$\beta_k = \left(1 + \frac{1}{N}\right)^2 \left(1 - \left(\frac{k}{N+1}\right)^2\right) \left(4 - \frac{1}{k^2}\right)^{-1}, \quad k = 1, 2, \ldots, N.$$

(You do not have to prove this.) Define $\gamma_k = \|\pi_k\|^2 = (\pi_k, \pi_k)$, which is known to be equal to $\beta_0\beta_1 \cdots \beta_k$ (cf. Ex. 20).

(b) Using the recurrence formula with coefficients α_k, β_k given in (a), generate an array $\boldsymbol{\pi}$ of dimension $(N + 2, N + 1)$ containing $\pi_k(t_\ell)$, $k = 0, 1, \ldots, N + 1$; $\ell = 0, 1, \ldots, N$. (Here k is the row index and ℓ the column index.) Define $\mu_k = \max_{0\le\ell\le N} |\pi_k(t_\ell)|$, $k = 1, 2, \ldots, N+1$. Print β_k, γ_k, and μ_{k+1} for $k = 0, 1, 2, \ldots, N$, where $N = 10$. Comment on the results.

(c) With $\hat{p}_n(t) = \sum_{k=0}^{n} \hat{c}_k\pi_k(t)$, $n = 0, 1, \ldots, N$, denoting the least squares approximation of degree $\le n$ to the function f on [–1,1], define

$$\|e_n\|_2 = \|\hat{p}_n - f\| = (\hat{p}_n - f, \hat{p}_n - f)^{1/2},$$

$$\|e_n\|_\infty = \max_{0\le i\le N} |\hat{p}_n(t_i) - f(t_i)|.$$

Using the array $\boldsymbol{\pi}$ generated in part (b), compute $\hat{c}_n$, $\|e_n\|_2$, $\|e_n\|_\infty$, $n = 0, 1, \ldots, N$, for the following four functions:

$$f(t) = \mathrm{e}^{-t}, \quad f(t) = \ln(2 + t), \quad f(t) = \sqrt{1 + t}, \quad f(t) = |t|.$$

Be sure you compute $\|e_n\|_2$ as accurately as possible. For $N = 10$ and for each f, print $\hat{c}_n$, $\|e_n\|_2$, and $\|e_n\|_\infty$ for $n = 0, 1, 2, \ldots, N$. Comment on your results. In particular, from the information provided in the output, discuss to what extent the computed coefficients $\hat{c}_k$ may be corrupted by rounding errors.

5. (a) A Sobolev-type least squares approximation problem results if the inner product is defined by

$$(u, v) = \int_{\mathbb{R}} u(t)v(t)\mathrm{d}\lambda_0(t) + \int_{\mathbb{R}} u'(t)v'(t)\mathrm{d}\lambda_1(t),$$

where $\mathrm{d}\lambda_0$, $\mathrm{d}\lambda_1$ are positive measures. What does this type of approximation try to accomplish?

(b) Letting $\mathrm{d}\lambda_0(t) = \mathrm{d}t,\ \mathrm{d}\lambda_1(t) = \lambda \mathrm{d}t$ on [0,2], where $\lambda > 0$ is a parameter, set up the normal equations for the Sobolev-type approximation in (a) of the function $f(t) = \mathrm{e}^{-t^2}$ on [0,2] by means of a polynomial of degree $n-1$. Use the basis $\pi_j(t) = t^{j-1},\ j = 1, 2, \ldots, n$. {*Hint*: express the components b_i of the right-hand vector of the normal equations in terms of the "incomplete gamma function" $\gamma(a, x) = \int_0^x t^{a-1}e^{-t}\mathrm{d}t$ with $x = 4,\ a = i/2$.}

(c) Use Matlab to solve the normal equations for $n = 2 : 5$ and $\lambda = 0, .5, 1, 2$. Print

$$\|\hat{\varphi}_n - f\|_\infty \quad \text{and} \quad \|\hat{\varphi}_n' - f'\|_\infty, \quad n = 2, 3, 4, 5$$

(or a suitable approximation thereof) along with the condition numbers of the normal equations. {Use the following values for the incomplete gamma function: $\gamma(\frac{1}{2}, 4) = 1.764162781524843,\ \gamma(1, 4) = .9816843611112658, \gamma(\frac{3}{2}, 4) = .8454501129849537, \gamma(2, 4) = .9084218055563291,\ \gamma(\frac{5}{2}, 4) = 1.121650058367554$.} Comment on the results.

6. With $\omega_n(x) = \prod_{k=0}^n (x-k)$, let M_n be the largest, and m_n the smallest, relative maximum of $|\omega_n(x)|$. For $n = 5 : 5 : 30$ calculate M_n, m_n, and M_n/m_n, using Newton's method (cf. Chap 4, Sect. 4.6), and print also the respective number of iterations.

7. (a) Write a subroutine that produces the value of the interpolation polynomial $p_n(f; x_0, x_1, \ldots, x_n; t)$ at any real t, where $n \geq 0$ is a given integer, x_i are $n+1$ distinct nodes, and f is any function available in the form of a function subroutine. Use Newton's interpolation formula and exercise frugality in the use of memory space when generating the divided differences. It is possible, indeed, to generate them "in place" in a single array of dimension $n+1$ that originally contains the values $f(x_i),\ i = 0, 1, \ldots, n$. {*Hint*: generate the divided differences from the bottom up.}

(b) Run your routine on the function $f(t) = \frac{1}{1+t^2},\ -5 \leq t \leq 5$, using $x_i = -5 + 10\frac{i}{n},\ i = 0, 1, \ldots, n$, and $n = 2 : 2 : 8$ (Runge's example). Plot the polynomials against the exact function.

8. (a) Write a Matlab function `y=tridiag(n,a,b,c,v)` for solving a tridiagonal (nonsymmetric) system

$$\begin{bmatrix} a_1 & c_1 & & & & 0 \\ b_1 & a_2 & c_2 & & & \\ & b_2 & a_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & c_{n-1} \\ 0 & & & & b_{n-1} & a_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix}$$

by Gauss elimination without pivoting. Keep the program short.

(b) Write a program for computing the natural spline interpolant $s_{\text{nat}}(f;\,\cdot\,)$ on an arbitrary partition $a = x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b$ of $[a, b]$. Print $\{i, \text{errmax}(i);\ i = 1, 2, \ldots, n-1\}$, where

$$\text{errmax}(i) = \max_{1 \le j \le N} |s_{\text{nat}}(f; x_{i,j}) - f(x_{i,j})|\,, \quad x_{i,j} = x_i + \frac{j-1}{N-1}\,\Delta x_i.$$

(You will need the function `tridiag`.) Test the program for cases in which the error is zero (what are these, and why?).

(c) Write a second program for computing the complete cubic spline interpolant $s_{\text{compl}}(f;\,\cdot\,)$ by modifying the program in (b) with a minimum of changes. Highlight the changes in the program listing. Apply (and justify) a test similar to that of (b).

(d) Run the programs in (b) and (c) for $[a, b] = [0, 1]$, $n = 11$, $N = 51$, and

(i) $x_i = \frac{i-1}{n-1}$, $i = 1, 2, \ldots, n$; $f(x) = \mathrm{e}^{-x}$ and $f(x) = x^{5/2}$;

(ii) $x_i = \left(\frac{i-1}{n-1}\right)^2$, $i = 1, 2, \ldots, n$; $f(x) = x^{5/2}$.

Comment on the results.

ANSWERS TO THE EXERCISES AND MACHINE ASSIGNMENTS OF CHAPTER 2

ANSWERS TO EXERCISES

1. Let $\|f-\varphi\| = \delta(c)$.

   (a) We have

$$\delta(c) = \int_{-1}^{1} |f-\varphi| \mathrm{d}t = \int_{-1}^{0} |-1-c| \mathrm{d}t + \int_{0}^{1} |1-c| \mathrm{d}t = |1+c| + |1-c|$$

$$= \begin{cases} 1+c+1-c = 2 & \text{if} \quad -1 \le c \le 1, \\ 1+c+c-1 = 2c > 2 & \text{if} \quad c > 1, \\ -c-1+1-c = -2c > 2 & \text{if} \quad c < -1. \end{cases}$$

   Thus, $c_{\text{opt}} = c$ for any $-1 \le c \le 1$, with $\delta_{\text{opt}} = 2$. There is *nonuniqueness.*

   (b) We have $\delta(c) = \sum_{i=1}^{N} |f(t_i) - c|$.

   Case 1: N even. Here,

$$\delta(c) = \frac{N}{2} (|1+c| + |1-c|),$$

   and by (a) again $c_{\text{opt}} = c$ for any $-1 \le c \le 1$, with $\delta_{\text{opt}} = N$. There is *nonuniqueness.*

   Case 2: N odd. Here,

$$\delta(c) = \frac{N-1}{2} (|1+c| + |1-c|) + |-c|$$

$$= \begin{cases} N-1+|c| & \text{if} \quad -1 \le c \le 1, \\ (N-1)c + c = Nc & \text{if} \quad c > 1, \\ -(N-1)c - c = -Nc & \text{if} \quad c < -1. \end{cases}$$

   Thus, $c_{\text{opt}} = 0$ and $\delta_{\text{opt}} = N-1$. There is *uniqueness.*

   (c) We have

$$\delta^2(c) = \int_{-1}^{0} (1+c)^2 \mathrm{d}t + \int_{0}^{1} (1-c)^2 \mathrm{d}t = (1+c)^2 + (1-c)^2 = 2(1+c^2).$$

   Thus, $c_{\text{opt}} = 0$ and $\delta_{\text{opt}} = \sqrt{2}$. There is *uniqueness.*

(d) We have $\delta^2(c) = \sum_{i=1}^{N} [f(t_i) - c]^2$.

Case 1: N even. Here,

$$\delta^2(c) = \frac{N}{2}\left((1+c)^2 + (1-c)^2\right) = N(1+c^2),$$

and $c_{\text{opt}} = 0$, $\delta_{\text{opt}} = \sqrt{N}$, *uniquely.*

Case 2: N odd. Here,

$$\delta^2(c) = \frac{N-1}{2}\left((1+c)^2 + (1-c)^2\right) + c^2 = (N-1)(1+c^2) + c^2,$$

giving $c_{\text{opt}} = 0$, with $\delta_{\text{opt}} = \sqrt{N-1}$, *uniquely.*

(e) We have

$$\delta(c) = \max\left(|-1-c|, |c|, |1-c|\right) = \max\left(|1+c|, |c|, |1-c|\right)$$
$$= \begin{cases} 1+c & \text{if} \quad c \geq 0, \\ 1-c & \text{if} \quad c < 0, \end{cases}$$

giving $c_{\text{opt}} = 0$, with $\delta_{\text{opt}} = 1$, *uniquely.*

(f) *Case* 1: N even. Here,

$$\delta(c) = \max\left(|1+c|, |1-c|\right).$$

Case 2: N odd. Here,

$$\delta(c) = \max\left(|1+c|, |c|, |1-c|\right).$$

In either case, $c_{\text{opt}} = 0$, $\delta_{\text{opt}} = 1$, *uniquely.*

2. (a) Clearly, $c = (1+y)/2$, since any different value of c would necessarily increase the L_∞ error.

(b) The square of the L_2 error, $(N-1)(c-1)^2 + (c-y)^2$ (a quadratic function in c), is minimized if $2(N-1)(c-1) + 2(c-y) = 0$, that is, if

$$c = 1 + \frac{y-1}{N}\,.$$

(c) The L_∞ approximant is sensitive to an "outlier", the least-squares approximant much less so. Indeed, as $N \to \infty$, the latter tends to 1, that is, it ignores the outlier altogether, whereas the former is strongly influenced by the outlier.

3. With notations as in the *Hint*, we have

$$\begin{aligned} f(x) - p(x) &= f(x) - p_n(f;x) - \omega_n(x)q(x) \\ &= \omega_n(x)\left\{\frac{f(x) - p_n(f;x)}{\omega_n(x)} - q(x)\right\}. \end{aligned}$$

Note that, by the rule of Bernoulli-L'Hospital, the fraction in curled brackets has finite limits as $x \to x_i$, since $f \in C^1[a,b]$:

$$\lim_{x\to x_i} \frac{f(x) - p_n(f;x)}{\omega_n(x)} = \frac{f'(x_i) - p_n'(f;x_i)}{\omega_n'(x_i)}, \quad i = 0, 1, \ldots, n.$$

Thus,

$$\frac{f(x) - p_n(f;x)}{\omega_n(x)} \ \in \ C[a,b].$$

Now

$$\|f - p\|_\infty \le \|\omega_n\|_\infty \left\| \frac{f - p_n(f;\cdot)}{\omega_n} - q \right\|_\infty$$

and $\|\omega_n\|_\infty \le (b-a)^{n+1}$. By Weierstrass's theorem, there exists an $m = m(\varepsilon)$ sufficiently large and $q \in \mathbb{P}_m$ such that

$$\left\| \frac{f - p_n(f;\cdot)}{\omega_n} - q \right\|_\infty \le \frac{\varepsilon}{(b-a)^{n+1}}.$$

There follows

$$\|f - p\|_\infty \le (b-a)^{n+1} \frac{\varepsilon}{(b-a)^{n+1}} = \varepsilon.$$

Clearly, $p = p_n(f;\cdot) + \omega_n q$ satisfies $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$, for any q.

4. *The case* $p = \infty$.

We have

$$E_\infty(c) = \max_{0\le t\le 1} |t^\alpha - c| = \max(c, 1-c).$$

This is minimized for $c_\infty = \frac{1}{2}$ and yields $E_\infty(c_\infty) = \frac{1}{2}$.

The case $p = 2$.

We have

$$[E_2(c)]^2 = \int_0^1 (t^\alpha - c)^2 \mathrm{d}t.$$

To find the minimum, differentiate with respect to c and set the result equal to zero:

$$\begin{aligned} 2\int_0^1 (t^\alpha - c)(-1)\mathrm{d}t &= -2\int_0^1 (t^\alpha - c)\mathrm{d}t = -2\left(\frac{1}{\alpha+1}\, t^{\alpha+1}\Big|_0^1 - c\right) \\ &= -2\left(\frac{1}{\alpha+1} - c\right) = 0, \end{aligned}$$

giving $c = c_2 := 1/(\alpha+1)$. Furthermore,

$$\begin{aligned}
[E_2(c_2)]^2 &= \int_0^1 \left(t^\alpha - \frac{1}{\alpha+1}\right)^2 \mathrm{d}t \\
&= \int_0^1 \left(t^{2\alpha} - \frac{2}{\alpha+1}\, t^\alpha + \frac{1}{(\alpha+1)^2}\right) \mathrm{d}t \\
&= \left[\frac{1}{2\alpha+1}\, t^{2\alpha+1} - \frac{2}{(\alpha+1)^2}\, t^{\alpha+1} + \frac{1}{(\alpha+1)^2}\, t\right]_0^1 \\
&= \frac{1}{2\alpha+1} - \frac{2}{(\alpha+1)^2} + \frac{1}{(\alpha+1)^2} = \frac{\alpha^2}{(2\alpha+1)(\alpha+1)^2}\,,
\end{aligned}$$

that is,

$$E_2(c_2) = \frac{\alpha}{\alpha+1}\, \frac{1}{\sqrt{2\alpha+1}}\,.$$

The case $p = 1$.

We have, since $0 < c < 1$,

$$\begin{aligned}
E_1(c) &= \int_0^{c^{1/\alpha}} (c - t^\alpha)\mathrm{d}t + \int_{c^{1/\alpha}}^1 (t^\alpha - c)\mathrm{d}t \\
&= \; c \cdot c^{1/\alpha} - \frac{1}{\alpha+1}\, t^{\alpha+1}\Big|_0^{c^{1/\alpha}} + \frac{1}{\alpha+1}\, t^{\alpha+1}\Big|_{c^{1/\alpha}}^1 - c(1 - c^{1/\alpha}) \\
&= c^{(\alpha+1)/\alpha} - \frac{1}{\alpha+1}\, c^{(\alpha+1)/\alpha} + \frac{1}{\alpha+1}\, (1 - c^{(\alpha+1)/\alpha}) \\
&\quad -c + c^{(\alpha+1)/\alpha} \\
&= 2\, \frac{\alpha}{\alpha+1}\, c^{(\alpha+1)/\alpha} - c + \frac{1}{\alpha+1}\,.
\end{aligned}$$

Setting the derivative (with respect to c) equal to zero yields

$$\begin{aligned}
2\, \frac{\alpha}{\alpha+1}\, \frac{\alpha+1}{\alpha}\, c^{1/\alpha} - 1 = 0, \\
c^{1/\alpha} = \frac{1}{2}, \quad c = c_1 := 2^{-\alpha}.
\end{aligned}$$

Furthermore,

$$\begin{aligned} E_1(c_1) &= 2\,\frac{\alpha}{\alpha+1}\,(2^{-\alpha})^{(\alpha+1)/\alpha} - 2^{-\alpha} + \frac{1}{\alpha+1} \\ &= 2\,\frac{\alpha}{\alpha+1}\,2^{-(\alpha+1)} - 2^{-\alpha} + \frac{1}{\alpha+1} \\ &= \left(\frac{\alpha}{\alpha+1} - 1\right) 2^{-\alpha} + \frac{1}{\alpha+1} \\ &= \frac{1}{\alpha+1}\,(1 - 2^{-\alpha}). \end{aligned}$$

5. (a) We want $\mathrm{e}^x - 1 \approx cx$ best in the L_2 sense. We have $\pi_1(x) = x$, $f(x) = \mathrm{e}^x - 1$, $n = 1$, so that the "normal equation" (2.18) simply is

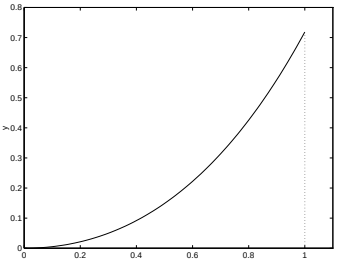
$$(\pi_1, \pi_1)c = (\pi_1, f), \quad \text{i.e.,} \quad \tfrac{1}{3}\,c = \int_0^1 x(\mathrm{e}^x - 1)\mathrm{d}x.$$

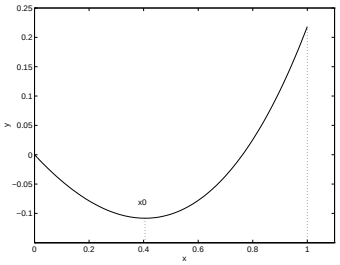
Here, using integration by parts,

$$\begin{aligned} \int_0^1 x(\mathrm{e}^x - 1)\mathrm{d}x &= x(\mathrm{e}^x - x)\Big|_0^1 - \int_0^1 (\mathrm{e}^x - x)\mathrm{d}x \\ &= \mathrm{e} - 1 - (\mathrm{e}^x - \tfrac{1}{2}\,x^2)\big|_0^1 = \mathrm{e} - 1 - (\mathrm{e} - \tfrac{1}{2}) + 1 = \tfrac{1}{2}\,, \end{aligned}$$

so that $c = \frac{3}{2}$.

(b)





$e_1(x) \uparrow$ on $[0,1]$
$e_1(1) = \mathrm{e} - 2 = .718\ldots$
$\max_{[0,1]} |e_1(x)| = .718\ldots$

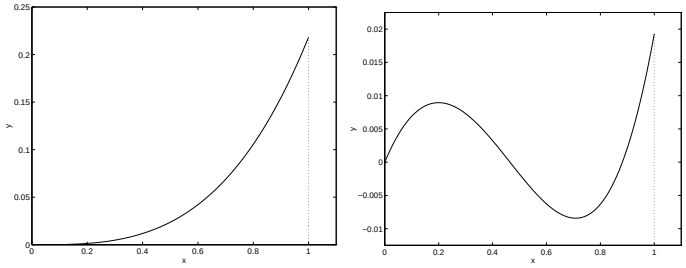
$e_2'(x_0) = 0$ at $x_0 = \ln(3/2) = .405\ldots$
$e_2(x_0) = \frac{1}{2} - \frac{3}{2}\,x_0 = -.108\ldots$
$e_2(1) = \mathrm{e} - 1 - \frac{3}{2} = .218\ldots$
$\max_{[0,1]} |e_2(x)| = .218\ldots$

(c) We want $\mathrm{e}^x - 1 \approx c_1 x + c_2 x^2$, hence $\pi_1(x) = x$, $\pi_2(x) = x^2$, $f(x) = \mathrm{e}^x - 1$, $n = 2$. The normal equations are now

$$\tfrac{1}{3}\,c_1 + \tfrac{1}{4}\,c_2 = \tfrac{1}{2}\,,$$

$$\tfrac{1}{4}\,c_1 + \tfrac{1}{5}\,c_2 = \mathrm{e} - \tfrac{7}{3}\,,$$

giving $c_1 = 164 - 60\mathrm{e} = .903090292\ldots$, $c_2 = 80\mathrm{e} - \frac{650}{3} = .795879610\ldots$.



$\max_{[0,1]} |e_1(x)| = e_1(1)$
$= \mathrm{e} - 2.5 = .218\ldots$
(same as for e_2
in problem (b))

$\max_{[0,1]} |e_2(x)| = e_2(1)$
$= \mathrm{e} - 1 - c_1 - c_2 = .0193\ldots$

6. By the positive definiteness of the inner product one has

$$(u + tv, u + tv) \geq 0, \quad \text{all } t \in \mathbb{R},$$

that is, multiplying out,

$$(u, u) + 2t(u, v) + t^2(v, v) \geq 0, \quad t \in \mathbb{R}.$$

Therefore, the discriminant of the quadratic function must be nonpositive,

$$4(u, v)^2 - 4(u, u) \cdot (v, v) \leq 0,$$

that is,

$$(u, v)^2 \leq (u, u) \cdot (v, v), \quad |(u, v)| \leq \|u\| \cdot \|v\|.$$

7. Let the least squares approximant to f from $\mathbb{P}_{n-1}$ be $\hat{p}_{n-1}$. We have

for $n = 1$: unique $\hat{p}_0 \in \mathbb{P}_0$, namely $\hat{p}_0 = \frac{1}{2}(f_1 + f_2)$;

for $n = 2$: unique $\hat{p}_1 \in \mathbb{P}_1$, namely $\hat{p}_1 = p_1(f; t_1, t_2; t)$ (cf. (2.51) for notation);

for $n \geq 3$: infinitely many $\hat{p}_{n-1} \in \mathbb{P}_{n-1}$, namely all polynomials $\hat{p}_{n-1}$ that interpolate f at t_1 and t_2,

$$\hat{p}_{n-1}(t) = \hat{p}_1(t) + (t - t_1)(t - t_2)q_{n-3}(t),$$

where q_{n-3} is an *arbitrary* polynomial of degree $\leq n - 3$.

8. We have $\pi_1(t) = \dfrac{1}{1+t}$, $\pi_2(t) = \dfrac{1}{(1+t)^2}$, and therefore

$$
\begin{aligned}
(\pi_1, \pi_1) &= \int_0^1 \frac{\mathrm{d}t}{(1+t)^2} = -(1+t)^{-1}\Big|_0^1 = 1 - \tfrac{1}{2} = \tfrac{1}{2}\,, \\
(\pi_1, \pi_2) &= \int_0^1 \frac{\mathrm{d}t}{(1+t)^3} = -\frac{1}{2}(1+t)^{-2}\Big|_0^1 = \tfrac{1}{2}\left(1 - \tfrac{1}{4}\right) = \tfrac{3}{8}\,, \\
(\pi_2, \pi_2) &= \int_0^1 \frac{\mathrm{d}t}{(1+t)^4} = -\frac{1}{3}\,(1+t)^{-3}\Big|_0^1 = \tfrac{1}{3}\,\left(1 - \tfrac{1}{8}\right) = \tfrac{7}{24}\,.
\end{aligned}
$$

The matrix of the normal equations and its inverse are

$$
\boldsymbol{A} = \frac{1}{24}\begin{bmatrix} 12 & 9 \\ 9 & 7 \end{bmatrix},
$$

$$
\boldsymbol{A}^{-1} = 24\begin{bmatrix} 12 & 9 \\ 9 & 7 \end{bmatrix}^{-1} = 24\begin{bmatrix} \frac{7}{3} & -3 \\ -3 & 4 \end{bmatrix} = 8\begin{bmatrix} 7 & -9 \\ -9 & 12 \end{bmatrix}.
$$

Therefore,

$$
\|\boldsymbol{A}\|_\infty = \frac{21}{24} = \frac{7}{8}\,, \quad \|\boldsymbol{A}^{-1}\|_\infty = 8 \cdot 21, \quad \operatorname{cond}_\infty \boldsymbol{A} = \frac{7}{8} \cdot 8 \cdot 21 = 147.
$$

The mth exponential integral (cf. *Point of information*), after the change of variables $xt = \tau$, becomes

$$
E_m(x) = \frac{1}{x}\int_x^\infty \left(\frac{\tau}{x}\right)^{-m} \mathrm{e}^{-\tau}\mathrm{d}\tau = x^{m-1}\int_x^\infty t^{-m}\mathrm{e}^{-t}\mathrm{d}t,
$$

hence,

$$
\int_x^\infty t^{-m}\mathrm{e}^{-t}\mathrm{d}t = x^{1-m}E_m(x), \quad x > 0.
$$

Therefore, with the change of variables $1 + t = \tau$,

$$
\begin{aligned}
(\pi_1, f) &= \int_0^1 \frac{\mathrm{e}^{-t}}{1+t}\,\mathrm{d}t = \int_1^2 \frac{\mathrm{e}^{-(\tau-1)}}{\tau}\,\mathrm{d}\tau = \mathrm{e}\int_1^2 \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t \\
&= \mathrm{e}\left[\int_1^\infty \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t - \int_2^\infty \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t\right] = \mathrm{e}[E_1(1) - E_1(2)].
\end{aligned}
$$

From Table 5.1 of Abramowitz and Stegun [1964], or using the Matlab routine `expint`, we find $E_1(1) = .219383934$, $E_1(2) = .048900511$, giving

$$
(\pi_1, f) = .463421991.
$$

Similarly,

$$(\pi_2, f) = \int_0^1 \frac{\mathrm{e}^{-t}}{(1+t)^2}\,\mathrm{d}t = \mathrm{e}\int_1^2 \frac{\mathrm{e}^{-t}}{t^2}\,\mathrm{d}t$$
$$= \mathrm{e}\left[\int_1^\infty \frac{\mathrm{e}^{-t}}{t^2}\,\mathrm{d}t - \int_2^\infty \frac{\mathrm{e}^{-t}}{t^2}\,\mathrm{d}t\right] = \mathrm{e}[E_2(1) - \tfrac{1}{2}\,E_2(2)].$$

From Table 5.4 of Abramowitz and Stegun [1964], we find $E_2(1) = .1484955$, $E_2(2) = .0375343$, giving

$$(\pi_2, f) = .3526382.$$

The solution of the normal equations is

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} (\pi_1, f) \\ (\pi_2, f) \end{bmatrix} = 8 \begin{bmatrix} 7(\pi_1, f) - 9(\pi_2, f) \\ -9(\pi_1, f) + 12(\pi_2, f) \end{bmatrix}$$

$$= 8 \begin{bmatrix} .07021020 \\ .06086040 \end{bmatrix} = \begin{bmatrix} .5616816 \\ .4868832 \end{bmatrix},$$

from which

$$f(0) - \varphi(0) = 1 - (c_1 + c_2) = -.0485648,$$
$$f(\tfrac{1}{2}) - \varphi(\tfrac{1}{2}) = \mathrm{e}^{-1/2} - \tfrac{2}{3}\,c_1 - \tfrac{4}{9}\,c_2 = .0156837,$$
$$f(1) - \varphi(1) = \mathrm{e}^{-1} - \tfrac{1}{2}\,c_1 - \tfrac{1}{4}\,c_2 = -.0346822.$$

9. We let $\pi_0(t) = 1$, $\pi_1(t) = t$ and $f(t) = \sqrt{1-t^2}$. Then, in the case $w(t) = (1-t^2)^{-1/2}$, we have, using the change of variables $t = \cos\theta$ and the *Points of information*,

$$(\pi_0, \pi_0) = \int_0^1 (1-t^2)^{-1/2}\mathrm{d}t = \int_0^{\pi/2} \mathrm{d}\theta = \frac{\pi}{2}\,,$$
$$(\pi_0, \pi_1) = \int_0^1 (1-t^2)^{-1/2} t\,\mathrm{d}t = \int_0^{\pi/2} \cos\theta\,\mathrm{d}\theta = \sin\theta\big|_0^{\pi/2} = 1,$$
$$(\pi_1, \pi_1) = \int_0^1 (1-t^2)^{-1/2} t^2\,\mathrm{d}t = \int_0^{\pi/2} \cos^2\theta\,\mathrm{d}\theta = \frac{\pi}{4}\,,$$
$$(\pi_0, f) = \int_0^1 \mathrm{d}t = 1, \qquad (\pi_1, f) = \int_0^1 t\,\mathrm{d}t = \frac{1}{2}\,,$$

so that the normal equations become

$$\tfrac{\pi}{2}\, c_0 + c_1 = 1,$$
$$c_0 + \tfrac{\pi}{4}\, c_1 = \tfrac{1}{2}\,,$$

and have the solution

$$c_0 = \frac{2(\pi-2)}{\pi^2-8} \;= 1.2212\ldots, \quad c_1 = \frac{2\pi-8}{\pi^2-8} \;= -.9182\ldots\ .$$

In the case $w(t) = 1$, we get

$$(\pi_0,\pi_0) = \int_0^1 \mathrm{d}t = 1, \quad (\pi_0,\pi_1) = \int_0^1 t\mathrm{d}t = \frac{1}{2}\,, \quad (\pi_1,\pi_1) = \int_0^1 t^2\mathrm{d}t = \frac{1}{3}\,,$$
$$(\pi_0, f) = \int_0^1 \sqrt{1-t^2}\mathrm{d}t = \int_0^{\pi/2} \sin^2\theta\mathrm{d}\theta = \frac{\pi}{4}\,,$$
$$(\pi_1, f) = \int_0^1 \sqrt{1-t^2}\, t\mathrm{d}t = \int_0^{\pi/2} \sin^2\theta\cos\theta\mathrm{d}\theta$$
$$= \int_0^{\pi/2} (1-\cos^2\theta)\cos\theta\mathrm{d}\theta = \sin\theta|_0^{\pi/2} - \int_0^{\pi/2} \cos^3\theta\mathrm{d}\theta = 1 - \frac{2}{3} = \frac{1}{3}\,,$$

and the normal equations are

$$c_0 + \tfrac{1}{2}\, c_1 = \tfrac{\pi}{4}\,,$$
$$\tfrac{1}{2}\, c_0 + \tfrac{1}{3}\, c_1 = \tfrac{1}{3}\,,$$

with solution

$$c_0 = \pi - 2 = 1.1415\ldots, \quad c_1 = 4 - \frac{3\pi}{2} = -.7123\ldots\ .$$

Thus, the line

$$\ell: \quad y(t) = c_0 + c_1 t$$

intersects the y-axis at $y = c_0$ and the t-axis at $t = -c_0/c_1$. For the two weight functions, the values are respectively

$$y = 1.2212\ldots, \quad t = 1.3298\ldots \quad (w(t) = (1-t^2)^{-1/2});$$
$$y = 1.1415\ldots, \quad t = 1.6024\ldots \quad (w(t) = 1).$$

Thus, the weight function $(1-t^2)^{-1/2}$ forces the line ℓ to be steeper and intersect the t-axis closer to 1.

10. (a) One easily shows, by the transformation of variables $t \longmapsto -\tau$, that

$$\int_a^b [f(t) - \hat{\varphi}_n(t)]^2 \omega(t) \mathrm{d}t = \int_a^b [f(t) - \hat{\varphi}_n(-t)]^2 \omega(t) \mathrm{d}t.$$

Hence, if $\hat{\varphi}_n(t)$ minimizes the L_2 error on $[a, b]$, so does $\hat{\varphi}_n(-t)$. By uniqueness, $\hat{\varphi}_n(t) \equiv \hat{\varphi}_n(-t)$ on $[a, b]$, i.e., $\hat{\varphi}_n$ is even.

(b) Here, $n = 3,\ \Phi_3 = \mathbb{P}_2$. Since f is even on [–1,1], so is $\hat{\varphi}_3$ by part (a). It suffices, therefore, to minimize

$$\int_0^1 [f(t) - \varphi(t)]^2 \mathrm{d}t \ \text{ over } \ \varphi(t) = c_0 + c_1 t^2.$$

Hence take as basis functions $\pi_0(t) = 1,\ \pi_1(t) = t^2$, and compute the normal equations:

$$(\pi_0, \pi_0) = \int_0^1 \mathrm{d}t = 1,\ (\pi_0, \pi_1) = \int_0^1 t^2 \mathrm{d}t = \frac{1}{3},\ (\pi_1, \pi_1) = \int_0^1 t^4 \mathrm{d}t = \frac{1}{5},$$
$$(\pi_0, f) = \int_0^1 (1-t) \mathrm{d}t = \frac{1}{2},\ (\pi_1, f) = \int_0^1 (1-t) t^2 \mathrm{d}t = \frac{1}{12},$$

that is,

$$c_0 + \tfrac{1}{3} c_1 = \tfrac{1}{2},$$
$$\tfrac{1}{3} c_0 + \tfrac{1}{5} c_1 = \tfrac{1}{12}.$$

The solution is

$$c_0 = \frac{13}{16},\quad c_1 = -\frac{15}{16},\quad \hat{\varphi}_3(t) = \frac{1}{16}(13 - 15t^2).$$

On $[0, 1]$ the error vanishes where $\hat{\varphi}_3(t) = 1 - t$, i.e., when

$$13 - 15t^2 = 16(1-t),\quad 15t^2 - 16t + 3 = 0,$$
$$t_{1,2} = \frac{1}{15}(8 \pm \sqrt{19}) = \begin{cases} .82392\ldots, \\ .24274\ldots. \end{cases}$$

11. See the text.

12. Using partial fraction expansion, one gets

$$\begin{aligned} (\pi_i, \pi_j) &= \int_{-1}^1 \frac{1}{t - a_i} \frac{1}{t - a_j} \mathrm{d}t \\ &= \frac{1}{a_i - a_j} \int_{-1}^1 \left(\frac{1}{t - a_i} - \frac{1}{t - a_j} \right) \mathrm{d}t = \frac{1}{a_i - a_j} \left(\ln \left| \frac{t - a_i}{t - a_j} \right| \right)_{-1}^1 \end{aligned}$$

$$= \frac{1}{a_i - a_j} \ln \left| \frac{1-a_i}{1-a_j} \frac{1+a_j}{1+a_i} \right| \quad \text{if } i \neq j,$$

$$(\pi_i, \pi_i) = \int_{-1}^{1} \frac{\mathrm{d}t}{(t-a_i)^2} = -(t-a_i)^{-1}\big|_{-1}^{1} = \frac{2}{a_i^2 - 1}.$$

Can π_i be orthogonal to π_j? If it were, we would have $(\pi_i, \pi_j) = 0$ for $i \neq j$, which is equivalent to

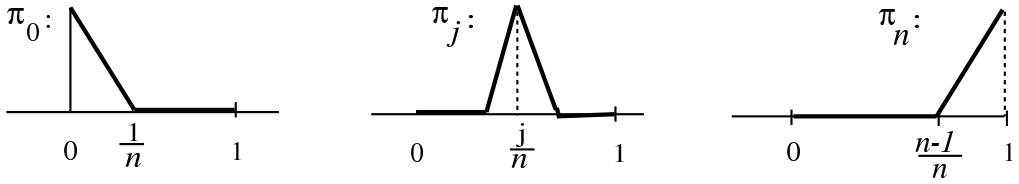
$$\left| \frac{1-a_i}{1-a_j} \frac{1+a_j}{1+a_i} \right| = 1 \text{ for } i \neq j,$$

or to

$$\left| \frac{1-a_i}{1+a_i} \right| = \left| \frac{1-a_j}{1+a_j} \right| \text{ for } i \neq j.$$

A graph of $\left| \frac{1-a}{1+a} \right|$ for $|a| > 1$ shows that the last statement is impossible.

13. (a) Graphs of the functions $\{\pi_j\}$.



Any linear combination $\pi(t)$ of $\{\pi_j\}_{j=0}^n$ is a continuous function that is piecewise linear on the subdivision Δ_n.

(b)

$$\pi_j(k/n) = \delta_{jk} = \begin{cases} 0 & \text{if } j \neq k, \\ 1 & \text{if } j = k. \end{cases}$$

(c) Suppose $\sum_{j=0}^n c_j \pi_j(t) \equiv 0$ on [0,1]. Put $t = k/n$, $k = 0, 1, \ldots, n$, to conclude, by (b), that

$$\sum_{j=0}^{n} c_j \pi_j(k/n) = \sum_{j=0}^{n} c_j \delta_{jk} = c_k = 0, \quad k = 0, 1, \ldots, n.$$

Thus, the system $\{\pi_j\}$ is linearly independent on $[0, 1]$. The argument is also valid if t is restricted to the subdivision points of Δ_n, proving linear independence also on this set of points.

(d) Straightforward calculation, distinguishing between $|i-j| \leq 1$ and $|i-j| > 1$, gives

$$\boldsymbol{A} = \frac{1}{6n} \begin{bmatrix} 2 & 1 & & & & & \mathbf{0} \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & 1 & 4 & 1 \\ \mathbf{0} & & & & & 1 & 2 \end{bmatrix}.$$

14. From Sect. 2.1.3, (2.29), we know that

$$e_{n,2}^2 = \|f\|^2 - \sum_{j=0}^{n} \frac{|(\pi_j, f)|^2}{\|\pi_j\|^2}\,.$$

Here, using the change of variables $t = \mathrm{e}^{-x}$, we have

$$\|f\|^2 = \int_0^1 [\ln(1/t)]^2 \mathrm{d}t = \int_\infty^0 x^2(-\mathrm{e}^{-x})\mathrm{d}x = \int_0^\infty x^2 \mathrm{e}^{-x}\mathrm{d}x = 2! = 2.$$

Moreover, following the *Hint*,

$$|(\pi_j, f)|^2 = \left(\int_0^1 \pi_j(t)\ln(1/t)\mathrm{d}t\right)^2 = \begin{cases} 1, & j=0, \\ \frac{1}{j^2(j+1)^2}\,, & j>0, \end{cases}$$

$$\|\pi_j\|^2 = \int_0^1 \pi_j^2(t)\mathrm{d}t = \frac{1}{2j+1}\,.$$

Therefore,

$$e_{n,2}^2 = 2 - 1 - \sum_{j=1}^{n} \frac{2j+1}{j^2(j+1)^2} = 1 - \sum_{j=1}^{n} \frac{2j+1}{j^2(j+1)^2}\,.$$

Using $2j+1 = (j+1)^2 - j^2$, we can simplify this to

$$e_{n,2}^2 = 1 - \sum_{j=1}^{n} \frac{1}{j^2} + \sum_{j=1}^{n} \frac{1}{(j+1)^2} = 1 - \sum_{j=1}^{n} \frac{1}{j^2} + \sum_{j=2}^{n+1} \frac{1}{j^2} = 1 - 1 + \frac{1}{(n+1)^2} = \frac{1}{(n+1)^2}\,,$$

so that

$$e_{n,2} = \frac{1}{n+1}\,.$$

15. From the discussion in Sect. 2.1.3, it follows that the error $e_n = \hat{p}_{n-1} - f$ is orthogonal to $\mathbb{P}_{n-1}$ and not identically zero by assumption. Now suppose the assertion is false. Then e_n changes sign in $[a, b]$ exactly k times, $k < n$, say at $\tau_1, \tau_2, \ldots, \tau_k$, where $a < \tau_1 < \tau_2 < \cdots < \tau_k < b$. Then (if $k = 0$, the empty product is defined, as usual, by 1)

$$\int_a^b e_n(t) \prod_{\kappa=1}^{k} (t - \tau_\kappa) \mathrm{d}\lambda(t), \qquad k < n,$$

has constant sign on $[a, b]$. But by the orthogonality of e_n to $\mathbb{P}_{n-1}$, the integral is zero. Contradiction!

16. See the text.

17. The problem is to minimize over all $\pi \in \mathbb{P}_n$ the integral

$$\int_a^b \left[f(t) - \frac{\pi(t)}{q(t)} \right]^2 w(t) \mathrm{d}t = \int_a^b [q(t)f(t) - \pi(t)]^2 \frac{w(t)}{q^2(t)} \mathrm{d}t.$$

In view of the second integral, this is the least squares problem of approximating $\overline{f}(t) = q(t)f(t)$ by a polynomial of degree $\leq n$ relative to the weight function $\overline{w}(t) = w(t)/q^2(t)$.

18. (a) Since $B_0^n(t) = (1-t)^n$, we have $B_0^n(0) = 1$.

Using Leibniz's rule of differentiation, we have for $t \to 0$

$$\frac{\mathrm{d}^r}{\mathrm{d}t^r} B_j^n(t) = \binom{n}{j} j(j-1)\cdots(j-r+1)t^{j-r}(1-t)^{n-j} + O(t^{j-r+1}),$$

hence, for $1 \leq j \leq n$,

$$\left.\frac{\mathrm{d}^r}{\mathrm{d}t^r} B_j^n(t)\right|_{t=0} = 0, \quad 0 \leq r < j; \qquad \left.\frac{\mathrm{d}^j}{\mathrm{d}t^j} B_j^n(t)\right|_{t=0} = \binom{n}{j} j! \neq 0.$$

(b) Observe that

$$B_j^n(1-t) = \binom{n}{j} (1-t)^j t^{n-j} = B_{n-j}^n(t),$$

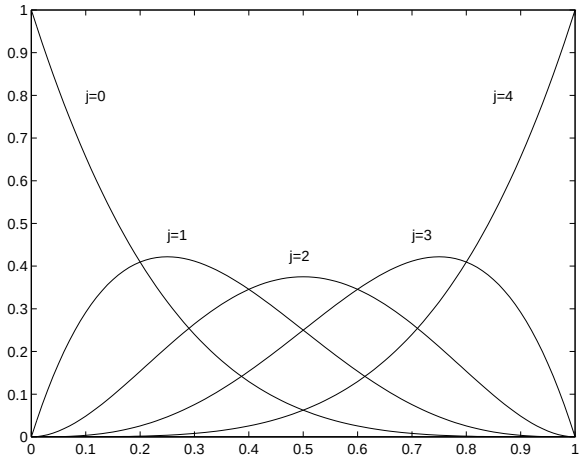
Therefore, if $x = 1 - t$, one gets

$$B_j^n(x) = B_{n-j}^n(1-x).$$

Using (a), one finds

$$B_0^n(0) = B_n^n(1) = 1,$$
$$\frac{\mathrm{d}^r}{\mathrm{d}x^r} B_j^n(x)\Big|_{x=0} = (-1)^r \frac{\mathrm{d}^r}{\mathrm{d}x^r} B_{n-j}^n(1-x)\Big|_{x=0}$$
$$= (-1)^r \frac{\mathrm{d}^r}{\mathrm{d}x^r} B_{n-j}^n(x)\Big|_{x=1} = 0, \quad 1 \le j \le n, \quad 0 \le r < j;$$
$$\frac{\mathrm{d}^j}{\mathrm{d}x^j} B_j^n(x)\Big|_{x=0} = (-1)^j \frac{\mathrm{d}^j}{\mathrm{d}x^j} B_{n-j}^n(1-x)\Big|_{x=0}$$
$$= (-1)^j \frac{\mathrm{d}^j}{\mathrm{d}x^j} B_{n-j}^n(x)\Big|_{x=1} = \binom{n}{j} j!, \quad 1 \le j \le n.$$

(c) Plot of $B_j^n(t)$ for $n = 4,\ 0 \le j \le 4$.



(d) Assume

$$\sum_{j=0}^{n} c_j B_j^n(t) \equiv 0.$$

Then, since $B_0^n(0) = 1$ and $B_j^n(0) = 0$ for $j > 0$, putting $t = 0$ in the above identity gives $c_0 = 0$. Now differentiate the identity and put $t = 0$; because of (a), this yields $c_1 = 0$. Next, differentiate the identity twice and put $t = 0$ to get $c_2 = 0$. Continuing in this manner yields successively $c_3 = 0, \ldots, c_n = 0$. This proves linear independence.

Let $p \in \mathbb{P}_n$ be arbitrary. Set

$$p(t) = \sum_{j=0}^{n} c_j B_j^n(t)$$

and use (a) to get

$$\begin{aligned} c_0 &= p(0), \\ c_0[B_0^n(0)]' + c_1[B_1^n(0)]' &= p'(0), \\ \cdots & \quad \cdots \\ c_0[B_0^n(0)]^{(n)} + c_1[B_1^n(0)]^{(n)} + \cdots + c_n[B_n^n(0)]^{(n)} &= p^{(n)}(0). \end{aligned}$$

Since $[B_j^n(0)]^{(j)} \neq 0$ for $j \geq 1$, this is a lower triangular nonsingular system for the coefficients c_j and therefore has a unique solution. This proves that $\{B_j^n\}_{j=0}^n$ spans $\mathbb{P}_n$.

(e) Using the binomial theorem, one gets

$$\sum_{j=0}^{n} B_j^n(t) = \sum_{j=0}^{n} \binom{n}{j} t^j (1-t)^{n-j} = (t + (1-t))^n = 1.$$

19. It suffices to show that

$$\boldsymbol{A}\boldsymbol{x} = \mathbf{0} \quad \text{for some} \quad \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x} \neq \mathbf{0}.$$

By assumption, there exists $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^T = [x_1, x_2, \ldots, x_n] \neq \mathbf{0^T}$, such that

$$\sum_{j=1}^{n} x_j \pi_j(t) \equiv 0 \quad \text{for} \quad t \in \operatorname{supp} d\lambda.$$

Taking the inner product of both sides with π_i gives

$$\sum_{j=1}^{n} x_j (\pi_i, \pi_j) = 0,$$

that is

$$\sum_{j=1}^{n} a_{ij} x_j = 0$$

as desired.

20. From Sect. 2.1.4, (2.41), we know that

$$\beta_k = \frac{\|\pi_k\|^2}{\|\pi_{k-1}\|^2}, \quad k = 1, 2, \ldots,$$

whereas by convention, $\beta_0 = \|\pi_0\|^2$. Therefore,

$$\begin{aligned} \|\pi_k\|^2 = \beta_k \|\pi_{k-1}\|^2 = \beta_k \beta_{k-1} \|\pi_{k-2}\|^2 = \cdots \\ = \beta_k \beta_{k-1} \cdots \beta_1 \|\pi_0\|^2 = \beta_k \beta_{k-1} \cdots \beta_1 \beta_0. \end{aligned}$$

21. (a) Inserting $\pi_k = \|\pi_k\|\tilde{\pi}_k$ into the three-term recurrence relation (2.38) for the monic orthogonal polynomials and dividing the result by $\|\pi_{k+1}\|$, one gets

$$\tilde{\pi}_{k+1}(t) = \frac{\|\pi_k\|}{\|\pi_{k+1}\|}(t-\alpha_k)\tilde{\pi}_k(t) - \frac{\|\pi_{k-1}\|}{\|\pi_{k+1}\|}\beta_k\tilde{\pi}_{k-1}(t).$$

Since, by (2.41),

$$\frac{\|\pi_k\|}{\|\pi_{k+1}\|} = 1/\sqrt{\beta_{k+1}}\,, \quad k \geq 0,$$
$$\frac{\|\pi_{k-1}\|}{\|\pi_{k+1}\|} = \frac{\|\pi_{k-1}\|}{\|\pi_k\|}\,\frac{\|\pi_k\|}{\|\pi_{k+1}\|} = 1/\sqrt{\beta_k\beta_{k+1}}\,, \quad k \geq 1,$$

the assertion follows upon multiplication by $\sqrt{\beta_{k+1}}$. The initial values follow from the definition $\pi_{-1} = 0$ and the convention about β_0.

(b) Write the recurrence relation of (a) in the form

$$t\tilde{\pi}_k(t) = \alpha_k\tilde{\pi}_k(t) + \sqrt{\beta_{k+1}}\tilde{\pi}_{k+1}(t) + \sqrt{\beta_k}\tilde{\pi}_{k-1}(t).$$

Multiply both sides by $\tilde{\pi}_k(x)$ and subtract the resulting relation from the same relation with x and t interchanged. This gives

$$\begin{aligned}(x-t)\tilde{\pi}_k(x)\tilde{\pi}_k(t) = {}& \sqrt{\beta_{k+1}}\,[\tilde{\pi}_{k+1}(x)\tilde{\pi}_k(t) - \tilde{\pi}_k(x)\tilde{\pi}_{k+1}(t)] \\ & -\sqrt{\beta_k}\,[\tilde{\pi}_k(x)\tilde{\pi}_{k-1}(t) - \tilde{\pi}_{k-1}(x)\tilde{\pi}_k(t)].\end{aligned}$$

Now summing both sides from $k = 0$ to $k = n$ and observing $\tilde{\pi}_{-1} = 0$ and the telescoping nature of the summation on the right yields the desired result.

22. (a) Writing, according to the *Hint*, $p = \pi_n + \sum_{j=0}^{n-1} c_j\pi_j$, we have by orthogonality

$$\begin{aligned}\int_{\mathbb{R}} p^2(t)\mathrm{d}\lambda(t) &= \int_{\mathbb{R}} \pi_n^2(t)\mathrm{d}\lambda(t) + 2\int_{\mathbb{R}} \sum_{j=0}^{n-1} c_j\pi_n(t)\pi_j(t)\mathrm{d}\lambda(t) \\ &\quad + \int_{\mathbb{R}} \left(\sum_{j=0}^{n-1} c_j\pi_j(t)\right)^2 \mathrm{d}\lambda(t) = \int_{\mathbb{R}} \pi_n^2(t)\mathrm{d}\lambda(t) + \int_{\mathbb{R}} \left(\sum_{j=0}^{n-1} c_j\pi_j(t)\right)^2 \mathrm{d}\lambda(t) \\ &\geq \int_{\mathbb{R}} \pi_n^2(t)\mathrm{d}\lambda(t).\end{aligned}$$

Equality holds if and only if

$$\sum_{j=0}^{n-1} c_j\pi_j(t) \equiv 0 \quad \text{for} \quad t \in \operatorname{supp}\mathrm{d}\lambda,$$

which implies (again by orthogonality) $c_j = 0$ for all j, that is, $p \equiv \pi_n$.

(b) By construction (definition of $\alpha_{N-1}, \beta_{N-1}$), the polynomial π_N is orthogonal to all π_j, $0 \le j \le N-1$. Let $p(t) = \prod_{j=1}^{N}(t - t_j)$. Since $p - \pi_N$ is a polynomial of degree $\le N-1$, one has $p = \pi_N + \sum_{j=0}^{N-1} c_j \pi_j$ for some c_j and, with $\|\cdot\| = \|\cdot\|_{\mathrm{d}\lambda_N}$,

$$\|p\|^2 = \|\pi_N\|^2 + \sum_{j=0}^{N-1} c_j^2 \, \|\pi_j\|^2.$$

Since $\|p\|^2 = 0$, there follows $\|\pi_N\|^2 = 0$, hence $\pi_N(t_j) = 0$ for $j = 1, 2, \ldots, N$.

23. (a) Let $(\cdot, \cdot)$ be the inner product induced by $\mathrm{d}\lambda$, and $\lambda_j := (\pi_j, \pi_j)$. Clearly, $\lambda_j > 0$. For $\{p_i\}$ to be an orthogonal system of polynomials, the following conditions must hold:

(i) $\boldsymbol{A}$ is lower triangular, and

(ii) $(p_i, p_k) = 0$ if $i \ne k$.

Condition (ii), by orthogonality of the π_j, implies

$$0 = \left(\sum_j a_{ij}\pi_j, \sum_\ell a_{k\ell}\pi_\ell \right) = \sum_{j,\ell} a_{ij} a_{k\ell} (\pi_j, \pi_\ell) = \sum_j a_{ij} a_{kj} \lambda_j, \quad i \ne k,$$

Thus, the matrix $\boldsymbol{A}$ must satisfy the conditions

$$\sum_j a_{ij} a_{kj} \lambda_j = 0 \quad \text{for} \quad i \ne k.$$

(b) The polynomials p_i are monic if and only if $\boldsymbol{A}$ is unit lower triangular. Let

$$\alpha_i = \sum_j a_{ij}^2 \lambda_j, \quad q_{ij} = \sqrt{\frac{\lambda_j}{\alpha_i}}\, a_{ij}.$$

Then $\boldsymbol{Q} = [q_{ij}]$ is orthogonal. Indeed, from (a(ii)) one has, for $i \ne k$,

$$\sum_j \sqrt{\frac{\alpha_i}{\lambda_j}}\, q_{ij} \sqrt{\frac{\alpha_k}{\lambda_j}}\, q_{kj} \lambda_j = 0,$$

$$\sqrt{\alpha_i \alpha_k} \sum_j q_{ij} q_{kj} = 0,$$

hence

$$\sum_j q_{ij} q_{kj} = 0, \quad i \ne k.$$

Moreover, if $i = k$,

$$\alpha_i = \sum_j a_{ij}^2 \lambda_j = \sum_j \frac{\alpha_i}{\lambda_j}\, q_{ij}^2 \lambda_j = \alpha_i \sum_j q_{ij}^2,$$

implying

$$\sum_j q_{ij}^2 = 1, \quad i = 1, 2, \ldots, n,$$

which proves orthogonality of $\boldsymbol{Q}$. But an orthogonal matrix that is also lower triangular (cf. (a(i)) and the definition of $\boldsymbol{Q}$) must be the identity matrix,

$$\boldsymbol{Q} = \boldsymbol{I}.$$

Then, $\boldsymbol{A}$ is diagonal, hence, being unit lower triangular, $\boldsymbol{A} = \boldsymbol{I}$.

(c) By assumption, $\lambda_j = 1$ and $\sum_j a_{ij} a_{kj} = \delta_{ik}$. Thus, $\boldsymbol{A}$ is orthogonal and, by (a), lower triangular, implying that $\boldsymbol{A} = \boldsymbol{I}$.

24. With $\pi_k^*(x)$ as defined in the *Hint*, it is clear that π_k^* is monic of degree k. Furthermore, if $i \neq j$, then

$$\begin{aligned}(\pi_i^*, \pi_j^*)^* &= \sum_{k=1}^N w_k \pi_i^*(x_k) \pi_j^*(x_k) \\ &= \left(\frac{b-a}{2}\right)^{i+j} \sum_{k=1}^N w_k \pi_i(t_k) \pi_j(t_k) = 0,\end{aligned}$$

so that $\{\pi_k^*\}$ are indeed orthogonal with respect to $(u, v)^*$. Now put $t = \frac{2}{b-a}\left(x - \frac{a+b}{2}\right)$ in the recurrence relation for the $\{\pi_k(t)\}$, and multiply through by $\left(\frac{b-a}{2}\right)^{k+1}$. The result is

$$\begin{aligned}\pi_{k+1}^*(x) &= \left[\frac{2}{b-a}\left(x - \frac{a+b}{2}\right) - \alpha_k\right] \frac{b-a}{2} \pi_k^*(x) \\ &\quad -\beta_k \left(\frac{b-a}{2}\right)^2 \pi_{k-1}^*(x) \\ &= \left[x - \left(\frac{a+b}{2} + \frac{b-a}{2}\alpha_k\right)\right] \pi_k^*(x) - \beta_k \left(\frac{b-a}{2}\right)^2 \pi_{k-1}^*(x),\end{aligned}$$

that is,

$$\alpha_k^* = \frac{a+b}{2} + \frac{b-a}{2}\alpha_k, \quad \beta_k^* = \left(\frac{b-a}{2}\right)^2 \beta_k,$$

along with $\pi_0^*(x) = 1$, $\pi_{-1}^*(x) = 0$.

25. The relations ($\star$) in matrix form are

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -(t-\alpha_0) & 1 & 0 & \cdots & 0 & 0 \\ \beta_1 & -(t-\alpha_1) & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \beta_{n-1} & -(t-\alpha_{n-1}) & 1 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

that is,

$$\boldsymbol{L}\boldsymbol{\pi} = \boldsymbol{e}_1, \quad \boldsymbol{e}_1^T = [1, 0, \ldots, 0],$$

with $\boldsymbol{L}$ unit lower triangular, as shown. Likewise, the relations for the u_k in ($\star\star$) are, in matrix form,

$$\boldsymbol{L}^T\boldsymbol{u} = \boldsymbol{c}, \quad \boldsymbol{u}^T = [u_0, u_1, \ldots, u_n], \quad \boldsymbol{c}^T = [c_0, c_1, \ldots, c_n].$$

Therefore,

$$p_n = \boldsymbol{c}^T\boldsymbol{\pi} = \boldsymbol{u}^T\boldsymbol{L}\boldsymbol{\pi} = \boldsymbol{u}^T\boldsymbol{e}_1 = u_0.$$

26. We have

$$\ell_i(x) = \prod_{j\neq i} \frac{x - x_j}{x_i - x_j}\,.$$

If $x = at + b$ maps x_i into t_i, we get

$$\ell_i(at+b) = \prod_{j\neq i} \frac{at + b - (at_j + b)}{at_i + b - (at_j + b)} = \prod_{j\neq i} \frac{t - t_j}{t_i - t_j}\,.$$

27. See the text.

28. (a) Since

$$\frac{\omega_n'(x)}{\omega_n(x)} = \frac{1}{x} + \frac{1}{x-1} + \cdots + \frac{1}{x-n}\,,$$

we have that x_n must satisfy

$$\frac{1}{x_n} = \frac{1}{1-x_n} + \frac{1}{2-x_n} + \cdots + \frac{1}{n-x_n}\,, \quad 0 < x_n < 1.$$

We show that $x_n \to 0$ by contradiction. Suppose that $x_n \geq a$, $0 < a < 1$, as $n \to \infty$. Then

$$\frac{1}{x_n} \leq \frac{1}{a}\,.$$

But since $0 < x_n < 1$, we have

$$\frac{1}{x_n} > 1 + \frac{1}{2} + \cdots + \frac{1}{n}\,.$$

Here the right-hand side, the nth partial sum of the harmonic series, tends to ∞ as $n \to \infty$, contradicting the previous inequality.

(b) We show that $x_{n+1} < x_n$, so that $x_n \downarrow 0$ as $n \to \infty$. Indeed

$$\omega'_{n+1}(x_n) = \omega_{n+1}(x_n)\left\{\frac{1}{x_n} + \frac{1}{x_n - 1} + \cdots + \frac{1}{x_n - n} + \frac{1}{x_n - (n+1)}\right\} = \omega_{n+1}(x_n)\ \frac{1}{x_n - (n+1)}$$

by what was proved in (a). Assume first n to be odd. Then ω_{n+1} is positive on (0,1) and has a local maximum at x_{n+1} with $\omega_{n+1}(x_{n+1}) > 0$. Since by the above equality $\omega'_{n+1}(x_n) < 0$, we see that x_n is located on the descending side of the "hump" $y = \omega_{n+1}(x)$, $0 \le x \le 1$, thus $x_n > x_{n+1}$. If n is even, one has similarly ω_{n+1} negative on $(0, 1)$ and $\omega'_{n+1}(x_n) > 0$, hence the same conclusion.

29. (a) We have

$$\omega_n(n - x) = \prod_{k=0}^{n}(n - x - k) = \prod_{j=0}^{n}(j - x) = (-1)^{n+1}\prod_{j=0}^{n}(x - j).$$

This implies symmetry (when n is odd) or skew symmetry (when n is even) with respect to the midpoint $x = \frac{1}{2}\,n$.

(b) One has

$$\omega_n(x+1) = \prod_{k=0}^{n}(x + 1 - k) = \prod_{j=-1}^{n-1}(x - j) = \frac{x+1}{x-n}\prod_{j=0}^{n}(x - j) = \frac{x+1}{x-n}\,\omega_n(x).$$

Therefore,

$$|\omega_n(x)| = \frac{n-x}{x+1}\ |\omega_n(x+1)| < |\omega_n(x+1)|$$

since $\dfrac{n-x}{x+1} < 1$ if $x > \dfrac{n-1}{2}$.

(c) Let

$$|\omega_n(\xi_k)| = \max_{k \le x \le k+1} |\omega_n(x)| = M_k, \quad 0 \le k < n.$$

Then, using (b) and $\xi_k > \dfrac{n-1}{2}$ if $k \ge \dfrac{n-1}{2}$, we have

$$M_k = |\omega_n(\xi_k)| < |\omega_n(\xi_k + 1)| \le \max_{k+1 \le x \le k+2} |\omega_n(x)| = M_{k+1},$$

whenever $\frac{n-1}{2} \le k < n - 1$.

30. The function φ defined in the *Hint* is linear between the nodes and can be taken constant in the rest of the interval $[a,b]$. Then, in particular, $\|\varphi\|_\infty = 1$. Furthermore, by construction,

$$\begin{aligned} p_n(\varphi; x_\infty) &= \sum_{i=0}^{n} \varphi(x_i)\ell_i(x_\infty) = \sum_{i=0}^{n} [\text{sgn } \ell_i(x_\infty)]\ell_i(x_\infty) \\ &= \sum_{i=0}^{n} |\ell_i(x_\infty)| = \lambda_n(x_\infty) = \|\lambda_n\|_\infty = \Lambda_n. \end{aligned}$$

Since $\|\varphi\|_\infty = 1$, one has

$$\frac{\|p_n(\varphi;\,\cdot\,)\|_\infty}{\|\varphi\|_\infty} \;\geq\; |p_n(\varphi; x_\infty)| = \Lambda_n.$$

But the opposite inequality is always true (cf. Sect. 2.2.1, (2.55)); therefore,

$$\|p_n(\varphi;\,\cdot\,)\|_\infty = \Lambda_n \|\varphi\|_\infty.$$

31. (a) By Lagrange's formula,

$$\begin{aligned} p_n(f^*; x) &= \sum_{i=0}^{n} f_i^* \ell_i(x) = \sum_{i=0}^{n} f_i \ell_i(x) + \sum_{i=0}^{n} \varepsilon_i \ell_i(x) \\ &= p_n(f; x) + \sum_{i=0}^{n} \varepsilon_i \ell_i(x). \end{aligned}$$

Hence, for $a \leq x \leq b$,

$$\begin{aligned} |p_n(f^*; x) - p_n(f; x)| &= \left| \sum_{i=0}^{n} \varepsilon_i \ell_i(x) \right| \leq \sum_{i=0}^{n} |\varepsilon_i|\, |\ell_i(x)| \\ &\leq \varepsilon \sum_{i=0}^{n} |\ell_i(x)| = \varepsilon \lambda_n(x). \end{aligned}$$

(b) $\lambda_n(x_j) = \sum_{i=0}^{n} |\ell_i(x_j)| = \sum_{i=0}^{n} \delta_{ij} = \delta_{jj} = 1,\ j = 0, 1, \ldots, n,$ where δ_{ij} is the Kronecker delta.

(c) We may take the three points to be $x_i = x_0 + ih,\ i = 0, \pm 1$. Letting $x = x_0 + th$ yields

$$\ell_{-1}(x) = -\frac{1}{2}t(1-t), \quad \ell_0(x) = 1 - t^2, \quad \ell_1(x) = \frac{1}{2}t(1+t),$$

so that, for $|t| \leq 1$,

$$\lambda_2(x) = \sum_{i=-1}^{1} |\ell_i(x)| = \frac{1}{2}|t|(1-t) + 1 - t^2 + \frac{1}{2}|t|(1+t).$$

By symmetry, it suffices to consider $0 \le t \le 1$, in which case

$$\lambda_2(x) = 1 + t(1-t), \quad 0 \le t \le 1.$$

This takes on the maximum value at $t = \frac{1}{2}$, so that

$$\lambda_2(x) \le \frac{5}{4} \quad \text{for} \quad x_{-1} \le x \le x_1.$$

(d) We have

$$\ell_0(x) = \frac{(x-1)(x-p)}{(-1)(-p)} = -\frac{(x-1)(p-x)}{p},$$

$$\ell_1(x) = \frac{x(x-p)}{1\cdot(1-p)} = \frac{x(p-x)}{p-1}, \quad \ell_2(x) = \frac{x(x-1)}{p(p-1)}.$$

On $1 \le x \le p$, therefore, using the *Hint*,

$$\lambda_2(x) = \frac{(x-1)(p-x)}{p} + \frac{x(p-x)}{p-1} + \frac{x(x-1)}{p(p-1)} = 1 + c(x-1)(p-x),$$

for some constant c. Comparing coefficients of x^2 on both sides gives

$$-\frac{1}{p} - \frac{1}{p-1} + \frac{1}{p(p-1)} = -c, \quad c = \frac{2}{p}.$$

Therefore,

$$\lambda_2(x) = 1 + \frac{2}{p}(x-1)(p-x), \quad 1 \le x \le p,$$

and

$$\lambda_2'(x) = \frac{2}{p}(p+1-2x) = 0 \quad \text{if} \quad x = \frac{p+1}{2}.$$

Note that $\lambda_2'' < 0$, so we have a maximum. Thus, for $1 \le x \le p$,

$$\lambda_2(x) \le \lambda_2\left(\frac{p+1}{2}\right) = 1 + \frac{2}{p}\left(\frac{p-1}{2}\right)^2 = \frac{p^2+1}{2p} \sim \frac{1}{2}p \quad \text{as} \quad p \to \infty.$$

32. The error e_1 in linear interpolation between x_0 and $x_1 = x_0 + h$ satisfies (cf. Sect. 2.2.2(1))

$$|e_1| \le \frac{h^2}{8} \max_{x_0 \le x \le x_1} |J_0''(x)|.$$

Differentiating J_0 twice (under the sign of integration) gives

$$J_0''(x) = -\frac{1}{\pi}\int_0^\pi \cos(x\sin\theta)\sin^2\theta \, \mathrm{d}\theta.$$

Therefore, for arbitrary real x, using the *Point of information*,

$$|J_0''(x)| \le \frac{1}{\pi}\int_0^{\pi} \sin^2\theta \mathrm{d}\theta = \frac{2}{\pi}\int_0^{\pi/2} \sin^2\theta \mathrm{d}\theta = \frac{1}{2}\,.$$

Thus,

$$|e_1| \le \frac{h^2}{16} = \left(\frac{h}{4}\right)^2.$$

To guarantee an error $< 10^{-6}$, take h such that

$$\left(\frac{h}{4}\right)^2 < 10^{-6}, \quad h < 4\times 10^{-3}.$$

33. The error of quadratic interpolation is (cf. Sect. 2.2.2(2))

$$e_2(x) = \frac{f^{(3)}(\xi(x))}{6}\,(x-x_0)(x-x_1)(x-x_2),\ x_0 < \xi(x) < x_2.$$

Here, $f(x) = \ln x$, $x = 11.1$, $x_0 = 10$, $x_1 = 11$, $x_2 = 12$, giving

$$|e_2(11.1)| = \left|\frac{2}{6\cdot[\xi(x)]^3}\,(1.1)\times(.1)\times(-.9)\right| = \frac{.033}{[\xi(x)]^3}$$
$$< \frac{.033}{x_0^3} = 3.3\times 10^{-5}.$$

The relative error can thus be estimated by

$$\frac{|e_2(11.1)|}{\ln 11.1} < \frac{3.3\times 10^{-5}}{2.4069} = 1.37\ldots\times 10^{-5}.$$

34. (a) The error is bounded by (cf. Sect. 2.2.2(1))

$$|e_1| \le \frac{h^2}{8}\max_{x_0\le x\le x_1}|y''(x)|.$$

Since $y''(x) = xy(x)$, one has by the stated properties of the Airy function $0 < y''(x) < x_1 y_0$ for $x_0 \le x \le x_1$, so that

$$|e_1| < \frac{h^2}{8}\,x_1 y_0.$$

(b) The error is bounded by (cf. Ex. 43)

$$|e_2| \le \frac{h^3}{9\sqrt{3}}\max_{x_0\le x\le x_2}|y'''(x)|.$$

Since $y'''(x) = y(x) + xy'(x)$, one has by the stated properties of the Airy function $y_2 + x_2 y_0' < y'''(x) < y_0 + x_0 y_2'$ on $[x_0, x_2]$, so that

$$|e_2| < \frac{h^3}{9\sqrt{3}}\max(|y_2 + x_2 y_0'|, |y_0 + x_0 y_2'|).$$

35. Since $f(x) = x^{-1}$, $f''(x) = 2x^{-3}$ and $p_1(f;x) = (2-x)f(1) + (x-1)f(2) = 2 - x + \frac{1}{2}(x-1) = \frac{1}{2}(3-x)$, we have

$$f(x) - p_1(f;x) = \frac{1}{x} - \frac{1}{2}(3-x) = \frac{(x-1)(x-2)}{2x} = (x-1)(x-2)\,\frac{1}{\xi^3(x)}\,.$$

The last equality gives

$$\xi^3(x) = 2x, \quad \xi(x) = (2x)^{1/3}.$$

Furthermore,

$$\max_{1\le x\le 2} \xi(x) = 4^{1/3} = 1.5874\ldots\ ; \quad \min_{1\le x\le 2} \xi(x) = 2^{1/3} = 1.2599\ldots\ .$$

36. See the text.

37. Let $x_j \le x \le x_{j+1}$ for some j with $0 \le j < n$. Then

$$\omega_n(x) = (x-x_j)(x-x_{j+1}) \prod_{\substack{i\ne j\\ i\ne j+1}} (x-x_i).$$

Clearly,

$$|(x-x_j)(x-x_{j+1})| \le \left(\frac{x_{j+1}-x_j}{2}\right)^2 \le \frac{H^2}{4} \quad \text{for } x_j \le x \le x_{j+1}.$$

Furthermore,

$$|x-x_i| \le (j-i+1)H \quad \text{for } i < j,$$
$$|x-x_i| \le (i-j)H \quad \text{for } i > j+1.$$

Therefore,

$$\left| \prod_{\substack{i\ne j\\ i\ne j+1}} (x-x_i) \right| \le 2\cdot 3\cdots(j+1)\cdot 2\cdot 3\cdots(n-j)\cdot H^{n-1} = (j+1)!(n-j)!H^{n-1}.$$

If $\alpha_j = (j+1)!(n-j)!$, then $\alpha_0 = \alpha_{n-1} = n!$ and α_j first decreases and then increases as j varies from $j=0$ to $j=n-1$. Therefore, $\alpha_j \le n!$, and we get

$$\|\omega_n\|_\infty \le \frac{H^{n+1}}{4}\, n!.$$

38. Let $\overset{\circ}{T}_n(x) = x^n - t_{n-1}(x)$ be the monic Chebyshev polynomial of degree n. Since $\|\overset{\circ}{T}_n\|_\infty \le 2^{-(n-1)}$, we have

$$\|x^n - t_{n-1}(x)\|_\infty \le 2^{-(n-1)},$$

showing that $t_{n-1}(x)$ is a linear combination of $1, x, \ldots, x^{n-1}$ of the kind desired.

39. Let $x = \frac{b-a}{2}t + \frac{a+b}{2}$. Then

$$|a_0x^n + a_1x^{n-1} + \cdots + a_n| = \left|a_0\left(\frac{b-a}{2}\right)^n t^n + b_1t^{n-1} + \cdots + b_n\right|,$$

where, with $a_1, a_2, \ldots, a_n$, also $b_1, b_2, \ldots, b_n$ are arbitrary real. By factoring out $a_0(\frac{b-a}{2})^n$ and noting that $a \le x \le b$ is equivalent to $-1 \le t \le 1$, one obtains from Chebyshev's theorem (cf. Theorem 2.2.1) immediately that the minimax in question is

$$|a_0|\left(\frac{b-a}{2}\right)^n \frac{1}{2^{n-1}} = 2|a_0|\left(\frac{b-a}{4}\right)^n.$$

40. Let $y_k = y_k^{(n)}$ be the extrema of T_n:

$$T_n(y_k) = (-1)^k, \quad k = 0, 1, \ldots, n.$$

Then

$$|\hat{p}_n(y_k)| = \|\hat{p}_n\|_\infty, \quad k = 0, 1, \ldots, n.$$

Suppose $p_n \in \mathbb{P}_n^a$ does better than $\hat{p}_n$:

$$\|p_n\|_\infty < \|\hat{p}_n\|_\infty, \quad p_n(a) = 1.$$

Define $d_n(x) = \hat{p}_n(x) - p_n(x)$. Then, since $T_n(a) > 0$,

$$(-1)^k d_n(y_k) > 0, \quad k = 0, 1, \ldots, n.$$

It follows that d_n has at least n distinct zeros on $[-1, 1]$. Since there is an additional zero at $x = a$, we have $n+1$ distinct zeros for d_n—a polynomial of degree at most n. Therefore, d_n vanishes identically, which contradicts the inequalities above. Our assumption on p_n, therefore, cannot be maintained.

41. We have

$$f(x) - p_{n-1}(f; x) = \frac{f^{(n)}(\xi)}{n!}\prod_{\nu=1}^{n}(x - x_\nu) = \frac{f^{(n)}(\xi)}{n!}\frac{1}{2^{n-1}}T_n(x),$$

where $-1 \le x \le 1$ and $\xi = \xi(x)$ is between –1 and 1. Since $|T_n(x)| \le 1$ on $[-1, 1]$ and

$$f^{(n)}(x) = \int_5^\infty \mathrm{e}^{-t}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\frac{1}{t-x}\mathrm{d}t = \int_5^\infty \frac{n!\mathrm{e}^{-t}}{(t-x)^{n+1}}\mathrm{d}t,$$

we get for $-1 \le x \le 1$

$$\begin{aligned}|f(x) - p_{n-1}(f; x)| &\le \frac{1}{2^{n-1}}\int_5^\infty \frac{\mathrm{e}^{-t}}{(t-\xi)^{n+1}}\mathrm{d}t \le \frac{1}{2^{n-1}}\int_5^\infty \frac{\mathrm{e}^{-t}}{(t-1)^{n+1}}\mathrm{d}t \\ &\le \frac{1}{2^{n-1}}\frac{1}{4^{n+1}}\int_5^\infty \mathrm{e}^{-t}\mathrm{d}t = \frac{1}{2^{3n+1}}\left[-\mathrm{e}^{-t}\right]_5^\infty = \frac{\mathrm{e}^{-5}}{2^{3n+1}},\end{aligned}$$

so that

$$\max_{-1\le x\le 1} |f(x) - p_{n-1}(f;x)| \le \frac{\mathrm{e}^{-5}}{2^{3n+1}}\,.$$

42. (a) We have

$$f(x) - p_{n-1}(x) = \frac{f^{(n)}(\xi(x))}{n!}\prod_{i=1}^{n}(x - x_i),$$

where x_i are the Chebyshev points on $[a,b]$. Transforming to the canonical interval,

$$\begin{aligned} &x = a + \frac{b-a}{2}(t+1), \qquad -1 \le t \le 1;\\ &x_i = a + \frac{b-a}{2}(t_i+1), \quad t_i = \cos\left(\frac{2i-1}{2n}\pi\right), \quad i = 1,2,\dots,n, \end{aligned}$$

we get

$$\begin{aligned} \prod_{i=1}^{n}(x - x_i) &= \prod_{i=1}^{n}\left(\frac{b-a}{2}(t - t_i)\right) = \left(\frac{b-a}{2}\right)^n \frac{1}{2^{n-1}}T_n(t)\\ &= 2\left(\frac{b-a}{4}\right)^n T_n(t), \end{aligned}$$

where T_n is the Chebyshev polynomial of degree n. Therefore,

$$\max_{a\le x\le b}\left|\prod_{i=1}^{n}(x - x_i)\right| = 2\left(\frac{b-a}{4}\right)^n \max_{-1\le t\le 1}|T_n(t)| = 2\left(\frac{b-a}{4}\right)^n,$$

giving for the maximum relative error

$$r_n \le \frac{2M_n}{m_0 n!}\left(\frac{b-a}{4}\right)^n.$$

(b) For $f(x) = \ln x$ we have $f^{(n)}(x) = (-1)^{n-1}(n-1)!x^{-n},\ n \ge 1$, and so, on I_r,

$$\begin{aligned} m_0 = \ln \mathrm{e}^r = r, \quad M_n = (n-1)!\mathrm{e}^{-rn},\\ b - a = \mathrm{e}^{r+1} - \mathrm{e}^r = \mathrm{e}^r(\mathrm{e}-1). \end{aligned}$$

The result in (a), therefore, yields

$$r_n \le \frac{2(n-1)!\mathrm{e}^{-rn}[\mathrm{e}^r(\mathrm{e}-1)]^n}{r\cdot n!\cdot 4^n} = \frac{2}{rn}\left(\frac{\mathrm{e}-1}{4}\right)^n.$$

Thus,

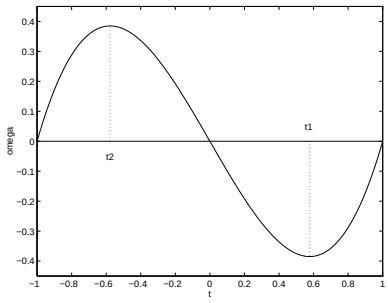
$$c = \frac{\mathrm{e}-1}{4} = .42957\dots\,, \quad \alpha(r,n) = \frac{2}{rn}\,.$$

(c) Since $\overline{x} = \mathrm{e}^{s-r}x$ for some $x \in I_r$, one gets $\ln\overline{x} = \ln x + s - r$.

43. (a) Exploiting symmetry, let $x = x_1 + th$, $-1 \le t \le 1$. Then

$$(x - x_0)(x - x_1)(x - x_2) = (1 + t)h \cdot th \cdot (t - 1)h = t(t^2 - 1)h^3.$$

The function $\omega(t) = t(t^2 - 1)$ has two extrema at



$t_{1,2} = \pm\sqrt{1/3}$, with values

$$|\,\omega(t_{1,2})| = \frac{1}{\sqrt{3}}\left(1 - \frac{1}{3}\right) = \frac{2}{3\sqrt{3}}\,.$$

Therefore,

$$|(x - x_0)(x - x_1)(x - x_2)| \le \frac{2}{3\sqrt{3}}\ h^3 \quad \text{for} \quad x_0 \le x \le x_2,$$

and

$$|f(x) - p_2(f;x)| = |(x - x_0)(x - x_1)(x - x_2)| \left|\frac{f'''(\xi)}{3!}\right|$$

$$\le \frac{2}{3\sqrt{3}}\ h^3 \frac{M_3}{3!} = \frac{M_3}{9\sqrt{3}}\ h^3.$$

(b) The Chebyshev points on $[x_0, x_2]$ are

$$\hat{x}_i = x_1 + \hat{t}_i h, \quad \hat{t}_i = \cos\ (2i + 1)\frac{\pi}{6}, \quad i = 0, 1, 2.$$

Letting (as in (a)) $x = x_1 + th$, we get

$$(x - \hat{x}_0)(x - \hat{x}_1)(x - \hat{x}_2) = (t - \hat{t}_0)h \cdot (t - \hat{t}_1)h \cdot (t - \hat{t}_2)h$$

$$= (t - \hat{t}_0)(t - \hat{t}_1)(t - \hat{t}_2)h^3 = \mathring{T}_3(t)h^3 = \frac{1}{4}\,T_3(t)h^3.$$

Therefore,

$$|(x - \hat{x}_0)(x - \hat{x}_1)(x - \hat{x}_2)| \le \frac{1}{4}\,h^3 \quad \text{for} \quad x_0 \le x \le x_2,$$

as compared to $\le \dfrac{2}{3\sqrt{3}}\,h^3 = .38490\ldots h^3$ in the equally spaced case.

44. (a) One has

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{k=0}^{n} (x - x_k), \quad -1 < \xi < 1.$$

Now, for $f(x) = \ln(2 + x)$, the derivatives are

$$f'(x) = \frac{1}{2+x}\,, \;\; f''(x) = -\frac{1}{(2+x)^2}\,, \ldots, f^{(n+1)}(x) = \frac{(-1)^n n!}{(2+x)^{n+1}}\,.$$

Therefore,

$$f(x) - p_n(x) = (-1)^n \frac{n!}{(2+\xi)^{n+1}(n+1)!} \prod_{k=0}^{n} (x - x_k)$$

$$= \frac{(-1)^n}{(2+\xi)^{n+1}(n+1)} \cdot \frac{1}{2^n} T_{n+1}(x).$$

There follows, since $\xi > -1$ and $|T_{n+1}(x)| \le 1$,

$$\|f - p_n\|_\infty \le \frac{1}{(n+1)2^n}\,.$$

(b) From Taylor's formula,

$$f(x) - t_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\, x^{n+1}, \quad |\xi| \in (0, |x|).$$

Hence, for $|x| \le 1$,

$$\|f - t_n\|_\infty \le \frac{n!|x|^{n+1}}{(2+\xi)^{n+1}(n+1)!} \le \frac{1}{n+1}\,.$$

The bound is larger by a factor of 2^n compared to the bound in (a). However, by using the full Taylor expansion of f, one gets

$$\|f - t_n\|_\infty \le \sum_{k=n+1}^{\infty} \left| \frac{f^{(k)}(0)}{k!} \right| = \sum_{k=n+1}^{\infty} \frac{(k-1)!}{k!2^k}$$
$$= \sum_{k=n+1}^{\infty} \frac{1}{k \cdot 2^k} \le \frac{1}{(n+1)2^n}\,,$$

which is the same bound as in (a).

45. We have, by Sect. 2.2.4, (2.99),

$$(T_i, T_j) = \int_{-1}^{1} T_i(t) T_j(t) \frac{\mathrm{d}t}{\sqrt{1-t^2}} = \begin{cases} 0 & \text{if } i \ne j, \\ \pi & \text{if } i = j = 0, \\ \frac{1}{2}\pi & \text{if } i = j > 0. \end{cases}$$

Furthermore, using the change of variables $t = \cos\theta$,

$$(T_i, f) = \int_{-1}^{1} f(t)T_i(t)\frac{\mathrm{d}t}{\sqrt{1-t^2}} = \int_{\pi}^{0} \theta \cdot \cos(i\theta)\frac{-\sin\theta}{\sin\theta}\mathrm{d}\theta$$

$$= \int_0^{\pi} \theta\cos(i\theta)\mathrm{d}\theta.$$

Integration by parts, if $i > 0$, gives

$$\int_0^{\pi} \theta\cos(i\theta)\mathrm{d}\theta = \theta\cdot\frac{1}{i}\sin(i\theta)\Big|_0^{\pi} - \frac{1}{i}\int_0^{\pi}\sin(i\theta)\mathrm{d}\theta$$

$$= -\frac{1}{i}\left[-\frac{1}{i}\cos(i\theta)\right]_0^{\pi} = -\frac{1}{i^2}(1-(-1)^i).$$

If $i = 0$, then $\int_{-1}^{1} f(t)T_0(t)\frac{\mathrm{d}t}{\sqrt{1-t^2}} = \int_0^{\pi}\theta\mathrm{d}\theta = \frac{1}{2}\pi^2$. Therefore,

$$(T_i, f) = \begin{cases} \frac{1}{2}\pi^2 & \text{if } i = 0, \\ 0 & \text{if } i > 0 \text{ is even}, \\ -2/i^2 & \text{if } i \text{ is odd}. \end{cases}$$

The normal equations $(T_j, T_j)c_j = (T_j, f),\ j = 0, 1, \ldots, n$, thus reduce to

$$\pi c_0 = \frac{1}{2}\pi^2, \quad \frac{1}{2}\pi c_j = \begin{cases} 0 & \text{if } j > 0 \text{ is even}, \\ -2/j^2 & \text{if } j \text{ is odd}. \end{cases}$$

Thus,

$$c_0 = \frac{1}{2}\pi, \quad c_{2j} = 0 \ \ (j > 0), \quad c_{2j+1} = -\frac{4}{\pi}\frac{1}{(2j+1)^2} \ \ (j \geq 0),$$

and

$$\hat{\varphi}_n(t) = \frac{1}{2}\pi - \frac{4}{\pi}\sum_{j=0}^{\lfloor\frac{n-1}{2}\rfloor}\frac{1}{(2j+1)^2}T_{2j+1}(t).$$

46. From the identity $T_n(\cos\theta) = \cos n\theta$, differentiating with respect to θ, one gets

$$T_n'(\cos\theta)(-\sin\theta) = -n\sin n\theta,$$

hence

$$T_n'(\cos\theta) = n\frac{\sin n\theta}{\sin\theta}.$$

Putting $\theta = \pi/2$ gives

$$T_n'(0) = n\,\frac{\sin(n\pi/2)}{\sin(\pi/2)} = n\sin(n\pi/2) = \begin{cases} 0, & n \text{ even}, \\ n(-1)^{\frac{n-1}{2}}, & n \text{ odd}. \end{cases}$$

47. See the text.

48. Put $x = \cos\theta$. Then

$$T_n(T_m(\cos\theta)) = T_n(\cos m\theta) = \cos(n \cdot m\theta) = T_{nm}(\cos\theta).$$

Therefore,

$$T_n(T_m(x)) = T_{nm}(x).$$

49. Since $|T_n(x)| \le 1$, all roots are in $[-1, 1]$. Therefore, set $x = \cos\theta,\ 0 \le \theta \le \pi$. Then

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

that is,

$$\sin\frac{1}{2}(n+1)\theta\,\sin\frac{1}{2}(n-1)\theta = 0.$$

This equation has the roots

$$\theta_k' = \frac{2k}{n+1}\,\pi, \qquad k = 0, 1, \ldots, \left\lfloor \frac{n+1}{2} \right\rfloor;$$

$$\theta_r'' = \frac{2r}{n-1}\,\pi, \qquad r = 1, 2, \ldots, \left\lfloor \frac{n-1}{2} \right\rfloor,$$

all contained in $[0, \pi]$. Hence, the original equation has the roots $x_k' = \cos\theta_k'$, $x_r'' = \cos\theta_r''$. If n is even, this accounts for exactly n distinct roots. Indeed, if we had $\theta_k' = \theta_r''$, that is, $\frac{2k}{n+1} = \frac{2r}{n-1}$, this would imply

$$(k-r)n = r + k.$$

Here, the right-hand side is between 1 and $n-1$, hence not divisible by n, whereas the left-hand side is. Contradiction! If n is odd, then $\theta'_{(n+1)/2} = \theta''_{(n-1)/2} = \pi$. Hence, one of these must be deleted; the remaining ones are again distinct, for the same reason as before.

50. Let $\sqrt{x} = \cos\theta$. Then

$$T_n(2x-1) = T_n(2\cos^2\theta - 1) = T_n(\cos 2\theta) = T_{2n}(\cos\theta) = T_{2n}(\sqrt{x}).$$

51. We have

$$f(0) - p_{2n-1}(0) = \frac{\prod_{k=1}^{n}(0-k^2)}{(2n)!}\, h^{2n} f^{(2n)}(\xi_n) = (-1)^n \frac{n!^2}{(2n)!}\, h^{2n} f^{(2n)}(\xi_n).$$

Since by assumption $|f^{(2n)}(\xi_n)| \le 1$ for all $n \ge 1$, we get, as $n \to \infty$, by Stirling's formula,

$$|f(0) - p_{2n-1}(0)| \le \frac{n!^2}{(2n)!}\, h^{2n} \sim \frac{2\pi n \cdot (n/\mathrm{e})^{2n}}{\sqrt{2\pi \cdot 2n}(2n/\mathrm{e})^{2n}}\, h^{2n} = \sqrt{\pi n}\left(\frac{h}{2}\right)^{2n}.$$

The bound on the far right clearly tends to zero as $n \to \infty$ if (and only if) $h < 2$. The convergence theory developed in Sect. 2.2.3 does *not* apply since the interpolation nodes $x_k = kh$, $k = \pm 1, \pm, 2, \ldots, \pm n$, do *not* remain in a bounded interval $[a, b]$ as $n \to \infty$ for fixed h.

52. (a) The Chebyshev points on $[a, b]$ are, by definition,

$$t_i^C = \frac{b-a}{2}\, x_i^C + \frac{b+a}{2}.$$

Therefore, letting $t = \frac{b-a}{2}\, x + \frac{b+a}{2}$, we get

$$\prod_{i=0}^{n}(t - t_i^C) = \prod_{i=0}^{n}\left[\frac{b-a}{2}\,(x - x_i^C)\right] = \left(\frac{b-a}{2}\right)^{n+1} \mathring{T}_{n+1}(x),$$

where $\mathring{T}_{n+1}$ is the monic Chebyshev polynomial of degree $n+1$. Taking the maximum modulus for $a \le t \le b$, hence for $-1 \le x \le 1$, we obtain

$$\max_{a \le t \le b}\left|\prod_{i=0}^{n}(t - t_i^C)\right| \le \left(\frac{b-a}{2}\right)^{n+1} 2^{-n} = 2\left(\frac{b-a}{4}\right)^{n+1}.$$

(b) One has, for arbitrary $t_i^{(n)} \in [a, b]$,

$$\begin{aligned} |f(t) - p_n(t)| &= \left|\frac{f^{(n+1)}(\tau)}{(n+1)!}\, \prod_{i=0}^{n}(t - t_i^{(n)})\right| \\ &\le \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \max_{a \le t \le b} \prod_{i=0}^{n} |t - t_i^{(n)}| \\ &\le \frac{(b-a)^{n+1}}{(n+1)!}\, \|f^{(n+1)}\|_\infty. \end{aligned}$$

If $f(t) = \ln t$, one has $f^{(n+1)}(t) = \dfrac{(-1)^n n!}{t^{n+1}}$, hence $\|f^{(n+1)}\|_\infty \le \dfrac{n!}{a^{n+1}}$, giving

$$\|f - p_n\|_\infty \le \left(\frac{b-a}{a}\right)^{n+1} \frac{1}{n+1}.$$

This tends to zero as $n \to \infty$ if $(b-a)/a < 1$, i.e., $b < 2a$.

(c) For Chebyshev points t_i^C, the analysis in (b), in view of (a), can be refined to give

$$\|f - p_n\|_\infty \le 2\left(\frac{b-a}{4}\right)^{n+1} \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \le 2\left(\frac{b-a}{4a}\right)^{n+1} \frac{1}{n+1},$$

which converges to zero if $(b-a)/4a < 1$, i.e., $b < 5a$.

53. (a) A solution, trivially, is

$$p(x) = \sum_{i=0}^{n} f_i\, \ell_i^2(x),$$

where $\ell_i \in \mathbb{P}_n$ are the elementary Lagrange interpolation polynomials for the nodes $x_0, x_1, \ldots, x_n$.

(b) For the data given in the *Hint*, we must have by Lagrange's interpolation formula

$$p(x) = \prod_{i=1}^{n}(x - x_i) \cdot q(x)$$

for some polynomial q. To ensure $p(x) \ge 0$, the polynomial q must change sign at each x_i, $i = 1, 2, \ldots, n$, so that $q(x) = \prod_{i=1}^{n}(x-x_i)\cdot r(x)$. Thus,

$$p(x) = \prod_{i=1}^{n}(x - x_i)^2 \cdot r(x),$$

and since $p(x_0) = 1$, the polynomial r does not vanish identically. It follows that $\deg p \ge 2n$.

54. Proof by induction on k: The assertion is true for $k = 0$. Assume it true for some k. Then

$$\begin{aligned}
\Delta^{k+1} f(x) &= \Delta(\Delta^k f(x)) = \Delta(k! h^k [x_0, x_1, \ldots, x_k] f) \\
&= k! h^k ([x_1, x_2, \ldots, x_{k+1}] f - [x_0, x_1, \ldots, x_k] f) \\
&= k! h^k (x_{k+1} - x_0)[x_0, x_1, \ldots, x_{k+1}] f \\
&= k! h^k \cdot (k+1) h [x_0, x_1, \ldots, x_{k+1}] f \\
&= (k+1)! h^{k+1} [x_0, x_1, \ldots, x_{k+1}] f,
\end{aligned}$$

as desired. For backward differences, $\nabla^k f(x) = k! h^k [x_0, x_{-1} \ldots, x_{-k}] f$, the proof is similar.

55. We have

$$[\,0, 1, 1, 1, 2, 2] f = \frac{f^{(5)}(\xi)}{5!}, \qquad 0 < \xi < 2,$$

for any $f \in C^5[0, 2]$, and therefore, in particular, for $f(x) = x^7$. For this function, the appropriate table of divided differences is

x	f					
0	0					
1	1	1				
1	1	7	6			
1	1	7	21	15		
2	128	127	120	99	42	
2	128	448	321	201	102	30

Therefore, from (2.117),

$$[\,0,1,1,1,2,2]f = \frac{7\cdot 6\cdot 5\cdot 4\cdot 3}{1\cdot 2\cdot 3\cdot 4\cdot 5}\,\xi^2 = 21\xi^2 = 30,$$

giving, uniquely (in the interval [0,2]),

$$\xi = \sqrt{\frac{30}{21}} = \sqrt{\frac{10}{7}} = 1.1952286\ldots\,.$$

56. (a) The assertion is true for $n = 0$. Assume it to be true for some n. Then

$$\begin{aligned}
[t, t+1, \ldots, t+n, t+n+1]f &= \frac{[t+1,\ldots,t+n+1]f - [t,\ldots,t+n]f}{(t+n+1)-t} \\
&= \frac{1}{n+1}\left\{\frac{(\mathrm{e}-1)^n}{n!}\mathrm{e}^{t+1} - \frac{(\mathrm{e}-1)^n}{n!}\mathrm{e}^t\right\} \text{ by induction assumption} \\
&= \frac{1}{(n+1)!}(\mathrm{e}-1)^n\mathrm{e}^t(\mathrm{e}-1) \\
&= \frac{(\mathrm{e}-1)^{n+1}}{(n+1)!}\,\mathrm{e}^t,
\end{aligned}$$

showing the validity of the assertion for $n+1$.

(b) Using the result in (a) for $t = 0$, we have

$$[0,1,\ldots,n]f = \frac{(\mathrm{e}-1)^n}{n!} = \frac{f^{(n)}(\xi)}{n!} = \frac{\mathrm{e}^\xi}{n!}\,.$$

There follows

$$\mathrm{e}^\xi = (\mathrm{e}-1)^n, \quad \xi = n\ln(\mathrm{e}-1) = .54132\ldots n.$$

We see that ξ is to the right of the midpoint $n/2$.

57. See the text.

58. By definition of the derivative, and of divided differences,

$$\begin{aligned}\frac{\partial}{\partial x_0}[x_0, x_1, \ldots, x_n]f &= \lim_{h\to 0} \frac{[x_0+h, x_1, \ldots, x_n]f - [x_0, x_1, \ldots, x_n]f}{h} \\ &= \lim_{h\to 0} \frac{[x_0+h, x_1, \ldots, x_n]f - [x_0, x_1, \ldots, x_n]f}{(x_0+h)-x_0} \\ &= \lim_{h\to 0} [x_0+h, x_0, x_1, \ldots, x_n]f \\ &= [x_0, x_0, x_1, \ldots, x_n]f.\end{aligned}$$

By the symmetry of divided differences, the analogous result holds for any of the other variables.

59. (a) Equate the leading coefficient in Newton's interpolation formula (2.111) with that of the Lagrange interpolation formula (2.52).

(b) We have, using the result of (a) and $g_j(x_j) = 0$,

$$\begin{aligned}[x_0, x_1, \ldots, x_n](fg_j) &= \sum_{\nu=0}^{n} \frac{f(x_\nu)g_j(x_\nu)}{\prod_{\substack{\mu=0\\ \mu\neq\nu}}^{n}(x_\nu - x_\mu)} \\ &= \sum_{\substack{\nu=0\\ \nu\neq j}}^{n} \frac{f(x_\nu)g_j(x_\nu)}{\prod_{\mu\neq\nu}(x_\nu - x_\mu)} \\ &= \sum_{\substack{\nu=0\\ \nu\neq j}}^{n} \frac{f(x_\nu)}{\prod_{\substack{\mu=0\\ \mu\neq\nu\\ \mu\neq j}}^{n}(x_\nu - x_\mu)} \\ &= [x_0, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n]f.\end{aligned}$$

60. The assertion is true (see Ex. 59(a)) for $m = 1$ and arbitrary n. Assume it to be true for some m and arbitrary n. Then, if we replace $\{x_1, \ldots, x_n\}$ by $\{t_0, t_1, \ldots, t_n\}$, adding one more point, we can write

$$\begin{aligned}[\underbrace{x_0, \ldots, x_0}_{m \text{ times}}, t_0, t_1, \ldots, t_n]f &= \frac{[\overbrace{x_0, \ldots, x_0}^{m \text{ times}}]f}{\prod_{\mu=0}^{n}(x_0 - t_\mu)} \\ &\quad - \sum_{\nu=0}^{n} \frac{[\overbrace{x_0, \ldots, x_0}^{(m-1) \text{ times}}, t_\nu]f}{(x_0 - t_\nu)\prod_{\substack{\mu=0\\ \mu\neq\nu}}^{n}(t_\nu - t_\mu)}.\end{aligned}$$

Now, use a partial fraction decomposition

$$\frac{1}{\prod_{\mu=0}^{n}(x_0-t_\mu)} = \sum_{\nu=0}^{n} \frac{1}{x_0-t_\nu} \; \frac{1}{\prod_{\substack{\mu=0\\ \mu\neq\nu}}^{n}(t_\nu-t_\mu)}$$

to obtain

$$[\underbrace{x_0,\ldots,x_0}_{m},t_0,t_1,\ldots t_n]f = \sum_{\nu=0}^{n} \frac{[\overbrace{x_0,\ldots,x_0}^{m \text{ times}}]f - [\overbrace{x_0,\ldots,x_0}^{(m-1) \text{ times}},t_\nu]f}{(x_0-t_\nu)\prod_{\substack{\mu=0\\ \mu\neq\nu}}^{n}(t_\nu-t_\mu)}$$

$$= \sum_{\nu=0}^{n} \frac{[\overbrace{x_0,\ldots,x_0}^{m \text{ times}},t_\nu]f}{\prod_{\substack{\mu=0\\ \mu\neq\nu}}^{n}(t_\nu-t_\mu)} \; .$$

Letting $t_0 \to x_0$ and $t_\nu = x_\nu$ for $\nu = 1,2,\ldots,n$, we get

$$[\underbrace{x_0,\ldots,x_0}_{(m+1) \text{ times}},x_1,\ldots,x_n]f = \frac{[\overbrace{x_0,\ldots,x_0}^{(m+1) \text{ times}}]f}{\prod_{\mu=1}^{n}(x_0-x_\mu)}$$

$$+\sum_{\nu=1}^{n} \frac{[\overbrace{x_0,\ldots,x_0}^{m \text{ times}},x_\nu]f}{\prod_{\substack{\mu=0\\ \mu\neq\nu}}^{n}(x_\nu-x_\mu)} \; ,$$

which is the assertion with m replaced by $m+1$.

61. (a) The computation of the divided differences for $n+1$ data points requires

$$2(n+(n-1)+\cdots+1) = 2\cdot\frac{1}{2}n(n+1) = n(n+1) \quad \text{additions,}$$

and

$$\frac{1}{2}n(n+1) \quad \text{divisions.}$$

(b) The auxiliary quantities $\lambda_0^{(n)},\ldots,\lambda_n^{(n)}$ in the barycentric formula require $\frac{1}{2}n(n+1)$ additions (the differences $x_i - x_k$ in the formula for $\lambda_i^{(k)}$ need

to be computed only once and can be re-used in the formula for $\lambda_k^{(k)}$), $\frac{1}{2}(n-1)n$ multiplications, and $\frac{1}{2}n(n+3)$ divisions, thus approximately the same number $\frac{3}{2}n^2 + O(n)$ of arithmetic operations as the table of divided differences.

(c) To evaluate the Newton polynomial

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots$$
$$+ a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

efficiently, given the coefficients a_k, one proceeds backwards as follows (in pseudo Matlab code):

```
s=a_n;
for k=n-1:-1:0
   s=a_k+(x-x_k)s;
end
```

This requires $2n$ additions and n multiplications, compared to $n+1+2n = 3n+1$ additions, $n+1$ multiplications, and $n+2$ divisions for the barycentric formula. These are only $O(n)$ arithmetic operations; neglecting them, both methods, according to (a) and (b), require about the same amount of work.

62. (a) From the table of divided differences

x	$f(x)$			
0	5			
1	3	–2		
3	5	1	1	
4	12	7	2	$\frac{1}{4}$

one obtains by Newton's formula

$$p_3(f;x) = 5 - 2x + x(x-1) + \tfrac{1}{4}\,x(x-1)(x-3),$$

or else,

$$p_3(f;x) = 5 - \tfrac{9}{4}\,x + \tfrac{1}{4}\,x^3.$$

(b) Setting the derivative of p_3 (instead of f) to zero gives

$$3x^2 = 9, \quad x_{\min} \approx \sqrt{3} = 1.73205\ldots\ .$$

This is indeed a minimum of p_3 since $p_3''(x) > 0$ for any positive x.

63. By induction on k, one finds that

$$[i, i+1, \ldots, i+k]f = \frac{a^k}{k!}\,f(i).$$

Therefore, using this for $i = 0$, one gets

$$p_n(f;x) = \sum_{k=0}^{n} [0,1,\ldots,k]f \prod_{j=0}^{k-1}(x-j)$$
$$= \sum_{k=0}^{n} \frac{a^k}{k!}\binom{x}{k} \cdot k! = \sum_{k=0}^{n} \binom{x}{k} a^k.$$

64. (a) The appropriate table of divided differences is

x	$f(x)$				
0	1				
1	2	1			
1	2	–1	–2		
3	0	–1	0	$\frac{2}{3}$	
3	0	0	$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{5}{36}$

The Hermite interpolant, therefore, is

$$p(x) = 1 + x - 2x(x-1) + \tfrac{2}{3}x(x-1)^2 - \tfrac{5}{36}x(x-1)^2(x-3),$$

which at $x = 2$ yields the following estimate of $f(2)$:

$$p(2) = 1 + 2 - 4 + \tfrac{2}{3} \cdot 2 - \tfrac{5}{36} \cdot 2(-1) = \tfrac{11}{18}\,.$$

(b) We have

$$|f(2) - p(2)| = \left|\frac{f^{(5)}(\xi)}{5!}(2-0)(2-1)^2(2-3)^2\right| \le \frac{2M}{120} = \frac{M}{60}\,.$$

65. (a) The table of divided differences is

x	p				
–1	0				
–1	0	0			
0	1	1	1		
1	0	–1	–1	–1	
1	0	0	1	1	1

By Newton's formula, therefore,

$$p(x) = (x+1)^2 - (x+1)^2x + (x+1)^2x(x-1)$$
$$= (x+1)^2(1 - x + x(x-1)) = (1+x)^2(1-x)^2$$
$$= (1-x^2)^2.$$

(b1) The error is $e(x) = (x+1)^2 x(x-1)^2 \dfrac{f^{(5)}(\xi)}{5!} = x(x^2-1)^2 \dfrac{f^{(5)}(\xi)}{5!}$, $-1 < \xi < 1$.

(b2) We have $f'(x) = 2\cos\left(\frac{\pi}{2}x\right)\cdot\left(-\frac{\pi}{2}\sin\frac{\pi}{2}x\right) = -\frac{\pi}{2}\sin\pi x$, hence $f^{(5)}(x) = -\frac{\pi}{2}\cdot\pi^4\sin\pi x = -\frac{\pi^5}{2}\sin\pi x$, so that

$$|e(x)| \le x(1-x^2)^2\,\frac{\pi^5}{2\cdot 5!}\,.$$

(b3) With $\omega(x) = x(1-x^2)^2$, we have $\omega'(x) = (1-x^2)(1-5x^2)$, which vanishes at $x = \pm 1/\sqrt{5}$, where $|\omega|$ has global maxima on [–1,1]. Since $\omega(1/\sqrt{5}) = 16/(25\sqrt{5})$, one gets

$$\max_{-1\le x\le 1}|e(x)| \le \frac{16}{25\sqrt{5}}\,\frac{\pi^5}{2\cdot 5!} = \frac{\pi^5}{375\sqrt{5}} = .0011925\ldots\pi^5 = .3649\ldots\ .$$

66. It is not a Lagrange interpolation problem because of the presence of derivatives, and, if $n = 2$ and $x_1 \ne x_0$ or $n > 2$, not a Hermite interpolation problem because of the absence of the values $p(x_i)$, $i = 1, 2, \ldots, n$.

Let $q(x) = p'(x)$. Then $q \in \mathbb{P}_{n-1}$ and $q(x_i) = f_i'$, $i = 1, 2, \ldots, n$, which, being a Lagrange interpolation problem, has a unique solution. The desired polynomial p is given by

$$p(x) = p(x_0) + \int_{x_0}^{x} p'(t)\mathrm{d}t = f_0 + \int_{x_0}^{x} q(t)\mathrm{d}t.$$

67. (a) We have

$$(\pi_0,\pi_0) = 1, \quad (\pi_0,\pi_1) = \tfrac{1}{2}\,, \quad (\pi_1,\pi_1) = \tfrac{1}{3}$$

and

$$(\pi_0, f) = \int_{\frac{1}{2}}^{1} \mathrm{d}t = \tfrac{1}{2}\,, \quad (\pi_1, f) = \int_{\frac{1}{2}}^{1} t\mathrm{d}t = \tfrac{1}{2}(1-\tfrac{1}{4}) = \tfrac{3}{8}\,.$$

The normal equations for $p_1(t) = c_0 + c_1\pi_1(t)$ are

$$\begin{aligned} c_0 + \tfrac{1}{2}c_1 &= \tfrac{1}{2}\,,\\ \tfrac{1}{2}c_0 + \tfrac{1}{3}c_1 &= \tfrac{3}{8}\,, \end{aligned}$$

giving $c_0 = -\frac{1}{4}\,,\ c_1 = \frac{3}{2}\,$. Thus,

$$\hat{p}_1(t) = -\tfrac{1}{4} + \tfrac{3}{2}t.$$

(b) No: Since $(B_0, f) = 0$, $(B_1, f) = (B_2, f) = \frac{1}{4}\,$, the normal equations for $s(t) = c_0B_0(t) + c_1B_1(t) + c_2B_2(t)$ are (cf. Ex. 13(d))

$$\begin{aligned} 2c_0 \quad +c_1 \quad\quad\ \ &= 0,\\ c_0 \quad +4c_1 \quad +c_2 \ &= 3,\\ \quad\quad c_1 \quad +2c_2 \ &= 3, \end{aligned}$$

The solution is $c_0 = -\frac{1}{4},\ c_1 = \frac{1}{2},\ c_2 = \frac{5}{4}$, giving

$$s(t) = -\tfrac{1}{4}B_0(t) + \tfrac{1}{2}B_1(t) + \tfrac{5}{4}B_2(t).$$

It so happens that this is a single linear function on $[0, 1]$ (not piecewise linear) and in fact the same linear function found in (a), $s(t) = \hat{p}_1(t)$.

68. Since all derivatives of order $\leq m$ of $s \in \mathbb{S}_m^m(\Delta)$ are continuous at each interior point t_i of Δ, the mth-degree Taylor polynomial of s centered at t_i gives the same polynomial on the left and on the right of t_i. Thus, s is a single polynomial of degree $\leq m$ on $[a, b]$.

69. We have

$$\begin{aligned}
&s(1-0) = 2 - 1 = 1 = s(1+0),\\
&s'(1-0) = -3 + 6 - 3 = 0 = s'(1+0),\\
&s''(1-0) = 6 - 6 = 0 = s''(1+0),\\
&s'''(1-0) = -6 \neq s'''(1-0),
\end{aligned}$$

so that $s \in C^2$ at $x = 1$. Similarly,

$$\begin{aligned}
&s(2+0) = 1 = s(2-0),\\
&s'(2+0) = 0 = s'(2-0),\\
&s''(2+0) = -\frac{3}{2} \neq s''(2-0),
\end{aligned}$$

so that $s \in C^1$ at $x = 2$. Since s is a cubic polynomial on each subinterval, we have that

$$s \in \mathbb{S}_3^1 \quad \text{(i.e., } m = 3, k = 1\text{)}.$$

70. Let

$$p(x) = a_1 x + a_2 x^2 + a_3 x^3.$$

Then we must have

$$\begin{aligned}
p(1) &= a_1 + a_2 + a_3 = 1,\\
p'(1) &= a_1 + 2a_2 + 3a_3 = -3,\\
p''(1) &= 2a_2 + 6a_3 = 6.
\end{aligned}$$

This gives

$$a_1 = 12, \quad a_2 = -18, \quad a_3 = 7.$$

Thus,

$$p(x) = 12x - 18x^2 + 7x^3.$$

Since $p''(0) = -36 \neq 0$, the spline s is *not* natural.

71. The number of parameters is: $(n-1)(m+1)$, the number of constraints (smoothness): $(n-2)(k+1)$, so that the dimension (degree of freedom) is:

$$(n-1)(m+1)-(n-2)(k+1) = n(m-k)+2k-m+1.$$

72. A basis of $\boldsymbol{S}$, in the notation of Sect. 2.3.2, is $B_2, B_3, \ldots, B_{n-1}$. Indeed, $s(x) = \sum_{i=2}^{n-1} c_i B_i(x)$ is in $\boldsymbol{S}$ for arbitrary constants c_i, and given any $s \in \boldsymbol{S}$, we have $s(x) = \sum_{i=2}^{n-1} s(x_i)B_i(x)$. Thus, $\operatorname{span}(B_2, \ldots, B_{n-1}) = \boldsymbol{S}$. Clearly, $B_2, \ldots, B_{n-1}$ are linearly independent, since $\sum_{i=2}^{n-1} c_i B_i(x) \equiv 0$, putting $x = x_j$, $j = 2, \ldots, n-1$, yields $c_j = 0$, $j = 2, 3, \ldots, n-1$. This proves that $B_2, B_3, \ldots, B_{n-1}$ is a basis of $\boldsymbol{S}$.

73. (a) The $3(n-1)$ parameters available are subject to $2(n-2)+2$ conditions of interpolation and $n-2$ conditions for continuity of the first derivative. The degree of freedom, therefore, is $3(n-1)-2(n-2)-2-(n-2) = 1$. Thus we need one additional condition to determine the spline uniquely.

(b) With the usual notation $\Delta x_i = x_{i+1} - x_i$, the appropriate table of divided differences, and the desired polynomial p_i, are

x	f		
x_i	f_i		
x_i	f_i	m_i	
x_{i+1}	f_{i+1}	$[x_i, x_{i+1}]f$	$\dfrac{1}{\Delta x_i}\left([x_i, x_{i+1}]f - m_i\right)$

$$p_i(x) = f_i + m_i(x-x_i) + \frac{1}{\Delta x_i}\left([x_i, x_{i+1}]f - m_i\right)(x-x_i)^2, \quad 1 \le i \le n-1.$$

(c) We want $p_i'(x_{i+1}) = m_{i+1}$, $i = 1, 2, \ldots, n-2$. Thus,

$$m_i + \frac{1}{\Delta x_i}\left([x_i, x_{i+1}]f - m_i\right) \cdot 2\Delta x_i = m_{i+1},$$
$$m_i + 2[x_i, x_{i+1}]f - 2m_i = m_{i+1},$$

that is,

$$\begin{cases} m_{i+1} = 2[x_i, x_{i+1}]f - m_i, & i = 1, 2, \ldots, n-2, \\ m_1 = f'(a). \end{cases}$$

74. The number of parameters at disposal is:

$$6(n-1),$$

the number of conditions imposed:

interpolation (which implies C^0)	$2(n-1)$,
C^k, $k = 1, 2, 3, 4$	$4(n-2)$.

Therefore, there are

$$6(n-1) - 2(n-1) - 4(n-2) = 4$$

additional conditions needed.

75. See the text

76. We must have

$$s_2(0) = s_1(0) = 1 + c; \quad s_2'(0) = s_1'(0) = 3c;$$
$$s_2''(0) = s_1''(0) = 6c; \quad s_2''(1) = 0.$$

This suggests

$$s_2(x) = 1 + c + 3cx + \frac{6c}{2!}x^2 + Ax^3.$$

The last condidion $[s_2''(1) = 0]$ gives $6c + 6A = 0$, hence $A = -c$. Thus,

$$s_2(x) = 1 + c + 3cx + 3cx^2 - cx^3.$$

The requirement $s(1) = -1$ is equivalent to $s_2(1) = -1$, that is,

$$1 + c + 3c + 3c - c = 1 + 6c = -1, \quad c = -\frac{1}{3}.$$

77. Let $a_{ij} = \int_a^b B_i(x)B_j(x)\mathrm{d}x$. Since $a_{ij} = 0$ if $|i-j| > 1$, we need only compute $a_{i,i-1}$, a_{ii}, $a_{i,i+1}$, $i = 1, 2, \ldots, n$, where $a_{i0} = a_{n,n+1} = 0$. For $1 < i < n$ we have (cf. Figure 2.11)

$$a_{i,i-1} = \int_{x_{i-1}}^{x_i} B_i(x)B_{i-1}(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{\Delta x_{i-1}} \, \frac{x_i - x}{\Delta x_{i-1}}\mathrm{d}x.$$

Here we conveniently change variables, $x = x_{i-1} + t\Delta x_{i-1}$, to get

$$a_{i,i-1} = \int_0^1 t(1-t) \cdot \Delta x_{i-1}\mathrm{d}t = \frac{1}{6}\,\Delta x_{i-1}.$$

Therefore also

$$a_{i,i+1} = a_{i+1,i} = \frac{1}{6}\,\Delta x_i.$$

Similarly,

$$\begin{aligned} a_{ii} &= \int_{x_{i-1}}^{x_i} \left(\frac{x - x_{i-1}}{\Delta x_{i-1}}\right)^2 \mathrm{d}x + \int_{x_i}^{x_{i+1}} \left(\frac{x_{i+1} - x}{\Delta x_i}\right)^2 \mathrm{d}x \\ &= \int_0^1 t^2 \cdot \Delta x_{i-1}\mathrm{d}t + \int_0^1 (1-t)^2 \cdot \Delta x_i \mathrm{d}t = \frac{1}{3}(\Delta x_{i-1} + \Delta x_i). \end{aligned}$$

Finally,

$$a_{11} = \int_{x_1}^{x_2} \left(\frac{x_2 - x}{\Delta x_1} \right)^2 \mathrm{d}x = \frac{1}{3}\,\Delta x_1, \quad a_{nn} = \int_{x_{n-1}}^{x_n} \left(\frac{x - x_{n-1}}{\Delta x_{n-1}} \right)^2 \mathrm{d}x = \frac{1}{3}\,\Delta x_{n-1}.$$

The right-hand sides are clearly

$$b_i = \int_{x_{i-1}}^{x_{i+1}} B_i(x) f(x) \mathrm{d}x, \qquad i = 1, 2, \ldots, n,$$

with the usual interpretation of $x_0 = x_1$ and $x_{n+1} = x_n$.

78. By Newton's interpolation formula, the polynomial in question is

$$p(x) = f_{i-1} + (x - x_{i-1})[x_{i-1}, x_i]f + (x - x_{i-1})(x - x_i)[x_{i-1}, x_i, x_{i+1}]f,$$

so that

$$\begin{aligned} m_i = p'(x_i) &= [x_{i-1}, x_i]f + (x_i - x_{i-1})[x_{i-1}, x_i, x_{i+1}]f \\ &= [x_{i-1}, x_i]f + \Delta x_{i-1} \cdot [x_{i-1}, x_i, x_{i+1}]f. \end{aligned}$$

79. See the text.

80. (a) We have, for any $g \in C[a, b]$,

$$\|s_1(g\,;\,\cdot)\|_\infty = \max_{1 \le i \le n} |s_1(g; x_i)| = \max_{1 \le i \le n} |g(x_i)| \le \|g\|_\infty.$$

(b) Using the additivity of $s_1(f\,;\,\cdot)$ with respect to f, we get for any $s \in \mathbb{S}_1^0(\Delta)$, since $s_1(s\,;\cdot) \equiv s$,

$$\begin{aligned} \|f - s_1(f\,;\,\cdot)\|_\infty &= \|f - s - s_1(f - s;\,\cdot)\|_\infty \\ &\le \|f - s\|_\infty + \|s_1(f - s;\,\cdot)\|_\infty \le 2\|f - s\|_\infty, \end{aligned}$$

where in the last step we have used (a) applied to $g = f - s$.

(c) Choose s in (b) so that $\| f - s \|_\infty$ is minimized over $\mathbb{S}_1^0(\Delta)$. Then

$$\|f - s_1(f;\,\cdot)\|_\infty \le 2\,\mathrm{dist}(f, \mathbb{S}_1^0(\Delta)),$$

i.e., the piecewise linear interpolant is nearly optimal (up to a factor of 2).

81. (a) Since f is an even function, so is $s = s_{\mathrm{nat}} \in \mathbb{S}_3^2(\Delta)$. In particular, s' is odd and s'' is even. It suffices to find $s|_{[0,1]}$ by the conditions $s(0) = 1$, $s'(0) = 0$, $s(1) = 0$, $s''(1) = 0$. Writing $s(x) = 1 + c_2 x^2 + c_3 x^3$ yields $1 + c_2 + c_3 = 0$, $2c_2 + 6c_3 = 0$, hence $c_2 = -\frac{3}{2}$, $c_3 = \frac{1}{2}$. Thus,

$$s(x) = 1 - \tfrac{3}{2}\,x^2 + \tfrac{1}{2}\,x^3, \quad s(-x) = s(x), \quad 0 \le x \le 1.$$

(b) We have

$$\begin{aligned}\int_{-1}^{1}[s''_{\rm nat}(x)]^2\mathrm{d}x &= 2\int_{0}^{1}[s''_{\rm nat}(x)]^2\mathrm{d}x\\ &= 2\int_{0}^{1}[3(x-1)]^2\mathrm{d}x = 18\int_{0}^{1}(x-1)^2\mathrm{d}x\\ &= 18\frac{(x-1)^3}{3}\Big|_0^1 = 18\cdot\frac{1}{3} = 6.\end{aligned}$$

On the other hand, for $p_2(f;-1,0,1;x) = 1-x^2$, one has

$$\int_{-1}^{1}[p_2''(f;x)]^2\mathrm{d}x = 2\int_{0}^{1}[-2]^2\mathrm{d}x = 8 > 6,$$

whereas for $g(x) = f(x)$ one gets

$$\begin{aligned}\int_{-1}^{1}[f''(x)]^2\mathrm{d}x &= 2\int_{0}^{1}\left[\left(\cos\frac{\pi}{2}x\right)''\right]^2\mathrm{d}x\\ &= 2\int_{0}^{1}\left[-\left(\frac{\pi}{2}\right)^2\cos\frac{\pi}{2}x\right]^2\mathrm{d}x\\ &= \frac{\pi^4}{8}\int_{0}^{1}\left[\cos\frac{\pi}{2}x\right]^2\mathrm{d}x = \frac{\pi^4}{8}\int_{0}^{\pi/2}\cos^2 t\cdot\frac{2}{\pi}\,\mathrm{d}t\\ &= \frac{\pi^3}{4}\int_{0}^{\pi/2}\cos^2 t\,\mathrm{d}t = \frac{\pi^4}{16} = 6.088068\ldots > 6,\end{aligned}$$

since $\int_0^{\pi/2}\cos^2 t\,\mathrm{d}t = \frac{\pi}{4}$.

(c) By symmetry of f, the complete interpolating spline $s = s_{\rm compl}$ must be even, hence $s \in C^3[-1,1]$, and therefore $s \in \mathbb{P}_3$ on $[-1,1]$. It follows that $p_3(f;-1,0,1,1;x) \equiv s_{\rm compl}(x)$, which illustrates the case of equality in Theorem 2.3.1. Since $f'(1) = -\frac{\pi}{2}$, one can obtain s by Hermite interpolation:

x	s			
0	1			
0	1	0		
1	0	-1	-1	
1	0	$-\frac{\pi}{2}$	$1-\frac{\pi}{2}$	$2-\frac{\pi}{2}$

$$s(x) = 1-x^2+\left(2-\frac{\pi}{2}\right)x^2(x-1) = 1-\left(3-\frac{\pi}{2}\right)x^2+\left(2-\frac{\pi}{2}\right)x^3 \quad \text{on} \quad 0\le x\le 1.$$

Therefore, $s''(x) = -2\left(3 - \frac{\pi}{2}\right) + 6\left(2 - \frac{\pi}{2}\right)x$, and

$$\int_{-1}^{1} [s''(x)]^2 \mathrm{d}x = 2\int_{0}^{1} [s''(x)]^2 \mathrm{d}x = 2\left[4\left(3 - \frac{\pi}{2}\right)^2 - 12\left(3 - \frac{\pi}{2}\right)\left(2 - \frac{\pi}{2}\right) + 12\left(2 - \frac{\pi}{2}\right)^2\right] = 8\left[3 - 3\frac{\pi}{2} + \frac{\pi^2}{4}\right] = 6.040096\ldots ,$$

which is indeed smaller than $\int_{-1}^{1}[f''(x)]^2\mathrm{d}x = 6.088\ldots$, but not by much.

ANSWERS TO MACHINE ASSIGNMENTS

1. (a) and (b) The problem of uniform best (discrete) approximation of a function g on [0,1] by a constant b is easily solved: if x_i are the discrete values of the variable, then $\max_i |g(x_i) - b|$ becomes a minimum precisely for

$$b = \frac{1}{2}\left(\min_i g(x_i) + \max_i g(x_i)\right).$$

Increasing or decreasing this value will indeed increase the maximum error. Solving the above problem with $g(x) = f(x) - ax$, using a discrete set of a-values, yields (approximately) the best uniform approximation of $f(x)$ by a linear function $ax + b$. A natural set of a-values consists of equally spaced values between $\min_{0\le x\le 1} f'(x)$ and $\max_{0\le x\le 1} f'(x)$. This is implemented in the program below and run for the seven functions suggested in (b).

```
 PROGRAM

%MAII_1AB
%
f0='%12.8f %12.8f %12.8f\n';
disp('     aopt         bopt         eopt')
n=500; m=300; eopt=1e20; i=0:n; x=i/n;
for ia=0:m
  a=1+(exp(1)-1)*ia/m;
%  a=-1+3*ia/(4*m);
%  a=pi*ia/(2*m);
%  a=2*ia/m;
%  a=3*ia/m;
%   a=4*ia/m;
%  a=5*ia/m;
  g=exp(x)-a*x;
%  g=1./(1+x)-a*x;
```

```
%  g=sin(pi*x/2)-a*x;
%  g=x.^2-a*x;
%  g=x.^3-a*x;
%  g=x.^4-a*x;
%  g=x.^5-a*x;
  gmin=min(g); gmax=max(g);
  b=(gmin+gmax)/2; e=(gmax-gmin)/2;
  if e<eopt
    aopt=a; bopt=b; eopt=e;
  end
end
fprintf(f0,aopt,bopt,eopt)


OUTPUT (for the seven functions f in (b))

>> MAII_1AB
   aopt         bopt         eopt
 1.71595076   0.89586226   0.10646881
-0.50000000   0.95710679   0.04289321
 1.00007366   0.10519917   0.10527283
 1.00000000  -0.12500000   0.12500000
 1.00000000  -0.19244972   0.19244972
 1.00000000  -0.23623520   0.23623520
 1.00000000  -0.26749530   0.26749530
 >>
```

What is interesting is that in all cases (at least approximately)

$$a = f(1) - f(0),$$

i.e., the slope of the linear approximation equals the slope of the secant of f through the endpoints. An explanation for this is given below.

(c) By the Principle of Alternation, we have, for three points $0 \le x_0 < x_1 < x_2 \le 1$ that

(1) $f_0 - (ax_0 + b) = \lambda,$

(2) $f_1 - (ax_1 + b) = -\lambda,$

(3) $f_2 - (ax_2 + b) = \lambda,$

where $f_i = f(x_i)$. Subtracting the first from the last equation gives

$$f_2 - f_0 - a(x_2 - x_0) = 0, \quad a = \frac{f_2 - f_0}{x_2 - x_0}\ .$$

In all our examples (as is often the case), x_0 and x_2 are the endpoints of [0,1], so that indeed $a = f(1) - f(0)$. With this assumed true,

the problem can be solved explicitly. One simply observes that the derivative of the error function $f(x) - (ax + b)$ must vanish at the interior point x_1,

$$f'(x_1) - a = 0. \tag{4}$$

Given that $x_0 = 0$, $x_2 = 1$, we now have four equations (1) – (4) in four unknowns, x_1, a, b, λ. From (4) and $a = f_2 - f_0$, we get

$$f'(x_1) = f_2 - f_0, \tag{5}$$

which can be solved for x_1. Adding (1) and (2), and solving for b gives

$$b = \tfrac{1}{2}(f_0 + f_1 - x_1(f_2 - f_0)). \tag{6}$$

Therefore, by (1), we get

$$\lambda = \tfrac{1}{2}(f_0 - f_1 + x_1(f_2 - f_0)). \tag{7}$$

Examples:

- $f(x) = \mathrm{e}^x$. Here, (5) gives $x_1 = \ln(\mathrm{e} - 1)$, and by (7),

$$|\lambda| = 1 - \frac{1}{2}\mathrm{e} + \frac{1}{2}(\mathrm{e} - 1)\ln(\mathrm{e} - 1) = .10593341\ldots\ .$$

- $f(x) = 1/(x + 1)$. Here, $x_1 = \sqrt{2} - 1$, and

$$|\lambda| = \frac{1}{2}\left(\frac{3}{2} - \sqrt{2}\right) = .04289321\ldots\ .$$

- $f(x) = \sin\left(\frac{\pi}{2}x\right)$. Here, (5) gives $x_1 = (2/\pi)\cos^{-1}(2/\pi)$, and by (7),

$$|\lambda| = \left|\frac{1}{2}\left(-\sqrt{1 - \left(\frac{2}{\pi}\right)^2} + \frac{2}{\pi}\cos^{-1}\frac{2}{\pi}\right)\right| = .10525683\ldots\ .$$

- $f(x) = x^\alpha$. Here, $x_1 = \left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}}$ and $|\lambda| = \left|\frac{\alpha-1}{2\alpha}\right|\left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}}$. The relevant table of $|\lambda|$ is given below.

α	$\lvert\lambda\rvert$
2	.12500000...
3	.19245008...
4	.23623519...
5	.26749612...

All values of the minimax error $|\lambda|$ derived here agree rather well with the numerical values (in the last column of the `OUTPUT`) found by our program.

2. (a) The matrix elements of the normal equations are

$$\begin{aligned} a_{ij} &= \int_0^1 B_i^n(t) B_j^n(t) \mathrm{d}t \\ &= \binom{n}{i}\binom{n}{j} \int_0^1 t^{i+j}(1-t)^{2n-i-j} \mathrm{d}t \\ &= \frac{n!}{i!(n-i)!}\,\frac{n!}{j!(n-j)!}\,\frac{(i+j)!(2n-i-j)!}{(2n+1)!}\,. \end{aligned}$$

(b)

```
 PROGRAM

%MAII_2B
%
f0='%6.0f %12.4e %12.4e\n';
disp('     n     error         cond')
for n=5:5:25
  A=zeros(n+1); e=ones(n+1,1); err=zeros(n+1,1);
  f=factorial(n)^2/factorial(2*n+1);
  for i=0:n
    for j=i:n
      A(i+1,j+1)=factorial(i+j)*factorial(2*n-i-j)*f/ ...
      (factorial(i)*factorial(n-i)*factorial(j) ...
      *factorial(n-j));
      if j>i, A(j+1,i+1)=A(i+1,j+1); end
    end
  end
  b=e/(n+1); c=A\b; err=norm(c-e,inf); cd=cond(A);
  fprintf(f0,n,err,cd)
end


 OUTPUT

 >> MAII_2B
  n     error         cond
  5   7.8826e-15   4.6200e+02
 10   1.9704e-11   3.5272e+05
 15   3.9159e-09   3.0054e+08
 20   8.5651e-06   2.6913e+11
 25   1.7482e-02   2.4927e+14
 >>
```

For $f(t) \equiv 1$, the right-hand vector $\boldsymbol{b}$ of the normal equations has

components (cf. *Hint*)

$$b_i = \int_0^1 B_i^n(t)\mathrm{d}t = \binom{n}{i} \int_0^1 t^i(1-t)^{n-i}\mathrm{d}t$$

$$= \frac{n!}{i!(n-i)!}\,\frac{i!(n-i)!}{(n+1)!} = \frac{1}{n+1}\,.$$

The solution of the normal equations $\boldsymbol{Ac} = \boldsymbol{b}$, in this case, is $c_j = 1$, all j, since $\sum_{j=0}^n B_j^n(t) \equiv 1$ (cf. Ex. 18(e)), and thus the least squares approximant is $\hat{\varphi}(t) \equiv 1$ ($\equiv f(t)$). The computer results show a gradual deterioration of accuracy caused by the worsening of the condition of the normal equations as n increases.

3. The normal equations have matrix elements

$$(\pi_i, \pi_j) = \sum_{k=1}^N t_k^{i+j}(1-t_k)^2, \quad i,j = 1,2,\ldots,n,$$

and right-hand sides

$$(\pi_i, f-t) = \sum_{k=1}^N t_k^i(1-t_k)\left(\sin(\tfrac{\pi}{2}t_k) - t_k\right), \quad i = 1,2,\ldots,n.$$

PROGRAM

```
%MAII_3
%
f0='%8.0f %12.4e %12.4e %12.4e %12.4e    N=%2.0f\n';
f1='%8.0f %12.4e %12.4e %12.4e %12.4e\n';
disp(['       n      cond         errc' ...
'         emin         emax'])
for N=[5 10 20 4]
  k=1:N; t=k/(N+1);
  for n=1:5
    Ad=zeros(n); bd=zeros(n,1);
    As=zeros(n,'single'); bs=zeros(n,1,'single');
    for i=1:n
      bd(i)=sum((1-t).*t.^i.*(sin(pi*t/2)-t));
      bs(i)=single(bd(i));
      for j=i:n
        Ad(i,j)=sum((1-t).^2.*t.^(i+j));
        As(i,j)=single(Ad(i,j));
        if j>i
          Ad(j,i)=Ad(i,j); As(j,i)=As(i,j);
        end
```

```
      end
    end
    cd=Ad\bd; cs=As\bs; condition=cond(Ad);
    errc=max(abs(cd-cs));
    e=cd(n);
    for j=n-1:-1:1
      e=t.*e+cd(j);
    end
    err=abs(t+t.*(1-t).*e-sin(pi*t/2));
    emin=min(err); emax=max(err);
    if n==1
      fprintf(f0,n,condition,errc,emin,emax,N)
    else
      fprintf(f1,n,condition,errc,emin,emax)
    end
  end
  fprintf('\n')
end
```

OUTPUT

```
>> MAII_3
       n      cond         errc         emin         emax
       1   1.0000e+00   3.5144e-08   1.5354e-03   2.2054e-02    N= 5
       2   4.4930e+01   1.4126e-06   1.7890e-04   1.9168e-03
       3   1.8299e+03   1.5461e-05   4.8546e-06   1.0729e-04
       4   8.9667e+04   3.1329e-04   1.4564e-06   4.8546e-06
Warning: Matrix is close to singular or badly scaled.
         Results may be inaccurate. RCOND = 6.459382e-08.
> In MAII_3new at 23
       5   7.7433e+06   1.3181e-02   4.3299e-15   2.3759e-14

       1   1.0000e+00   4.1935e-08   3.5366e-03   2.2397e-02    N=10
       2   4.4322e+01   1.3953e-07   4.8084e-04   2.0390e-03
       3   1.6622e+03   8.8779e-07   4.4644e-05   1.6535e-04
       4   6.0556e+04   2.2959e-04   7.7304e-07   9.1376e-06
       5   2.2733e+06   1.4950e-02   1.1406e-07   3.9254e-07

       1   1.0000e+00   5.6581e-10   1.1004e-03   2.2499e-02    N=20
       2   4.4268e+01   2.8157e-07   1.8954e-04   2.0642e-03
       3   1.6471e+03   7.5256e-06   2.0562e-05   1.6198e-04
       4   5.8238e+04   3.1635e-04   1.7055e-07   1.0988e-05
       5   2.0143e+06   1.2200e-02   1.0863e-08   6.3789e-07

       1   1.0000e+00   5.9213e-08   9.5863e-03   2.2564e-02    N= 4
```

```
       2   4.5652e+01   1.9080e-06   9.4474e-04   1.5867e-03
       3   2.0467e+03   4.5924e-05   4.2396e-05   8.4792e-05
       4   1.4657e+05   2.7802e-03   1.1102e-16   3.8858e-16
Warning: Matrix is close to singular or badly scaled.
         Results may be inaccurate. RCOND = 6.997197e-18.
> In MAII_3new at 22
Warning: Matrix is close to singular or badly scaled.
         Results may be inaccurate. RCOND = 2.341298e-09.
> In MAII_3new at 23
       5   1.4370e+17   1.7664e+00   1.1102e-16   6.6613e-16
>>
```

Comments

- As predicted by the condition numbers, the accuracy of the c deteriorates with increasing n. The maximum errors of the approximations keep decreasing because they are computed in double precision.
- When $n = N$, the maximum error should be zero since we are then interpolating. This is confirmed numerically in the case $N = 5$. (When $N = 10$, then the error for $n = 5$ is no longer zero.)
- Comparing the results for $N = 10$ and $N = 20$, one can see that the maximum errors reached a point of saturation: rather than smaller for $N = 20$, they are almost all slightly larger compared to $N = 10$. It seems that we reached the limiting case $N = \infty$, i.e., the case of *continuous* least squares approximation.
- In the case $N = 4$, the maximum errors for $1 \leq n \leq 3$ are practically the same as those in the case $N = 5$, except for $n = 5$. Here, the normal equations should be singular, which is correctly signaled by the routine. Interestingly, the maximum error of the approximation is still practically zero, suggesting a correlation of errors in the c that cancel out when φ_n is computed.

4. See the text.

5. (a) Since

$$\|f - \varphi\|_2^2 = \int_{\mathbb{R}} [f(t) - \varphi(t)]^2 \mathrm{d}\lambda_0(t) + \int_{\mathbb{R}} [f'(t) - \varphi'(t)]^2 \mathrm{d}\lambda_1(t),$$

by minimizing this quantity over a class of functions φ, one tries to approximate simultaneously f and its derivative f'.

(b) The matrix of the normal equations has elements

$$(\pi_i, \pi_j) = \int_0^2 t^{i+j-2} \mathrm{d}t + \lambda \cdot (i-1)(j-1) \int_0^2 t^{i+j-4} \mathrm{d}t,$$

where the second term is 0 if either $i = 1$ or $j = 1$ or both. Thus,

$$(\pi_i, \pi_j) = \frac{2^{i+j-1}}{i+j-1} \quad \text{if } i = 1 \text{ or } j = 1 \text{ or } i = j = 1,$$

and

$$(\pi_i, \pi_j) = 2^{i+j-1} \left\{ \frac{1}{i+j-1} + \frac{\lambda}{4} \frac{(i-1)(j-1)}{i+j-3} \right\} \quad \text{if } i > 1 \text{ and } j > 1.$$

The right-hand vector of the normal equations has elements

$$(\pi_i, f) = \int_0^2 t^{i-1} \mathrm{e}^{-t^2} \mathrm{d}t + \lambda \cdot (i-1) \int_0^2 t^{i-2} (-2t\mathrm{e}^{-t^2}) \mathrm{d}t,$$

where again the second term is 0 if $i = 1$. Using the change of variables $t^2 = x,\ 2t\mathrm{d}t = \mathrm{d}x$, one obtains

$$\begin{aligned}(\pi_i, f) &= \frac{1}{2} \int_0^4 x^{\frac{i-2}{2}} \mathrm{e}^{-x} \mathrm{d}x - \lambda \cdot (i-1) \int_0^4 x^{\frac{i-2}{2}} \mathrm{e}^{-x} \mathrm{d}x \\ &= \left(\frac{1}{2} - (i-1)\lambda \right) \gamma(\tfrac{i}{2}, 4), \quad i \geq 1.\end{aligned}$$

(c)

```
 PROGRAM

%MAII_5C
%
f0='%8.0f %12.4e %12.4e %12.4e   lambda=%1.1f\n';
f1='%8.0f %12.4e %12.4e %12.4e\n';
for lam=[0 .5 1 2]
  b(1)=.5*1.764162781524843;
  b(2)=(.5-lam)*.9816843611112658;
  b(3)=(.5-2*lam)*.8454501129849537;
  b(4)=(.5-3*lam)*.9084218055563291;
  b(5)=(.5-4*lam)*1.121650058367554;
  for i=1:5
    for j=1:5
      A(i,j)=1/(i+j-1);
      if i~=1 & j~=1
        A(i,j)=A(i,j)+.25*lam*(i-1)*(j-1)/(i+j-3);
      end
      A(i,j)=2^(i+j-1)*A(i,j);
    end
  end
  for n=2:5
```

```
    An=A(1:n,1:n);
    bn=b(1:n)';
    cd=cond(An);
    c=An\bn;
    p=c(n:-1:1);
    pd=(n-1:-1:1)'.*c(n:-1:2);
    x=(0:.02:2);
    y=polyval(p,x);
    yd=polyval(pd,x);
    e=abs(y-exp(-x.*x));
    ed=abs(yd+2*x.*exp(-x.*x));
    emax=max(e);
    edmax=max(ed);
    if n==2
      fprintf(f0,n,emax,edmax,cd,lam)
    else
      fprintf(f1,n,emax,edmax,cd)
    end
  end
  fprintf('\n')
end
```

```
 OUTPUT

 >> MAII_5C
   n      emax         edmax          cond
   2   1.6413e-01   5.8686e-01   1.4263e+01   lambda=0.0
   3   1.3701e-01   9.1418e-01   3.3624e+02
   4   4.3186e-02   4.0300e-01   1.0408e+04
   5   6.6937e-03   1.2996e-01   3.7710e+05

   2   1.0652e-01   5.2925e-01   7.5000e+00   lambda=0.5
   3   9.0779e-02   7.4807e-01   7.1594e+01
   4   3.6692e-02   2.8192e-01   1.1681e+03
   5   4.0083e-03   7.3378e-02   3.2203e+04

   2   1.1286e-01   5.1485e-01   6.1713e+00   lambda=1.0
   3   9.9882e-02   7.2688e-01   5.5220e+01
   4   3.7822e-02   2.7485e-01   9.4511e+02
   5   4.0542e-03   7.2610e-02   2.8007e+04

   2   1.2027e-01   5.0456e-01   5.8775e+00   lambda=2.0
   3   1.0654e-01   7.1288e-01   4.7966e+01
   4   3.8442e-02   2.7098e-01   8.3145e+02
   5   4.0814e-03   7.2215e-02   2.5864e+04
```

```
>>
```

It can be seen that increasing λ consistently decreases the error in the approximation for the derivative f', and, for $\lambda > .5$, slightly increases the error in the function f.

6. One easily shows (cf. Ex. 29(c)) that the relative maxima decrease moving from the ends of the interval $[0, n]$ toward the center. Therefore, by symmetry with respect to the midline $x = n/2$,

$$M_n = \max_{0<x<1} |\,\omega_n(x)|$$

and

$$m_n = \begin{cases} \max\limits_{n/2-1<x<n/2} |\,\omega_n(x)| & \text{if } n \text{ is even,} \\ \max\limits_{(n-1)/2<x<(n+1)/2} |\,\omega_n(x)| = |\,\omega_n(n/2)| & \text{if } n \text{ is odd.} \end{cases}$$

Here,

$$|\,\omega_n(n/2)| = 2^{-2n} \left(\frac{n!}{((n-1)/2)!} \right)^2 .$$

To calculate the (nontrivial) maxima, we note that

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln |\,\omega_n(x)| = \sum_{k=0}^{n} \frac{1}{x-k} =: f(x).$$

We must compute the zero ξ of $f(x)$ on $(0, 1)$ and, if n is even, the zero ξ' on $(n/2 - 1, n/2)$, to obtain

$$\begin{aligned} M_n &= |\,\omega_n(\xi)|, \\ m_n &= |\,\omega_n(\xi')| \quad \text{if } n \text{ is even.} \end{aligned}$$

For this, we may use Newton's method

$$\begin{aligned} x^{[\nu+1]} &= x^{[\nu]} - \frac{f(x^{[\nu]})}{f'(x^{[\nu]})} \\ &= x^{[\nu]} + \frac{\sum_{k=0}^{n} 1/(x^{[\nu]} - k)}{\sum_{k=0}^{n} 1/(x^{[\nu]} - k)^2} , \end{aligned}$$

using $x^{[0]} = 1/2$ for M_n and $x^{[0]} = (n-1)/2$ for m_n.

PROGRAM

```
%MAII_6
%
f0='%6.0f %20.12e %20.12e %20.12e %6.0f %6.0f\n';
disp(['  n          M_n                  m_n' ...
'               M_n/m_n        it  it1']);
for n=5:5:30
  k=0:n;
  x=0; x1=1/2; it=0;
  while abs(x-x1)>10^2*eps
    it=it+1;
    x=x1;
    x1=x+sum(1./(x-k))/sum(1./(x-k).^2);
  end
  M=abs(prod(x1-k));
  if 2*floor(n/2)==n
    x=0; x1=(n-1)/2; it1=0;
    while abs(x-x1)>10^2*eps
      it1=it1+1;
      x=x1;
      x1=x+sum(1./(x-k))/sum(1./(x-k).^2);
    end
    m=abs(prod(x1-k));
  else
    it1=0;
    m=(factorial(n)/(2^n*factorial((n-1)/2)))^2;
  end
  fprintf(f0,n,M,m,M/m,it,it1)
end
```

OUTPUT

```
>> MAII_6
  n         M_n                  m_n               M_n/m_n        it  it1
  5  1.690089432738e+01  3.515625000000e+00  4.807365497566e+00   6   0
 10  4.166144502892e+05  4.804865017268e+03  8.670679588124e+01   6   4
 15  1.348544769269e+11  6.269577561378e+07  2.150934023970e+03   7   0
 20  2.336679071562e+17  4.294534559061e+12  5.441053132596e+04   7   4
 25  1.413559513696e+24  9.313601380330e+17  1.517736755066e+06   7   0
 30  2.319528568743e+31  5.532808912085e+23  4.192316426611e+07   8   4
>>
```

7. See the text.

8. See the text.

EXERCISES AND MACHINE ASSIGNMENTS TO CHAPTER 3

EXERCISES

1. From Eqs. (3.4)–(3.6) with $n = 3$ we know that

   $$f'(x_0) = [x_0, x_1]f + (x_0 - x_1)[x_0, x_1, x_2]f + (x_0 - x_1)(x_0 - x_2)[x_0, x_1, x_2, x_3]f + e_3,$$

   where $e_3 = (x_0 - x_1)(x_0 - x_2)(x_0 - x_3)\frac{f^{(4)}(\xi)}{4!}$. Apply this to

   $$x_0 = 0, \quad x_1 = \tfrac{1}{8}, \quad x_2 = \tfrac{1}{4}, \quad x_3 = \tfrac{1}{2}$$

   and express $f'(0)$ as a linear combination of $f_k = f(x_k)$, $k = 0, 1, 2, 3$, and e_3. Also estimate the error e_3 in terms of $M_4 = \max_{0 \le x \le \frac{1}{2}} |f^{(4)}(x)|$.

2. Derive a formula for the error term $r_n'(x)$ of numerical differentiation analogous to (3.5) but for $x \neq x_0$. {*Hint*: use Ch. 2, (2.116) in combination with Ch. 2, Ex. 58.}

3. Let x_i, $i = 0, 1, \ldots, n$, be $n + 1$ distinct points with $H = \max_{1 \le i \le n} |x_i - x_0|$ small.

   (a) Show that for $k = 0, 1, \ldots, n$ one has

   $$\frac{\mathrm{d}^k}{\mathrm{d}x^k} \prod_{i=1}^{n} (x - x_i) \bigg|_{x = x_0} = O(H^{n-k}) \quad \text{as} \quad H \to 0.$$

   (b) Prove that

   $$f^{(n)}(x_0) = n!\, [x_0, x_1, \ldots, x_n]f + e_n,$$

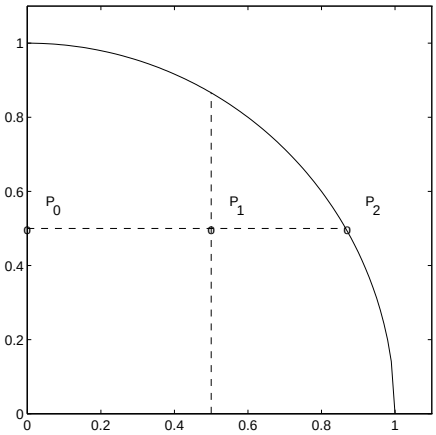
   where

   $$e_n = \begin{cases} O(H^2) & \text{if } \; x_0 = \frac{1}{n}\sum_{i=1}^{n} x_i, \\ O(H) & \text{otherwise,} \end{cases}$$

   assuming that f is sufficiently often (how often?) differentiable in the interval spanned by the x_i. {*Hint*: use the Newton interpolation formula with remainder, in combination with Leibniz's rule of differentiation.}

   (c) Specialize the formula in (b) to equally spaced points x_i with spacing h and express the result in terms of either the nth forward difference $\Delta^n f_0$ or the nth backward difference $\nabla^n f_n$ of the values $f_i = f(x_i)$. {Here $\Delta f_0 = f_1 - f_0$, $\Delta^2 f_0 = \Delta(\Delta f_0) = \Delta f_1 - \Delta f_0 = f_2 - 2f_1 + f_0$, etc., and similarly for ∇f_1, $\nabla^2 f_2$, and so on.}

4. Approximate $\partial u/\partial x|_{P_1}$ in terms of $u_0 = u(P_0)$, $u_1 = u(P_1)$, $u_2 = u(P_2)$ (see figure, where the curve represents a quarter arc of the unit circle). Estimate the error.



5. (a) Use the central difference quotient approximation $f'(x) \approx [f(x+h) - f(x-h)]/(2h)$ of the first derivative to obtain an approximation of $\dfrac{\partial^2 u}{\partial x \partial y}(x, y)$ for a function u of two variables.

   (b) Use Taylor expansion of a function of two variables to show that the error of the approximation derived in (a) is $O(h^2)$.

6. Consider the integral $I = \int_{-1}^{1} |x| \mathrm{d}x$, whose exact value is evidently 1. Suppose I is approximated (as it stands) by the composite trapezoidal rule $T(h)$ with $h = 2/n$, $n = 1, 2, 3, \ldots$.

   (a) Show (without any computation) that $T(2/n) = 1$ if n is even.

   (b) Determine $T(2/n)$ for n odd and comment on the speed of convergence.

7. Let

$$I(h) = \int_0^h f(x)\mathrm{d}x, \quad T(h) = \frac{h}{2}\,[f(0) + f(h)].$$

   (a) Evaluate $I(h)$, $T(h)$, and $E(h) = I(h) - T(h)$ explicitly for $f(x) = x^2 + x^{5/2}$.

   (b) Repeat for $f(x) = x^2 + x^{1/2}$. Explain the discrepancy that you will observe in the order of the error terms.

8. (a) Derive the "midpoint rule" of integration

$$\int_{x_k}^{x_k+h} f(x)\mathrm{d}x = hf(x_k + \tfrac{1}{2}h) + \frac{1}{24}h^3 f''(\xi), \quad x_k < \xi < x_k + h.$$

   {*Hint*: use Taylor's theorem centered at $x_k + \frac{1}{2}h$.}

   (b) Obtain the composite midpoint rule for $\int_a^b f(x)\mathrm{d}x$, including the error term, subdividing $[a, b]$ into n subintervals of length $h = \frac{b-a}{n}$.

9. (a) Show that the elementary Simpson's rule can be obtained as follows:

$$\int_{-1}^{1} y(t)\mathrm{d}t = \int_{-1}^{1} p_3(y; -1, 0, 0, 1; t)\mathrm{d}t + E^S(y).$$

   (b) Obtain a formula for the remainder $E^S(y)$, assuming $y \in C^4[-1, 1]$.

   (c) Using (a) and (b), derive the composite Simpson's rule for $\int_a^b f(x)\mathrm{d}x$, including the remainder term.

10. Let $E_n^S(f)$ be the remainder term of the composite Simpson's rule for $\int_0^{2\pi} f(x)\mathrm{d}x$ using n subintervals (n even). Evaluate $E_n^S(f)$ for $f(x) = \mathrm{e}^{imx}$ ($m = 0, 1, \dots$). Hence determine for what values of d Simpson's rule integrates exactly (on $[0, 2\pi]$) trigonometric polynomials of degree d.

11. Estimate the number of subintervals required to obtain $\int_0^1 \mathrm{e}^{-x^2}\mathrm{d}x$ to 6 correct decimal places (absolute error $\leq \frac{1}{2} \times 10^{-6}$)

   (a) by means of the composite trapezoidal rule,

   (b) by means of the composite Simpson's rule.

12. Let f be an arbitrary (continuous) function on [0,1] satisfying $f(x) + f(1 - x) \equiv 1$ for $0 \leq x \leq 1$.

   (a) Show that $\int_0^1 f(x)\mathrm{d}x = \frac{1}{2}$.

   (b) Show that the composite trapezoidal rule for computing $\int_0^1 f(x)\mathrm{d}x$ is exact.

   (c) Show, with as little computation as possible, that the composite Simpson's rule and more general symmetric rules are also exact.

13. (a) Construct a trapezoidal-like formula

$$\int_0^h f(x)\mathrm{d}x = af(0) + bf(h) + E(f), \quad 0 < h < \pi,$$

   which is exact for $f(x) = \cos x$ and $f(x) = \sin x$. Does this formula integrate constants exactly?

   (b) Show that a similar formula holds for $\int_c^{c+h} g(t)\mathrm{d}t$.

14. Given the subdivision Δ of $[0, 2\pi]$ into N equal subintervals, $0 = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = 2\pi$, $x_k = kh$, $h = 2\pi/N$, and a (2π)-periodic function f, construct a quadrature rule for the mth (complex) Fourier coefficients of f,

$$\frac{1}{2\pi}\int_0^{2\pi} f(x)\mathrm{e}^{-\mathrm{i}mx}\mathrm{d}x,$$

by approximating f by the spline interpolant $s_1(f;\cdot)$ from $\mathbb{S}_1^0(\Delta)$. Write the result in the form of a "modified" composite trapezoidal approximation. {*Hint*: express $s_1(f;\cdot)$ in terms of the hat functions defined in Chap. 2, (2.129).}

15. The composite trapezoidal rule for computing $\int_0^1 f(x)\mathrm{d}x$ can be generalized to subdivisions

$$\Delta: \quad 0 = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = 1$$

of the interval [0,1] in subintervals of arbitrary length $\Delta x_i = x_{i+1} - x_i$, $i = 0, 1, \ldots, n-1$, by approximating

$$\int_0^1 f(x)\mathrm{d}x \approx \int_0^1 s_1(f;x)\mathrm{d}x,$$

where $s_1(f;\cdot) \in \mathbb{S}_1^0(\Delta)$ is the piecewise linear continuous spline interpolating f at $x_0, x_1, \ldots, x_n$.

(a) Use the basis of hat functions $B_0, B_1, \ldots, B_n$ to represent $s_1(f;\cdot)$ and calculate $\int_0^1 s_1(f;x)\mathrm{d}x$.

(b) Discuss the error $E(f) = \int_0^1 f(x)\mathrm{d}x - \int_0^1 s_1(f;x)\mathrm{d}x$. In particular, find a formula of the type $E(f) = \text{const} \cdot f''(\xi)$, $0 < \xi < 1$, where the constant depends only on Δ.

16. (a) Construct the weighted Newton–Cotes formula

$$\int_0^1 f(x)x^\alpha \mathrm{d}x = a_0 f(0) + a_1 f(1) + E(f), \qquad \alpha > -1.$$

Explain why the formula obtained makes good sense.

(b) Derive an expression for the error term $E(f)$ in terms of an appropriate derivative of f.

(c) From the formulae in (a) and (b) derive an approximate integration formula for $\int_0^h g(t)t^\alpha \mathrm{d}t$ ($h > 0$ small), including an expression for the error term.

17. (a) Construct the weighted Newton–Cotes formula

$$\int_0^1 f(x) \cdot x\ln(1/x)\mathrm{d}x \approx a_0 f(0) + a_1 f(1).$$

{*Hint:* use $\int_0^1 x^r \ln(1/x)\mathrm{d}x = (r+1)^{-2}$, $r = 0, 1, 2, \ldots$.}

(b) Discuss how the formula in (a) can be used to approximate $\int_0^h g(t) \cdot t\ln(1/t)\mathrm{d}t$ for small $h > 0$. {*Hint*: make a change of variables.}

18. Let s be the function defined by

$$s(x) = \begin{cases} (x+1)^3 & \text{if } -1 \le x \le 0, \\ (1-x)^3 & \text{if } 0 \le x \le 1. \end{cases}$$

(a) With Δ denoting the subdivision of $[-1,1]$ into the two subintervals $[-1,0]$ and $[0,1]$, to what class $\mathbb{S}_m^k(\Delta)$ does the spline s belong?

(b) Estimate the error of the composite trapezoidal rule applied to $\int_{-1}^{1} s(x)\mathrm{d}x$, when $[-1,1]$ is divided into n subintervals of equal length $h = 2/n$ and n is even.

(c) What is the error of the composite Simpson's rule applied to $\int_{-1}^{1} s(x)\mathrm{d}x$, with the same subdivision of $[-1,1]$ as in (b)?

(d) What is the error resulting from applying the 2-point Gauss–Legendre rule to $\int_{-1}^{0} s(x)\mathrm{d}x$ and $\int_{0}^{1} s(x)\mathrm{d}x$ separately and summing?

19. (Gauss–Kronrod rule) Let $\pi_n(\,\cdot\,; w)$ be the (monic) orthogonal polynomial of degree n relative to a nonnegative weight function w on $[a,b]$, and $t_k^{(n)}$ its zeros. Use Theorem 3.2.1 to determine conditions on w_k, w_k^*, t_k^* for the quadrature rule

$$\int_a^b f(t)w(t)\mathrm{d}t = \sum_{k=1}^{n} w_k f(t_k^{(n)}) + \sum_{k=1}^{n+1} w_k^* f(t_k^*) + E_n(f)$$

to have degree of exactness at least $3n+1$; that is, $E_n(f) = 0$ for all $f \in \mathbb{P}_{3n+1}$.

20. (Turán quadrature formula) Let w be a nonnegative weight function on $[a,b]$. Prove: the quadrature formula

$$\int_a^b f(t)w(t)\mathrm{d}t = \sum_{k=1}^{n} [w_k f(t_k) + w_k' f'(t_k) + w_k'' f''(t_k)] + E_n(f)$$

has degree of exactness $d = 4n-1$ if and only if the following conditions are satisfied:

(a) The formula is (Hermite-) interpolatory; that is, $E_n(f) = 0$ if $f \in \mathbb{P}_{3n-1}$.

(b) The node polynomial $\omega_n(t) = \Pi_{k=1}^n (t - t_k)$ satisfies

$$\int_a^b [\omega_n(t)]^3 p(t)w(t)\mathrm{d}t = 0 \quad \text{for all } \ p \in \mathbb{P}_{n-1}.$$

{*Hint*: simulate the proof of Theorem 3.2.1.}

21. Consider $s > 1$ weight functions $w_\sigma(t)$, $\sigma = 1, 2, \ldots, s$, integers m_σ such that $\sum_{\sigma=1}^{s} m_\sigma = n$, and s quadrature rules

$$Q_\sigma: \quad \int_a^b f(t)w_\sigma(t)\mathrm{d}t = \sum_{k=1}^{n} w_{k,\sigma} f(t_k) + E_{n,\sigma}(f), \quad \sigma = 1, 2, \ldots, s,$$

which share n common nodes t_k but have individual weights $w_{k,\sigma}$. State necessary and sufficient conditions for Q_σ to have degree of exactness $n + m_\sigma - 1$, $\sigma = 1, 2, \ldots, s$, and explain why this is likely to be optimal.

22. Consider a quadrature formula of the form

$$\int_0^1 f(x)\mathrm{d}x \approx a_0 f(0) + a_1 f'(0) + \sum_{k=1}^{n} w_k f(x_k) + b_0 f(1).$$

(a) Call the formula "Hermite-interpolatory" if the right-hand side is obtained by integrating on the left instead of f the (Hermite) interpolation polynomial p satisfying

$$p(0) = f(0), \quad p'(0) = f'(0), \quad p(1) = f(1),$$
$$p(x_k) = f(x_k), \quad k = 1, 2, \ldots, n.$$

What degree of exactness does the formula have in this case (regardless of how the nodes x_k are chosen, as long as they are mutually distinct and strictly inside the interval $[0, 1]$) ?

(b) What is the maximum degree of exactness expected to be if all coefficients and nodes x_k are allowed to be freely chosen?

(c) Show that for the maximum degree of exactness to be achieved, it is necessary that $\{x_k\}$ are the zeros of the polynomial π_n of degree n which is orthogonal on $[0, 1]$ with respect to the weight function $w(x) = x^2(1 - x)$. Identify this polynomial in terms of one of the classical orthogonal polynomials.

(d) Show that the choice of the x_k in (c) together with the requirement of the quadrature formula to be Hermite-interpolatory, is sufficient for the maximum degree of exactness to be attained.

23. Show that the Gauss–Radau as well as the Gauss–Lobatto formulae are positive if the weight function w is nonnegative and not identically zero. {*Hint*: modify the proof given for the Gauss formula in Sect. 3.2.3(b).} What are the implications with regard to convergence as $n \to \infty$ of the formulae?

24. (Fejér, 1933). Let t_k, $k = 1, 2, \ldots, n$, be the zeros of

$$\omega_n(t) = P_n(t) + \alpha P_{n-1}(t) + \beta P_{n-2}(t), \quad n \geq 2,$$

where $\{P_k\}$ are the Legendre polynomials, and assume $\alpha \in \mathbb{R}$, $\beta \leq 0$, and the zeros t_k real and pairwise distinct. Show that the Newton–Cotes formula

$$\int_{-1}^{1} f(t)\mathrm{d}t = \sum_{k=1}^{n} w_k f(t_k) + E_n(f), \quad E_n(\mathbb{P}_{n-1}) = 0,$$

has all weights positive: $w_k > 0$ for $k = 1, 2, \ldots, n$. {*Hint*: define $\Delta_k(t) = [\ell_k(t)]^2 - \ell_k(t)$ and show that $\int_{-1}^{1} \Delta_k(t)\mathrm{d}t \leq 0$.}

25. (a) Determine by Newton's interpolation formula the quadratic polynomial p interpolating f at $x = 0$ and $x = 1$ and f' at $x = 0$. Also express the error in terms of an appropriate derivative (assumed continuous on [0,1]).

    (b) Based on the result of (a), derive an integration formula of the type

    $$\int_0^1 f(x)\mathrm{d}x = a_0 f(0) + a_1 f(1) + b_0 f'(0) + E(f).$$

    Determine a_0, a_1, b_0 and an appropriate expression for $E(f)$.

    (c) Transform the result of (b) to obtain an integration rule, with remainder, for $\int_c^{c+h} y(t)\mathrm{d}t$, where $h > 0$. {Do not rederive this rule from scratch.}

26. Imitate the procedures used in the Example of Ch. 2, Sect. 2.1.4(2) for monic Legendre polynomials to show orthogonality on $[0, \infty)$ relative to the Laguerre measure $\mathrm{d}\lambda(t) = t^\alpha \mathrm{e}^{-t}\mathrm{d}t$, $\alpha > -1$, of the (monic) polynomials

    $$\pi_k(t) = (-1)^k t^{-\alpha}\mathrm{e}^t \frac{\mathrm{d}^k}{\mathrm{d}t^k}(t^{\alpha+k}\mathrm{e}^{-t}), \quad k = 0, 1, 2, \dots ,$$

    and to derive explicit formulae for the recursion coefficients α_k, β_k. {*Hint*: express α_k and β_k in terms of the coefficients λ_k, μ_k in $\pi_k(t) = t^k + \lambda_k t^{k-1} + \mu_k t^{k-2} + \cdots$.}

27. Show that

    $$\pi_k(t) = \frac{(-1)^k}{2^k}\mathrm{e}^{t^2}\frac{\mathrm{d}^k}{\mathrm{d}t^k}(\mathrm{e}^{-t^2}), \quad k = 0, 1, 2, \dots ,$$

    are the monic orthogonal polynomials on $\mathbb{R}$ relative to the Hermite measure $\mathrm{d}\lambda(t) = \mathrm{e}^{-t^2}\mathrm{d}t$. Use this "Rodrigues formula" directly to derive the recurrence relation for the (monic) Hermite polynomials.

28. (a) Construct the quadratic (monic) polynomial $\pi_2(\,\cdot\,; w)$ orthogonal on $(0, \infty)$ with respect to the weight function $w(t) = \mathrm{e}^{-t}$. {*Hint*: use $\int_0^\infty t^m \mathrm{e}^{-t}\mathrm{d}t = m!$.}

    (b) Obtain the two-point Gauss–Laguerre quadrature formula,

    $$\int_0^\infty f(t)\mathrm{e}^{-t}\mathrm{d}t = w_1 f(t_1) + w_2 f(t_2) + E_2(f),$$

    including a representation for the remainder $E_2(f)$.

    (c) Apply the formula in (b) to approximate $I = \int_0^\infty \mathrm{e}^{-t}\mathrm{d}t/(t+1)$. Use the remainder term $E_2(f)$ to estimate the error, and compare your estimate with the true error {use $I = .596347361\ldots$}. Knowing the true error, identify the unknown quantity $\xi > 0$ contained in the error term $E_2(f)$.

29. Derive the 2-point Gauss–Hermite quadrature formula,

$$\int_{-\infty}^{\infty} f(t)\mathrm{e}^{-t^2}\,\mathrm{d}t = w_1 f(t_1) + w_2 f(t_2) + E_2(f),$$

including an expression for the remainder $E_2(f)$. {*Hint*: use $\int_0^\infty t^{2n}\mathrm{e}^{-t^2}\,\mathrm{d}t = \frac{(2n)!}{n!2^{2n}}\frac{\sqrt{\pi}}{2}$, $n = 0, 1, 2, \ldots$.}

30. Let $\pi_n(\,\cdot\,; w)$ be the nth-degree orthogonal polynomial with respect to the weight function w on $[a, b]$, $t_1, t_2, \ldots, t_n$ its n zeros, and $w_1, w_2, \ldots, w_n$ the n Gauss weights.

    (a) Assuming $n > 1$, show that the n polynomials $\pi_0, \pi_1, \ldots, \pi_{n-1}$ are also orthogonal with respect to the *discrete* inner product $(u, v) = \sum_{\nu=1}^{n} w_\nu u(t_\nu)v(t_\nu)$.

    (b) With $\ell_i(t) = \prod_{k\neq i}[(t - t_k)/(t_i - t_k)]$, $i = 1, 2, \ldots, n$, denoting the elementary Lagrange interpolation polynomials associated with the nodes $t_1, t_2, \ldots, t_n$, show that

$$\int_a^b \ell_i(t)\ell_k(t)w(t)\mathrm{d}t = 0 \quad \text{if} \quad i \neq k.$$

31. Consider a quadrature formula of the type

$$\int_0^\infty \mathrm{e}^{-x} f(x)\mathrm{d}x = af(0) + bf(c) + E(f).$$

    (a) Find a, b, c such that the formula has degree of exactness $d = 2$. Can you identify the formula so obtained? {*Point of information*: $\int_0^\infty \mathrm{e}^{-x}x^r\mathrm{d}x = r!$}

    (b) Let $p_2(x) = p_2(f; 0, 2, 2; x)$ be the Hermite interpolation polynomial interpolating f at the (simple) point $x = 0$ and the double point $x = 2$. Determine $\int_0^\infty \mathrm{e}^{-x}p_2(x)\mathrm{d}x$ and compare with the result in (a).

    (c) Obtain the remainder $E(f)$ in the form $E(f) = \text{const} \cdot f'''(\xi)$, $\xi > 0$.

32. In this problem, $\pi_j(\,\cdot\,; w)$ denotes the monic polynomial of degree j orthogonal on the interval $[a, b]$ relative to a weight function $w \geq 0$.

    (a) Show that $\pi_n(\,\cdot\,; w)$, $n > 0$, has at least one real zero in the interior of $[a, b]$ at which π_n changes sign.

    (b) Prove that all zeros of $\pi_n(\,\cdot\,; w)$ are real, simple, and contained in the interior of $[a, b]$. {*Hint*: put $r_0 = \max\{r \geq 1$: $t_{k_1}^{(n)}, t_{k_2}^{(n)}, \ldots, t_{k_r}^{(n)}$ are distinct real zeros of π_n in (a, b) at each of which π_n changes sign$\}$. Show that $r_0 = n$.}

33. Prove that the zeros of $\pi_n(\,\cdot\,; w)$ interlace with those of $\pi_{n+1}(\,\cdot\,; w)$.

34. Consider the Hermite interpolation problem: Find $p \in \mathbb{P}_{2n-1}$ such that

$$(*) \qquad p(\tau_\nu) = f_\nu, \quad p'(\tau_\nu) = f'_\nu, \quad \nu = 1, 2, \ldots, n.$$

There are "elementary Hermite interpolation polynomials" h_ν, k_ν such that the solution of (*) can be expressed (in analogy to Lagrange's formula) in the form

$$p(t) = \sum_{\nu=1}^{n} [h_\nu(t) f_\nu + k_\nu(t) f'_\nu].$$

(a) Seek h_ν and k_ν in the form

$$h_\nu(t) = (a_\nu + b_\nu t)\ell_\nu^2(t), \quad k_\nu(t) = (c_\nu + d_\nu t) l_\nu^2(t),$$

where ℓ_ν are the elementary Lagrange polynomials. Determine the constants a_ν, b_ν, c_ν, d_ν.

(b) Obtain the quadrature rule

$$\int_a^b f(t)w(t)\mathrm{d}t = \sum_{\nu=1}^{n} [\lambda_\nu f(\tau_\nu) + \mu_\nu f'(\tau_\nu)] + E_n(f)$$

with the property that $E_n(f) = 0$ for all $f \in \mathbb{P}_{2n-1}$.

(c) What conditions on the node polynomial $\omega_n(t) = \prod_{\nu=1}^{n}(t - \tau_\nu)$ (or on the nodes τ_ν) must be imposed in order that $\mu_\nu = 0$ for $\nu = 1, 2, \ldots, n$?

35. Show that $\int_0^1 (1-t)^{-1/2} f(t)\mathrm{d}t$, when f is smooth, can be computed accurately by Gauss–Legendre quadrature. {*Hint*: substitute $1 - t = x^2$.}

36. The Gaussian quadrature rule for the (Chebyshev) weight function $w(t) = (1 - t^2)^{-1/2}$ is known to be

$$\int_{-1}^{1} f(t)(1 - t^2)^{-1/2}\mathrm{d}t \approx \frac{\pi}{n} \sum_{k=1}^{n} f(t_k^C), \qquad t_k^C = \cos\left(\frac{2k-1}{2n}\pi\right).$$

(The nodes t_k^C are the n Chebyshev points.) Use this fact to show that the unit disk has area π.

37. Assuming f is a well-behaved function, discuss how the following integrals can be approximated by *standard* Gauss-type rules (i.e., with canonical intervals and weight functions).

(a) $\int_a^b f(x)\mathrm{d}x \quad (a < b)$.

(b) $\int_1^\infty \mathrm{e}^{-ax} f(x)\mathrm{d}x \quad (a > 0)$.

(c) $\int_{-\infty}^{\infty} \mathrm{e}^{-(ax^2+bx)} f(x)\mathrm{d}x \quad (a > 0)$. {*Hint*: complete the square.}

(d) $\int_0^\infty \frac{\mathrm{e}^{-xt}}{y+t}\mathrm{d}t$, $x>0$, $y>0$. Is the approximation you get for the integral too small or too large? Explain.

38. (a) Let $w(t)$ be an even weight function on $[a,b]$, $a<b$, $a+b=0$, i.e., $w(-t)=w(t)$ on $[a,b]$. Show that $(-1)^n\pi_n(-t;w)\equiv\pi_n(t;w)$, i.e., the (monic) nth-degree orthogonal polynomial relative to the weight function w is even [odd] for n even [odd].

(b) Show that the Gauss formula

$$\int_a^b f(t)w(t)\mathrm{d}t=\sum_{\nu=1}^n w_\nu f(t_\nu)+E_n(f)$$

for an even weight function w is symmetric, i.e.,

$$t_{n+1-\nu}=-t_\nu,\quad w_{n+1-\nu}=w_\nu,\quad \nu=1,2,\dots,n.$$

(c) Let w be the "hat function"

$$w(t)=\begin{cases}1+t & \text{if } -1\le t\le 0,\\ 1-t & \text{if } 0\le t\le 1.\end{cases}$$

Obtain the 2-point Gaussian quadrature formula $\int_{-1}^1 f(t)w(t)\mathrm{d}t = w_1f(t_1)+w_2f(t_2)+E_2(f)$ for this weight function w, including an expression for the error term under a suitable regularity assumption on f. {*Hint*: use (a) and (b) to simplify the calculations.}

39. Let $t_k^{(2n)}$ be th nodes, ordered monotonically, of the $(2n)$-point Gauss–Legendre quadrature rule and $w_k^{(2n)}$ the associated weights. Show that, for any $p\in\mathbb{P}_{2n-1}$, one has

$$\int_0^1 t^{-1/2}p(t)\mathrm{d}t=2\sum_{k=1}^n w_k^{(2n)}p([t_k^{(2n)}]^2).$$

40. Let f be a smooth function on $[0,\pi]$. Explain how best to evaluate

$$I_{\alpha,\beta}(f)=\int_0^\pi f(\theta)[\cos\tfrac{1}{2}\theta]^\alpha[\sin\tfrac{1}{2}\theta]^\beta\mathrm{d}\theta,\quad \alpha>-1,\quad \beta>-1.$$

41. Let $Q_nf, Q_{n^*}f$ be n-point, resp. n^*-point quadrature rules for $If=\int_a^b f(t)w(t)\mathrm{d}t$ and $Q_{n^*}f$ at least twice as accurate as Q_nf, i.e.,

$$|Q_{n^*}f-If|\le\tfrac{1}{2}|Q_nf-If|.$$

Show that the error of $Q_{n^*}f$ then satisfies

$$|Q_{n^*}f-If|\le|Q_nf-Q_{n^*}f|.$$

42. Given a nonnegative weight function w on $[-1,1]$ and $x>1$, let

$$(G) \qquad \int_{-1}^{1} f(t)\frac{w(t)}{x^2-t^2}\mathrm{d}t = \sum_{k=1}^{n} w_k^G f(t_k^G) + E_n^G(f)$$

be the n-point Gaussian quadrature formula for the weight function $\frac{w(t)}{x^2-t^2}$. (Note that $t_k^G,\ w_k^G$ both depend on n and x.) Consider the quadrature rule

$$\int_{-1}^{1} g(t)w(t)\mathrm{d}t = \sum_{k=1}^{n} w_k g(t_k) + E_n(g),$$

where $t_k = t_k^G,\ w_k = [x^2-(t_k^G)^2]w_k^G$. Prove:

(a) $E_n(g)=0$ if $g(t)=\frac{1}{t\pm x}$.

(b) If $n\geq 2$, then $E_n(g)=0$ whenever g is a polynomial of degree $\leq 2n-3$.

43. Let ξ_ν, $\nu=1,2,\ldots,2n$, be $2n$ preassigned distinct numbers satisfying $-1<\xi_\nu<1$, and let w be a positive weight function on [–1,1]. Define $\omega_{2n}(x)=\prod_{\nu=1}^{2n}(1+\xi_\nu x)$. (Note that ω_{2n} is positive on [–1,1].) Let $x_k^G,\ w_k^G$ be the nodes and weights of the n-point Gauss formula for the weight function $w^*(x)=\frac{w(x)}{\omega_{2n}(x)}$:

$$\int_{-1}^{1} p(x)w^*(x)\mathrm{d}x = \sum_{k=1}^{n} w_k^G p(x_k^G), \quad p\in\mathbb{P}_{2n-1}.$$

Define $x_k^* = x_k^G,\ w_k^* = w_k^G\omega_{2n}(x_k^G)$. Show that the quadrature formula

$$\int_{-1}^{1} f(x)w(x)\mathrm{d}x = \sum_{k=1}^{n} w_k^* f(x_k^*) + E_n^*(f)$$

is exact for the $2n$ rational functions

$$f(x)=\frac{1}{1+\xi_\nu x}, \quad \nu=1,2,\ldots,2n.$$

44. (a) Prove (3.46).

(b) Prove (3.47). {*Hint*: use the Christoffel–Darboux formula of Chap. 2, Ex. 21(b).}

45. Prove (3.50). {*Hint*: prove, more generally, $(I^{p+1}g)(s)=\int_0^s \frac{(s-t)^p}{p!}g(t)\mathrm{d}t$.}

46. (a) Use the method of undetermined coefficients to obtain an integration rule (having degree of exactness $d=2$) of the form

$$\int_0^1 y(s)\mathrm{d}s \approx ay(0)+by(1)-c[y'(1)-y'(0)].$$

(b) Transform the rule in (a) into one appropriate for approximating $\int_x^{x+h} f(t)\mathrm{d}t$.

(c) Obtain a composite integration rule based on the formula in (b) for approximating $\int_a^b f(t)\mathrm{d}t$. Interpret the result.

47. Determine the quadrature formula of the type

$$\int_{-1}^{1} f(t)\mathrm{d}t = \alpha_{-1}\int_{-1}^{-1/2} f(t)\mathrm{d}t + \alpha_0 f(0) + \alpha_1 \int_{1/2}^{1} f(t)\mathrm{d}t + E(f)$$

having maximum degree of exactness d. What is the value of d?

48. (a) Determine the quadratic spline $s_2(x)$ on $[-1,1]$ with a single knot at $x=0$ and such that $s_2(x)\equiv 0$ on $[-1,0]$ and $s_2(1)=1$.

(b) Consider a function $s(x)$ of the form

$$s(x) = c_0 + c_1 x + c_2 x^2 + c_3 s_2(x), \quad c_i = \text{const},$$

where $s_2(x)$ is as defined in (a). What kind of function is s? Determine s such that

$$s(-1) = f_{-1}, \quad s(0) = f_0, \quad s'(0) = f_0', \quad s(1) = f_1,$$

where $f_{-1} = f(-1),\ f_0 = f(0),\ f_0' = f'(0),\ f_1 = f(1)$ for some function f on $[-1,1]$.

(c) What quadrature rule does one obtain if one approximates $\int_{-1}^{1} f(x)\mathrm{d}x$ by $\int_{-1}^{1} s(x)\mathrm{d}x$, with s as obtained in (b)?

49. Prove that the condition (3.68) does not depend on the choice of the basis $\varphi_1, \varphi_2, \ldots, \varphi_n$.

50. Let E be a linear functional that annihilates all polynomials of degree $d \geq 0$. Show that the Peano kernel $K_r(t)$, $r \leq d$, of E vanishes for $t \notin [a,b]$, where $[a,b]$ is the interval of function values referenced by E.

51. Show that a linear functional E satisfying $Ef = e_{r+1} f^{(r+1)}(\overline{t})$, $\overline{t} \in [a,b]$, $e_{r+1} \neq 0$, for any $f \in C^{r+1}[a,b]$, is necessarily definite of order r if it has a continuous Peano kernel K_r.

52. Let E be a linear functional that annihilates all polynomials of degree d. Show that none of the Peano kernels $K_0, K_1, \ldots, K_{d-1}$ of E can be definite.

53. Suppose in (3.61) the function f is known to be only once continuously differentiable, i.e., $f \in C^1[0,1]$.

(a) Derive the appropriate Peano representation of the error functional Ef.

(b) Obtain an estimate of the form $|Ef| \leq c_0 \|f'\|_\infty$.

54. Assume, in Simpson's rule

$$\int_{-1}^{1} f(x)\mathrm{d}x = \frac{1}{3}[f(-1) + 4f(0) + f(1)] + E^S(f),$$

that f is only of class $C^2[-1,1]$ instead of class $C^4[-1,1]$ as normally assumed.

(a) Find an error estimate of the type

$$|E^S(f)| \leq \text{const} \cdot \|f''\|_\infty, \quad \|f''\|_\infty = \max_{-1\leq x\leq 1} |f''(x)|.$$

{*Hint*: apply the appropriate Peano representation of $E^S(f)$.}

(b) Transform the result in (a) to obtain Simpson's formula, with remainder estimate, for the integral

$$\int_{c-h}^{c+h} g(t)\mathrm{d}t, \quad g \in C^2[c-h, c+h], \quad h > 0.$$

(c) How does the estimate in (a) compare with the analogous error estimate for two applications of the trapezoidal rule,

$$\int_{-1}^{1} f(x)\mathrm{d}x = \frac{1}{2}\,[f(-1) + 2f(0) + f(1)] + E_2^T(f)?$$

55. Determine the Peano kernel $K_1(t)$ on $[a,b]$ of the error functional for the composite trapezoidal rule over the interval $[a,b]$ subdivided into n subintervals of equal length.

56. Consider the trapezoidal formula "with mean values,"

$$\int_0^1 f(x)\mathrm{d}x = \frac{1}{2}\left[\frac{1}{\varepsilon}\int_0^\varepsilon f(x)\mathrm{d}x + \frac{1}{\varepsilon}\int_{1-\varepsilon}^1 f(x)\mathrm{d}x\right] + E(f), \quad 0 < \varepsilon < \frac{1}{2}\,.$$

(a) Determine the degree of exactness of this formula.

(b) Express the remainder $E(f)$ by means of its Peano kernel K_1 in terms of f'', assuming $f \in C^2[0,1]$.

(c) Show that the Peano kernel K_1 is definite, and thus express the remainder in the form $E(f) = e_2 f''(\tau)$, $0 < \tau < 1$.

(d) Consider (and explain) the limit cases $\varepsilon \downarrow 0$ and $\varepsilon \to \frac{1}{2}$.

57. (a) Use the method of undetermined coefficients to construct a quadrature formula of the type

$$\int_0^1 f(x)\mathrm{d}x = af(0) + bf(1) + cf''(\gamma) + E(f)$$

having maximum degree of exactness d, the variables being a, b, c, and γ.

(b) Show that the Peano kernel K_d of the error functional E of the formula obtained in (a) is definite, and hence express the remainder in the form $E(f) = e_{d+1} f^{(d+1)}(\xi)$, $0 < \xi < 1$.

58. (a) Use the method of undetermined coefficients to construct a quadrature formula of the type

$$\int_0^1 f(x)\mathrm{d}x = -\alpha f'(0) + \beta f(\tfrac{1}{2}) + \alpha f'(1) + E(f)$$

that has maximum degree of exactness.

(b) What is the precise degree of exactness of the formula obtained in (a)?

(c) Use the Peano kernel of the error functional E to express $E(f)$ in terms of the appropriate derivative of f reflecting the result of (b).

(d) Transform the formula in (a) to one that is appropriate to evaluate $\int_c^{c+h} g(t)\mathrm{d}t$, and then obtain the corresponding composite formula for $\int_a^b g(t)\mathrm{d}t$, using n subintervals of equal length, and derive an error term. Interpret your result.

59. Consider a quadrature rule of the form

$$\int_0^1 x^\alpha f(x)\mathrm{d}x \approx Af(0) + B\int_0^1 f(x)\mathrm{d}x, \quad \alpha > -1, \quad \alpha \neq 0.$$

(a) Determine A and B such that the formula has degree of exactness $d = 1$.

(b) Let $E(f)$ be the error functional of the rule determined in (a). Show that the Peano kernel $K_1(t) = E_{(x)}((x-t)_+)$ of E is positive definite if $\alpha > 0$, and negative definite if $\alpha < 0$.

(c) Based on the result of (b), determine the constant e_2 in $E(f) = e_2 f''(\xi)$, $0 < \xi < 1$.

60. (a) Consider a quadrature formula of the type

$$(^*) \qquad \int_0^1 f(x)\mathrm{d}x = \alpha f(x_1) + \beta[f(1) - f(0)] + E(f)$$

and determine α, β, x_1 such that the degree of exactness is as large as possible. What is the maximum degree attainable?

(b) Use interpolation theory to obtain a bound on $|E(f)|$ in terms of $\|f^{(r)}\|_\infty = \max_{0\le x\le 1} |f^{(r)}(x)|$ for some suitable r.

(c) Adapt (*), including the bound on $|E(f)|$, to an integral of the form $\int_c^{c+h} f(t)\mathrm{d}t$, where c is some constant and $h > 0$.

(d) Apply the result of (c) to develop a composite quadrature rule for $\int_a^b f(t)\mathrm{d}t$ by subdividing $[a, b]$ into n subintervals of equal length $h = \frac{b-a}{n}$. Find a bound for the total error.

61. Construct a quadrature rule

$$\int_0^1 x^\alpha f(x)\mathrm{d}x \approx a_1 \int_0^1 f(x)\mathrm{d}x + a_2 \int_0^1 x f(x)\mathrm{d}x, \quad 0 < \alpha < 1,$$

(a) which is exact for all polynomials p of degree ≤ 1;

(b) which is exact for all $f(x) = x^{1/2}p(x)$, $p \in \mathbb{P}_1$.

62. Let

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b, \quad x_k = a + kh, \quad h = \frac{b-a}{n},$$

be a subdivision of $[a, b]$ into n equal subintervals.

(a) Derive an elementary quadrature formula for the integral $\int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x$, including a remainder term, by approximating f by the cubic Hermite interpolation polynomial $p_3(f;\, x_k,\, x_k,\, x_{k+1},\, x_{k+1};\, x)$ and then integrating over $[x_k,\, x_{k+1}]$. Interpret the result.

(b) Develop the formula obtained in (a) into a composite quadrature rule, with remainder term, for the integral $\int_a^b f(x)\mathrm{d}x$.

63. (a) Given a function $g(x, y)$ on the unit square $0 \leq x \leq 1$, $0 \leq y \leq 1$, determine a "bilinear polynomial" $p(x, y) = a + bx + cy + dxy$ such that p has the same values as g at the four corners of the square.

(b) Use (a) to obtain a cubature formula for $\int_0^1 \int_0^1 g(x, y)\mathrm{d}x\mathrm{d}y$ that involves the values of g at the four corners of the unit square. What rule does this reduce to if g is a function of x only (i.e., does not depend on y)?

(c) Use (b) to find a "composite cubature rule" for $\int_0^1 \int_0^1 g(x, y)\mathrm{d}x\mathrm{d}y$ involving the values $g_{i,j} = g(ih, jh)$, $i, j = 0, 1, \ldots, n$, where $h = 1/n$.

64. (a) Let $d_1(h) = (f(h) - f(0))/h$, $h > 0$, be the difference quotient of f at the origin. Describe how the extrapolation method based on a suitable expansion of $d_1(h)$ can be used to approximate $f'(0)$ to successively higher accuracy.

(b) Develop a similar method for calculating $f''(0)$, based on $d_2(h) = [f(h) - 2f(0) + f(-h)]/h^2$.

MACHINE ASSIGNMENTS

1. Let $f(x) = \frac{1}{1-\pi x}$ and $f_i = f(ih)$, $i = -2, -1, 0, 1, 2$. In terms of the four backward differences

$$\nabla f_1 = f_1 - f_0, \quad \nabla^2 f_1 = f_1 - 2f_0 + f_{-1},$$
$$\nabla^3 f_2 = f_2 - 3f_1 + 3f_0 - f_{-1}, \quad \nabla^4 f_2 = f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2},$$

define

$$e_n(h) = f^{(n)}(0) - \frac{1}{h^n}\,\nabla^n f_{\lfloor \frac{n+1}{2} \rfloor}, \quad n = 1, 2, 3, 4.$$

Try to determine the order of convergence of $e_n(h)$ as $h \to 0$ by printing, for $n = 1, \ldots, 4$,

$$e_n(h_k) \;\text{ and }\; r_k := \frac{e_n(h_k)}{e_n(h_{k-1})}, \quad k = 1, 2, \ldots, 10,$$

where $h_k = \frac{1}{4} \cdot 2^{-k}$, $k \geq 0$. Comment on the results.

2. Let

$$f(z) = \tan z, \qquad |z| < \tfrac{1}{2}\pi.$$

(a) Express

$$y_m(\theta) = \mathrm{Re}\,\{\mathrm{e}^{-\mathrm{i}m\theta} f(x_0 + r\mathrm{e}^{\mathrm{i}\theta})\}, \quad 0 < r < \tfrac{1}{2}\pi, \quad m = 1, 2, 3, \ldots$$

explicitly as a function of θ. {*Hint*: use Euler's identities $\sin z = (\mathrm{e}^{\mathrm{i}z} - \mathrm{e}^{-\mathrm{i}z})/(2\mathrm{i})$, $\cos z = (\mathrm{e}^{\mathrm{i}z} + \mathrm{e}^{-\mathrm{i}z})/2$, valid for arbitrary complex z.}

(b) Obtain the analogue to (3.19) for the mth derivative and thus write $f^{(m)}(x_0)$ as a definite integral over $[0, 2\pi]$.

(c) Use Matlab to compute $f^{(m)}(0)$ for $m = 1 : 5$ using the integral in (b) in conjunction with the composite trapezoidal rule (cf. Sect. 3.2.1) relative to a subdivision of $[0, 2\pi]$ into n subintervals. Use $r = \frac{1}{12}\rho\pi$, $\rho = 1, 2, 3, 4, 5$, and $n = 5 : 5 : 50$. For each m print a table whose columns contain r, n, the trapezoidal approximation $t_n^{(m)}$, and the (absolute) error, in this order. Comment on the results; in particular, try to explain the convergence behavior as r increases and the difference in behavior for n even and n odd. {*Hint*: prepare plots of the integrand; you may use the Matlab routine `spline` for cubic spline interpolation to do this.}

(d) Do the same as (c), but for $f^{(m)}(\frac{7}{16}\pi)$ and $r = \frac{1}{32}\pi$.

(e) Write and run a Matlab program for approximating $f^{(m)}(0)$, $m = 1 : 5$, by central difference formulae with steps $h = 1, \frac{1}{5}, \frac{1}{25}, \frac{1}{125}, \frac{1}{625}$. Comment on the results.

(f) Do the same as (e), but for $f^{(m)}(\frac{7}{16}\pi)$ and $h = \frac{1}{32}\pi, \frac{1}{160}\pi, \frac{1}{800}\pi, \frac{1}{4000}\pi$.

3. Given n distinct real nodes $x_k = x_k^{(n)}$, the interpolatory quadrature rule

$$(\mathrm{WNC}_n) \qquad \int_a^b f(x)w(x)\mathrm{d}x = \sum_{k=1}^{n} w_k^{(n)} f(x_k^{(n)}), \quad \text{all } f \in \mathbb{P}_{n-1},$$

is called a weighted (by the weight function w) Newton–Cotes formula (cf. Sect. 3.2.2). The weights $w_k^{(n)}$ can be generated by n_g-point Gauss integration,

$n_g = \lfloor (n+1)/2 \rfloor$, of the elementary Lagrange interpolation polynomials (see (3.39)),

$$(\mathrm{W}_n) \qquad w_k^{(n)} = \int_a^b \ell_k(x) w(x) \mathrm{d}x, \quad \ell_k(x) = \prod_{\substack{\ell=1 \\ \ell \neq k}}^{n} \frac{x - x_\ell}{x_k - x_\ell}\,.$$

This is implemented in the `OPQ` routine `NewtCotes.m` downloadable from the web site mentioned in MA 4. For reasons of economy, it uses the barycentric form (see Ch. 2, Sect. 2.2.5, (2.106)) of the Lagrange polynomials and the algorithm (*ibid.*, (2.108)) to compute the auxiliary quantities $\lambda_i^{(k)}$. Use the routine `NewtCotes.m` to explore the positivity of (WNC_n), i.e., $w_k^{(n)} > 0$, $k = 1, 2, \ldots, n$.

(a) Write a Matlab function `y=posNC(n,ab,ab0,eps0)`, which checks the positivity of the n-point Newton–Cotes formula with the abscissae being the zeros of the Jacobi polynomial $P_n^{(\alpha,\beta)}$ with parameters α, β, and using integration relative to the Jacobi weight function $w = w^{(\alpha_0,\beta_0)}$ with other parameters α_0, β_0. The former are selected by the `OPQ` routine `ab=r_jacobi(n,alpha,beta)`, from which the Jacobi abscissae can be obtained via the `OPQ` function `xw=gauss(n,ab)` as the first column of the $n \times 2$ array `xw`. The weight function w is provided by the routine `ab0=r_jacobi(floor((n+1)/2),alpha0,beta0)` which allows us to generate the required n_g-point Gaussian quadrature rule by the routine `xw=gauss(ng,ab0)`. The input parameter `eps0`, needed in the routine `NewtCotes.m`, is a number close to, but larger than, the machine precision `eps`, for example $\varepsilon_0 = .5 \times 10^{-14}$. Arrange the output parameter `y` to have the value 1 if all n weights of the Newton–Cotes formula (WNC_n) are positive, and the value 0 otherwise.

Use your routine for all $n \le N = 50$, $\alpha_0 = \beta_0 = 0, -1/2, 1/2$, and $\alpha = -1 + h : h : \alpha^+$, $\beta = \alpha : h : \beta^+$, where $\alpha^+ = \beta^+ = 3, 1.5, 4$ and $h = .05, .025, .05$ for the three values of α_0, β_0 respectively. Prepare plots in which a red plus sign is placed at the point (α, β) of the (α, β)-plane if positivity holds for all $n \le N$, and a blue dot otherwise. Explain why it suffices to consider only $\beta \ge \alpha$. {*Hint*: use the reflection formula $P_n^{(\beta,\alpha)}(x) = (-1)^n P_n^{(\alpha,\beta)}(-x)$ for Jacobi polynomials.} In a second set of plots show the exact upper boundary of the positivity domain created in the first plots; compute it by a bisection-type method (cf. Chap. 4, Sect. 4.3.1). (Running the programs for $N = 50$ may take a while. You may want to experiment with smaller values of N to see how the positivity domains vary.)

(b) The plots in (a) suggest that n-point Newton–Cotes formulae are positive for all $n \le N = 50$ on the line $0 \le \beta = \alpha$ up to a point $\alpha = \alpha_{\max}$. Use the same bisection-type method as in (a) to determine $\alpha_{\max}$ for the three values of α_0, β_0 and for $N = 20, 50, 100$ in each case.

(c) Repeat (a) with $\alpha = \beta = 0, -1/2, 1/2$ (Gauss–Legendre and Chebyshev abscissae of the first and second kinds) and $\alpha_0 = -.95 : .05 : \alpha_0^+$, $\beta_0 = \alpha_0 : .05 : \beta_0^+$, where $\alpha_0^+ = \beta_0^+ = 3.5, 3, 4$.

(d) Repeat (a) with $\alpha^+ = \beta^+ = 6, 4, 6$, but for weighted $(n+2)$-point Newton–Cotes formulae that contain as nodes the points ± 1 in addition to the n Jacobi abscissae.

(e) The plots in (d) suggest that the closed $(n+2)$-point Newton–Cotes formulae are positive for all $n \le N = 50$ on some line $\alpha_{\min} < \alpha = \beta < \alpha_{\max}$. Determine $\alpha_{\min}$ and $\alpha_{\max}$ similarly as $\alpha_{\max}$ in (b).

(f) Repeat (c) for the weighted closed $(n+2)$-point Newton–Cotes formula of (d) with $\alpha_0 = -1 + h : h : \alpha_0^+$, $\beta_0 = \alpha_0 : h : \beta_0^+$, where $\alpha_0^+ = \beta_0^+ = .5, -.2, 1$ and $h = .01, .01, .02$ for the three values of α, β.

4. Below are a number of suggestions as to how the following integrals may be computed,

$$I_c = \int_0^1 \frac{\cos x}{\sqrt{x}} \mathrm{d}x, \quad I_s = \int_0^1 \frac{\sin x}{\sqrt{x}} \mathrm{d}x.$$

(a) Use the composite trapezoidal rule with n intervals of equal length $h = 1/n$, "ignoring" the singularity at $x = 0$ (i.e., arbitrarily using zero as the value of the integrand at $x = 0$).

(b) Use the composite trapezoidal rule over the interval $[h, 1]$ with $n-1$ intervals of length $h = 1/n$ in combination with a weighted Newton–Cotes rule with weight function $w(x) = x^{-1/2}$ over the interval $[0, h]$. {Adapt the formula (3.43) to the interval $[0, h]$.}

(c) Make the change of variables $x = t^2$ and apply the composite trapezoidal rule to the resulting integrals.

(d) Use Gauss–Legendre quadrature on the integrals obtained in (c).

(e) Use Gauss–Jacobi quadrature with parameters $\alpha = 0$ and $\beta = -\frac{1}{2}$ directly on the integrals I_c and I_s.

{As a *point of information*, $I_c = \sqrt{2\pi}\, C\left(\sqrt{\frac{2}{\pi}}\right) = 1.809048475800\ldots$, $I_s = \sqrt{2\pi}\, S\left(\sqrt{\frac{2}{\pi}}\right) = .620536603446\ldots$, where $C(x)$, $S(x)$ are the Fresnel integrals.}

Implement and run the proposed methods for $n = 100 : 100 : 1000$ in (a) and (b), for $n = 20 : 20 : 200$ in (c), and for $n = 1 : 10$ in (d) and (e). Try to explain the results you obtain. {To get the required subroutines for Gaussian quadrature, download the `OPQ` routines `r_jacobi.m` and `gauss.m` from the web site `http://www.cs.purdue.edu/archives/2002/wxg/codes/OPQ.html`}

5. For a natural number p let

$$I_p = \int_0^1 (1-t)^p f(t) \mathrm{d}t$$

be (except for the factor $1/p!$) the pth iterated integral of f; cf. (3.50). Compare the composite trapezoidal rule based on n subintervals with the n-point Gauss–Jacobi rule on $[0,1]$ with parameters $\alpha = p$ and $\beta = 0$. Take, for example, $f(t) = \tan t$ and $p = 5 : 5 : 20$, and let $n = 10 : 10 : 50$ in the case of the trapezoidal rule, and $n = 2 : 2 : 10$ for the Gauss rule. {See MA 4 for instructions on how to download routines for generating the Gaussian quadrature rules.}

6. (a) Let $h_k = (b-a)/2^k$, $k = 0, 1, 2, \ldots$. Denote by

$$T_{h_k}(f) = h_k \left(\tfrac{1}{2} f(a) + \sum_{r=1}^{2^k - 1} f(a + r h_k) + \tfrac{1}{2} f(b) \right)$$

the composite trapezoidal rule and by

$$M_{h_k}(f) = h_k \sum_{r=1}^{2^k} f(a + (r - \tfrac{1}{2}) h_k)$$

the composite midpoint rule, both relative to a subdivision of $[a, b]$ into 2^k subintervals. Show that the first column $T_{k,0}$ of the Romberg array $\{T_{k,m}\}$ can be generated recursively as follows:

$$T_{0,0} = \frac{b-a}{2} [f(a) + f(b)],$$

$$T_{k+1,0} = \tfrac{1}{2} [T_{k,0} + M_{h_k}(f)], \quad k = 0, 1, 2, \ldots \ .$$

(b) Write a Matlab function for computing $\int_a^b f(x)\mathrm{d}x$ by the Romberg integration scheme, with $h_k = (b-a)/2^k$, $k = 0, 1, \ldots, n-1$.
Formal parameters: `a, b, n`; include f as a subfunction.
Output variable: the $n \times n$ Romberg array `T`.
Order of computation: Generate `T` row by row; generate the trapezoidal sums recursively as in part (a).
Program size: Keep it down to about 20 lines of Matlab code.

Output: $\mathtt{T}_{k,0}$, $\mathtt{T}_{k,k}$, $k = 0, 1, \ldots, n-1$.

(c) Call your subroutine (with $n = 10$) to approximate the following integrals.

1. $\displaystyle\int_1^2 \frac{\mathrm{e}^x}{x}\,\mathrm{d}x$ ("exponential integral")
2. $\displaystyle\int_0^1 \frac{\sin x}{x}\,\mathrm{d}x$ ("sine integral")
3. $\displaystyle\frac{1}{\pi}\int_0^\pi \cos(yx)\mathrm{d}x, \quad y = 1.7$

4. $\dfrac{1}{\pi}\displaystyle\int_0^{\pi} \cos(y\sin x)\mathrm{d}x, \quad y = 1.7$
5. $\displaystyle\int_0^{1} \sqrt{1-x^2}\,\mathrm{d}x$
6. $\displaystyle\int_0^{2} f(x)\mathrm{d}x, \quad f(x) = \begin{cases} x, & 0 \le x \le \sqrt{2}, \\ \dfrac{\sqrt{2}}{2-\sqrt{2}}\,(2-x), & \sqrt{2} \le x \le 2 \end{cases}$
7. $\displaystyle\int_0^{2} f(x)\mathrm{d}x, \quad f(x) = \begin{cases} x, & 0 \le x \le \dfrac{3}{4}, \\ \dfrac{3}{5}\,(2-x), & \dfrac{3}{4} \le x \le 2 \end{cases}$

(d) Comment on the behavior of the Romberg scheme in each of the seven cases in part (c).

ANSWERS TO THE EXERCISES AND MACHINE ASSIGNMENTS OF CHAPTER 3

ANSWERS TO EXERCISES

1. The table of divided differences is

x	f			
0	f_0			
$\frac{1}{8}$	f_1	$8(f_1-f_0)$		
$\frac{1}{4}$	f_2	$8(f_2-f_1)$	$32(f_2-2f_1+f_0)$	
$\frac{1}{2}$	f_3	$4(f_3-f_2)$	$\frac{32}{3}(f_3-3f_2+2f_1)$	$\frac{64}{3}(f_3-6f_2+8f_1-3f_0)$

   Therefore,

$$\begin{aligned} f'(0) &= 8(f_1-f_0) - \tfrac{1}{8}\cdot 32(f_2-2f_1+f_0) + \tfrac{1}{32}\cdot\tfrac{64}{3}\,(f_3-6f_2+8f_1-3f_0) + e_3 \\ &= \tfrac{1}{3}\,(2f_3-24f_2+64f_1-42f_0) + e_3, \end{aligned}$$

   where

$$|e_3| \le \frac{1}{8}\cdot\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{M_4}{4!} = \frac{M_4}{1536}\,.$$

2. From

$$f(x) = p_n(f;x) + r_n(x), \quad r_n(x) = [x_0, x_1, \ldots, x_n, x]f \cdot \omega_n(x),$$

   where $\omega_n(x) = \prod_{i=0}^{n}(x-x_i)$, we get (cf. Ch. 2, Ex. 58)

$$r_n'(x) = [x_0, x_1, \ldots, x_n, x]f \cdot \omega_n'(x) + [x_0, x_1, \ldots, x_n, x, x]f \cdot \omega_n(x),$$

   or

$$r_n'(x) = f^{(n+1)}(\xi)\,\omega_n'(x) + f^{(n+2)}(\xi')\,\omega_n(x),$$

   with ξ, ξ' contained in the span of $x_0, x_1, \ldots, x_n, x$, assuming that $f^{(n+2)}$ is continuous on that interval. Note that

$$\omega_n'(x) = \omega_n(x)\sum_{i=0}^{n}\frac{1}{x-x_i}\,, \quad x \ne x_i, \text{ all } i.$$

3. See the text.

4. We have $x_0 = 0$, $x_1 = \frac{1}{2}$, $x_2 = \frac{\sqrt{3}}{2}$, and the corresponding table of divided differences is

x	u		
0	u_0		
$\frac{1}{2}$	u_1	$2(u_1-u_0)$	
$\frac{\sqrt{3}}{2}$	u_2	$\frac{2}{\sqrt{3}-1}\,(u_2-u_1)$	$\frac{2}{\sqrt{3}}\left(\frac{2}{\sqrt{3}-1}\,(u_2-u_1)-2(u_1-u_0)\right)$

There follows from (3.4) (after renaming the nodes appropriately)

$$\left.\frac{\partial u}{\partial x}\right|_{P_1} \approx [x_0,x_1]u + (x_1-x_0)[x_0,x_1,x_2]u$$

$$= 2(u_1-u_0) + \frac{1}{2}\cdot\frac{2}{\sqrt{3}}\left(\frac{2}{\sqrt{3}-1}\,(u_2-u_1)-2(u_1-u_0)\right),$$

which can be simplified to

$$\left.\frac{\partial u}{\partial x}\right|_{P_1} \approx \frac{1}{3}(3+\sqrt{3})u_2 - (\sqrt{3}-1)u_1 - \frac{2}{3}\,(3-\sqrt{3})u_0.$$

For the error e we have from (3.5)

$$e = (x_1-x_0)(x_1-x_2)\,\frac{\dfrac{\partial^3 u}{\partial x^3}\left(\xi,\dfrac{1}{2}\right)}{6} = -\frac{\sqrt{3}-1}{24}\,\frac{\partial^3 u}{\partial x^3}\left(\xi,\frac{1}{2}\right),$$

hence

$$|e| \le \frac{\sqrt{3}-1}{24}\;M_3 = .030502\ldots M_3,\;\; M_3 = \max_{x_0\le x\le x_2}\left|\frac{\partial^3 u}{\partial x^3}\left(x;\frac{1}{2}\right)\right|.$$

5. (a) We have

$$\frac{\partial^2 u}{\partial x\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}u\right)$$

$$\approx \frac{\partial}{\partial y}\left\{\frac{1}{2h}[u(x+h,y)-u(x-h,y)]\right\}$$

$$\approx \frac{1}{2h}\left\{\frac{u(x+h,y+h)-u(x+h,y-h)}{2h} - \frac{u(x-h,y+h)-u(x-h,y-h)}{2h}\right\}$$

$$= \frac{1}{4h^2}\{u(x+h,y+h)-u(x+h,y-h)-u(x-h,y+h)+u(x-h,y-h)\}.$$

(b) Since the approximation obtained in (a) is an even function of h, its Taylor expansion involves only even powers of h. The constant term in

the expression between braces is clearly zero, and the remaining terms are

$$\tfrac{1}{2}h^2[(u_{xx}+2u_{xy}+u_{yy})-(u_{xx}-2u_{xy}+u_{yy})-(u_{xx}-2u_{xy}+u_{yy})$$
$$+(u_{xx}+2u_{xy}+u_{yy})]+O(h^4)=\tfrac{1}{2}h^2[8u_{xy}]+O(h^4)=4h^2u_{xy}+O(h^4),$$

where all partial derivatives are evaluated at (x,y). Thus, dividing by $4h^2$, we see that the approximation equals the exact value u_{xy} plus a term of $O(h^2)$.

6. (a) If n is even, then $T(2/n)$ is the trapezoidal rule applied to [–1,0] plus the trapezoidal rule applied to [0,1], each being exact.

   (b) Denoting by $T(h)[a,b]$ the composite trapezoidal rule applied to $[a,b]$ with steplength h, we have, when n is odd,

$$\begin{aligned}T(2/n)[-1,1]&=T(2/n)[-1,-1/n]+T(2/n)[1/n,1]+\frac{2}{n}\left[\frac{1}{2}\;\frac{1}{n}+\frac{1}{2}\;\frac{1}{n}\right]\\&=\int_{-1}^{-1/n}|x|\mathrm{d}x+\int_{1/n}^{1}|x|\mathrm{d}x+\frac{2}{n}\cdot\frac{1}{n}\\&=2\int_{1/n}^{1}x\mathrm{d}x+\frac{2}{n^2}=2\left[\frac{1}{2}\;x^2\right]_{1/n}^{1}+\frac{2}{n^2}\\&=1-\frac{1}{n^2}+\frac{2}{n^2}=1+\frac{1}{n^2}\;.\end{aligned}$$

   Thus, we still have the usual $O(h^2)$ convergence (cf. (3.25)), even though the integrand is only in $C[-1,1]$.

7. (a) One computes

$$I(h)=\int_0^h(x^2+x^{5/2})\mathrm{d}x=\frac{1}{3}h^3+\frac{2}{7}h^{7/2},$$
$$T(h)=\frac{h}{2}\;[h^2+h^{5/2}]=\frac{1}{2}h^3+\frac{1}{2}h^{7/2},$$

   so that

$$E(h)=-\frac{1}{6}h^3-\frac{3}{14}h^{7/2}=O(h^3)\quad\text{ as }\;h\to 0.$$

   (b) Likewise,

$$I(h)=\int_0^h(x^2+x^{1/2})\mathrm{d}x=\frac{1}{3}h^3+\frac{2}{3}h^{3/2},$$
$$T(h)=\frac{h}{2}\;[h^2+h^{1/2}]=\frac{1}{2}h^3+\frac{1}{2}h^{3/2},$$

   but now

$$E(h)=-\frac{1}{6}h^3+\frac{1}{6}h^{3/2}=O(h^{3/2})\quad\text{ as }\;h\to 0.$$

The error is larger by $1\frac{1}{2}$ orders, compared to the error in (a). The latter is typical for functions f, like the one in (a), which are in $C^2[0,h]$; the function f in (b) is only in $C[0,h]$.

8. (a) By Taylor's theorem, writing $x_{k+\frac{1}{2}} = x_k + \frac{1}{2}h$, we have

$$f(x) = f(x_{k+\frac{1}{2}}) + (x - x_{k+\frac{1}{2}})f'(x_{k+\frac{1}{2}}) + \tfrac{1}{2}(x - x_{k+\frac{1}{2}})^2 f''(\xi(x)),$$
$$x_k \le x \le x_k + h,$$

where $\xi(x) \in [x_k, x_k + h]$. Integrating gives

$$\int_{x_k}^{x_k+h} f(x)\mathrm{d}x = hf(x_{k+\frac{1}{2}}) + f'(x_{k+\frac{1}{2}}) \int_{x_k}^{x_k+h} (x - x_{k+\frac{1}{2}})\mathrm{d}x$$
$$+\frac{1}{2}\int_{x_k}^{x_k+h} (x - x_{k+\frac{1}{2}})^2 f''(\xi(x))\mathrm{d}x.$$

Here, the first integral on the right vanishes (by skew symmetry), and to the last integral we can apply the mean value theorem. This yields

$$\int_{x_k}^{x_k+h} f(x)\mathrm{d}x = hf(x_{k+\frac{1}{2}}) + \frac{1}{2}\ f''(\xi_k) \int_{x_k}^{x_k+h} (x - x_{k+\frac{1}{2}})^2 \mathrm{d}x,$$
$$x_k < \xi_k < x_k + h.$$

The integral on the right is most easily evaluated by making the change of variables $x = x_{k+\frac{1}{2}} + \frac{1}{2}th$, $-1 \le t \le 1$, giving

$$\int_{x_k}^{x_k+h} (x - x_{k+\frac{1}{2}})^2 \mathrm{d}x = \int_{-1}^{1} \left(\frac{1}{2}\ th\right)^2 \frac{1}{2}\ h\mathrm{d}t = \frac{h^3}{8}\int_{-1}^{1} t^2 \mathrm{d}t = \frac{h^3}{12},$$

so that

$$\int_{x_k}^{x_k+h} f(x)\mathrm{d}x = hf(x_{k+\frac{1}{2}}) + \frac{h^3}{24}\ f''(\xi_k), \quad x_k < \xi_k < x_k + h.$$

(b) Summing over all subintervals, we get

$$\int_a^b f(x)\mathrm{d}x = h\ \sum_{k=0}^{n-1} f(x_{k+\frac{1}{2}}) + \frac{h^2}{24}\ \frac{b-a}{n}\ \sum_{k=0}^{n-1}\ f''(\xi_k),$$

that is,

$$\int_a^b f(x)\mathrm{d}x = h\ \sum_{k=0}^{n-1} f(x_{k+\frac{1}{2}}) + \frac{h^2}{24}(b-a)f''(\xi), \quad a < \xi < b.$$

9. (a) The underlying Hermite interpolation problem can be solved by Newton's formula, using the table of divided differences

t	y			
-1	y_{-1}			
0	y_0	$y_0 - y_{-1}$		
0	y_0	y_0'	$y_0' - y_0 + y_{-1}$	
1	y_1	$y_1 - y_0$	$y_1 - y_0 - y_0'$	$\frac{1}{2}\,(y_1 - y_{-1} - 2y_0')$

The result is

$$p_3(y;t) = y_{-1} + (t+1)(y_0 - y_{-1}) + (t+1)t(y_0' - y_0 + y_{-1})$$
$$+\tfrac{1}{2}(t+1)t^2(y_1 - y_{-1} - 2y_0').$$

Integration yields

$$\int_{-1}^{1} p_3(y;t)\mathrm{d}t = y_{-1}\int_{-1}^{1}\mathrm{d}t + (y_0 - y_{-1})\int_{-1}^{1}(t+1)\mathrm{d}t$$
$$+(y_0' - y_0 + y_{-1})\int_{-1}^{1}(t+1)t\mathrm{d}t + \tfrac{1}{2}(y_1 - y_{-1} - 2y_0')\int_{-1}^{1}(t+1)t^2\mathrm{d}t$$
$$= 2y_{-1} + 2(y_0 - y_{-1}) + \tfrac{2}{3}(y_0' - y_0 + y_{-1}) + \tfrac{1}{3}(y_1 - y_{-1} - 2y_0')$$
$$= \tfrac{1}{3}(y_{-1} + 4y_0 + y_1).$$

(b) By interpolation theory, and the mean value theorem for integration,

$$E^S(y) = \int_{-1}^{1}(t+1)t^2(t-1)\frac{y^{(4)}(\tau(t))}{4!}\,\mathrm{d}t = -\frac{y^{(4)}(\tau)}{4!}\int_{-1}^{1}(1-t^2)t^2\mathrm{d}t$$
$$= -\frac{1}{4!}\,\frac{4}{15}\,y^{(4)}(\tau) = -\frac{1}{90}\,y^{(4)}(\tau), \quad -1 < \tau < 1.$$

(c) Let n be even, and $h = (b-a)/n$, $x_k = a + kh$, $f_k = f(x_k)$. With the change of variables $x = x_{k+1} + th$, $-1 \le t \le 1$, one gets

$$\int_{x_k}^{x_{k+2}} f(x)\mathrm{d}x = h\int_{-1}^{1} f(x_{k+1} + th)\mathrm{d}t.$$

Hence, letting $y(t) = f(x_{k+1} + th)$ in (a) and (b), we obtain

$$\int_{x_k}^{x_{k+2}} f(x)\mathrm{d}x = h\left\{\frac{1}{3}\,(f_k + 4f_{k+1} + f_{k+2}) - \frac{1}{90}\,h^4 f^{(4)}(\xi_k)\right\},$$
$$\xi_k = x_{k+1} + \tau h, \quad -1 < \tau < 1.$$

Summing over all even k from 0 to $n-2$, one gets

$$\int_a^b f(x)\mathrm{d}x = \sum_{\substack{k=0\\(k \text{ even})}}^{n-2} \int_{x_k}^{x_{k+2}} f(x)\mathrm{d}x$$

$$= h \sum_{\substack{k=0\\(k \text{ even})}}^{n-2} \frac{1}{3}(f_k + 4f_{k+1} + f_{k+2}) - \frac{1}{90}h^5 \sum_{\substack{k=0\\(k \text{ even})}}^{n-2} f^{(4)}(\xi_k)$$

$$= \frac{h}{3}\,(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{n-2} + 4f_{n-1} + f_n) + E_n^S(f),$$

where

$$E_n^S(f) = -\frac{1}{90}\,h^4 \;\cdot\; \frac{1}{2}\left[\frac{b-a}{n/2}\right] \sum_{\substack{k=0\\(k \text{ even})}}^{n-2} f^{(4)}(\xi_k).$$

Since the sum on the right contains exactly $n/2$ terms, division by $n/2$ yields a mean value of the fourth derivative, so that

$$E_n^S(f) = -\frac{b-a}{180}\,h^4 f^{(4)}(\xi),\ a < \xi < b.$$

10. With the notation $h = \frac{2\pi}{n}$ (n even), $x_k = \frac{2\pi}{n}\,k$, $f_k = f(x_k)$, $k = 0, 1, \ldots, n$, assuming f is 2π-periodic, one has

$$\begin{aligned} E_n^S(f) &= \int_0^{2\pi} f(x)\mathrm{d}x - \frac{h}{3} \sum_{k=0}^{(n-2)/2} (2f_{2k} + 4f_{2k+1}) \\ &= \int_0^{2\pi} f(x)\mathrm{d}x - \frac{h}{3}\left(2 \sum_{k=0}^{(n-2)/2} f_{2k} + 4 \sum_{k=0}^{(n-2)/2} f_{2k+1}\right). \end{aligned}$$

Clearly, $E_n^S(1) = 0$; hence, assume $f(x) = \mathrm{e}^{\mathrm{i}mx}$, $m \neq 0$. Then

$$\begin{aligned} E_n^S(\mathrm{e}^{\mathrm{i}mx}) &= -\frac{h}{3}\left(2 \sum_{k=0}^{(n-2)/2} \mathrm{e}^{\mathrm{i}mx_{2k}} + 4 \sum_{k=0}^{(n-2)/2} \mathrm{e}^{\mathrm{i}mx_{2k+1}}\right) \\ &= -\frac{h}{3}\left(2 \sum_{k=0}^{(n-2)/2} \mathrm{e}^{\mathrm{i}m\frac{2\pi}{n}\cdot 2k} + 4 \sum_{k=0}^{(n-2)/2} \mathrm{e}^{\mathrm{i}m\frac{2\pi}{n}(2k+1)}\right) \\ &= -\frac{h}{3}\left(2 \sum_{k=0}^{(n-2)/2} \left(\mathrm{e}^{\mathrm{i}m\frac{4\pi}{n}}\right)^k + 4\mathrm{e}^{\mathrm{i}m\frac{2\pi}{n}} \sum_{k=0}^{(n-2)/2} \left(\mathrm{e}^{\mathrm{i}m\frac{4\pi}{n}}\right)^k\right). \end{aligned}$$

If $m \neq 0 \pmod{\frac{n}{2}}$, then from the formula for geometric sums, $E_n^S(\mathrm{e}^{\mathrm{i}mx}) = 0$. If $m = \mu\frac{n}{2}$ ($\mu \neq 0$ an integer), then

$$\begin{aligned} E_n^S(\mathrm{e}^{\mathrm{i}mx}) &= -\frac{h}{3}\left(2\left(\frac{n-2}{2}\ +1\right) + 4\mathrm{e}^{\mathrm{i}\mu\pi}\left(\frac{n-2}{2}\ +1\right)\right) \\ &= -\frac{2}{3}\ h\ \frac{n}{2}\ (1+2(-1)^\mu) = -\frac{2\pi}{3}\,(1+2(-1)^\mu) \neq 0. \end{aligned}$$

Thus, Simpson's rule integrates exactly all trigonometric polynomials of degrees $d \leq \frac{n}{2} - 1$.

11. We have

$$f(x) = \mathrm{e}^{-x^2},\ f''(x) = 2(2x^2-1)\mathrm{e}^{-x^2},\ f^{(4)}(x) = 4(4x^4-12x^2+3)\mathrm{e}^{-x^2},$$

Here, $f'''(x) > 0$ on $[0,1]$, so that $|f''(x)| \leq 2$. Moreover, $f^{(5)}(x)$ is negative on $\left(0, \sqrt{\frac{5-\sqrt{10}}{2}}\right)$ and positive on $\left(\sqrt{\frac{5-\sqrt{10}}{2}}, 1\right]$, so that $f^{(4)}$ decreases from 12 to a negative minimum $-7.419\ldots$ at $x = \sqrt{\frac{5-\sqrt{10}}{2}}$ and then increases to $-20\mathrm{e}^{-1} = -7.357\ldots$ at $x = 1$. Therefore, $|f^{(4)}(x)| \leq 12$ on $[0,1]$

(a) Since

$$E_n^T = -\frac{b-a}{12}\ h^2 f''(\xi) = -\frac{1}{12n^2}\ f''(\xi), \quad 0 < \xi < 1,$$

we have

$$|E_n^T| \leq \frac{1}{6n^2}\ \leq\ \frac{1}{2}\ 10^{-6} \quad \text{if } n^2 \geq \frac{1}{3}\cdot 10^6, \ \text{ i.e., } \ n \geq \frac{10^3}{\sqrt{3}} = 577.35\ldots\ .$$

Therefore, the required number of subintervals is 578.

(b) Since

$$E_n^S = -\frac{b-a}{180}\ h^4 f^{(4)}(\xi) = -\frac{1}{180n^4}\ f^{(4)}(\xi), \quad 0 < \xi < 1,$$

we have

$$|E_n^S| \leq \frac{1}{180n^4}\cdot 12 = \frac{1}{15n^4} \leq \frac{1}{2}\ 10^{-6}$$

if

$$n^4 \geq \frac{2}{15}\cdot 10^6, \ \text{ i.e., } \ n \geq \left[\frac{200}{15}\right]^{1/4}\cdot 10 = 19.108\ldots\ .$$

Therefore, the required number of subintervals is 20.

12. (a) Write

$$\int_0^1 f(x)\mathrm{d}x = \int_0^{1/2} f(x)\mathrm{d}x + \int_{1/2}^1 f(x)\mathrm{d}x$$

and make the change of variables $x = 1 - t$ in the second integral. This gives

$$\int_0^1 f(x)\mathrm{d}x = \int_0^{1/2} f(x)\mathrm{d}x - \int_{1/2}^0 f(1-t)\mathrm{d}t = \int_0^{1/2} f(x)\mathrm{d}x$$
$$+\int_0^{1/2} f(1-t)\mathrm{d}t = \int_0^{1/2} [f(x) + f(1-x)]\mathrm{d}x = \int_0^{1/2} 1 \cdot \mathrm{d}x = \frac{1}{2}\,.$$

(b) *By explicit computation.* For n even, the composite trapezoidal sum, in the usual notation, is

$$h\left\{\frac{1}{2}(f(0)+f(1)) + \sum_{k=1}^{(n/2)-1} (f(x_k) + f(1-x_k)) + f(x_{n/2})\right\}$$

$$= h\left\{\frac{1}{2} + \left(\frac{n}{2} - 1\right) + \frac{1}{2}\right\} \quad \left(\text{since } f(x_{n/2}) = f\left(\frac{1}{2}\right) = \frac{1}{2}\right)$$

$$= h\,\frac{n}{2} = \frac{1}{2} \quad (\text{since } hn = 1).$$

For n odd, similarly,

$$h\left\{\frac{1}{2}\,(f(0)+f(1))\right\} + \sum_{k=1}^{(n-1)/2} (f(x_k) + f(1-x_k))$$
$$= h\left\{\frac{1}{2} + \frac{n-1}{2}\right\} = h\,\frac{n}{2} = \frac{1}{2}\,.$$

By a noncomputational argument. By symmetry, the composite trapezoidal rule must be of the form

$$h(\alpha_0 n + \alpha_1) = \alpha_0 + \alpha_1 h$$

for some constants α_0, α_1. Assuming $f \in C^2[0,1]$, the error $\frac{1}{2} - (\alpha_0 + \alpha_1 h)$ is $O(h^2)$, hence $\alpha_0 = \frac{1}{2}$, $\alpha_1 = 0$.

(c) Use the noncomputational argument in (b), which works for any symmetric rule with error at least $O(h^2)$, i.e., for any such rule whose nodes are symmetric with respect to the midpoint $\frac{1}{2}$ and whose weights are equal for symmetrically located nodes.

13. (a) Putting in turn $f(x) = \cos x$ and $f(x) = \sin x$ gives

$$a + b\cos h = \sin h,$$
$$b\sin h = 1 - \cos h,$$

which yields

$$
\begin{aligned}
b &= \frac{1-\cos h}{\sin h} = \frac{2\sin^2 \frac{1}{2}h}{2\sin \frac{1}{2}h \cos \frac{1}{2}h} = \tan \tfrac{1}{2}h, \\
a &= \sin h - b\cos h = 2\sin \tfrac{1}{2}h \cos \tfrac{1}{2}h - \tan \tfrac{1}{2}h \ [\cos^2 \tfrac{1}{2}h - \sin^2 \tfrac{1}{2}h] \\
&= \sin \tfrac{1}{2}h \cos \tfrac{1}{2}h + \sin^2 \tfrac{1}{2}h \tan \tfrac{1}{2}h \\
&= \sin \tfrac{1}{2}h \left[\cos \tfrac{1}{2}h + \frac{\sin^2 \frac{1}{2}h}{\cos \frac{1}{2}h}\right] \\
&= \sin \tfrac{1}{2}h \, \frac{\cos^2 \frac{1}{2}h + \sin^2 \frac{1}{2}h}{\cos \frac{1}{2}h} \\
&= \tan \tfrac{1}{2}h.
\end{aligned}
$$

Thus,

$$\int_0^h f(x)\mathrm{d}x = \tan \tfrac{1}{2}h \ [f(0) + f(h)] + E(f).$$

Since $\tan \frac{1}{2}h \sim \frac{1}{2}h$ for $h \to 0$, this is a modified trapezoidal rule. It does *not* integrate constants exactly, but

$$E(1) = h - 2\tan \tfrac{1}{2}h = h - 2(\tfrac{1}{2}h + \tfrac{1}{24}h^3 + \cdots\ ) = O(h^3) \quad \text{as} \quad h \to 0.$$

(b) We have, with the change of variables $t = c + x$,

$$\int_c^{c+h} g(t)\mathrm{d}t = \int_0^h g(c+x)\mathrm{d}x.$$

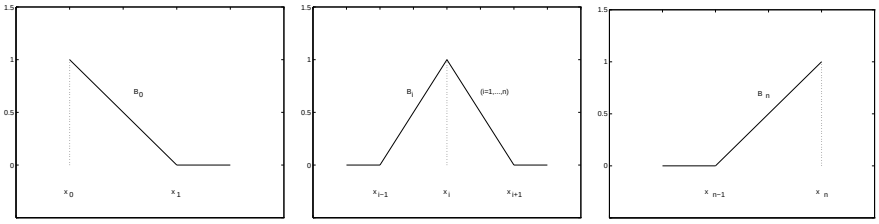
Since $\cos(c+x)$ and $\sin(c+x)$ are linear combinations of $\cos x$ and $\sin x$, one gets by (a)

$$\int_0^h g(c+x)\mathrm{d}x = \tan \tfrac{1}{2}h \ [g(c) + g(c+h)] + E(g),$$

where $E(g) = 0$ for $g(x) = \cos x$ and $g(x) = \sin x$.

14. See the text.

15. (a) The "hat functions" are defined as shown in the figure.



Then

$$s_1(f;x) = \sum_{i=0}^{n} f(x_i)B_i(x),$$
$$\int_0^1 s_1(f;x)\mathrm{d}x = \sum_{i=0}^{n} f(x_i)\int_0^1 B_i(x)\mathrm{d}x.$$

Now, on geometric grounds,

$$\int_0^1 B_0(x)\mathrm{d}x = \frac{1}{2}\,\Delta x_0, \qquad \int_0^1 B_n(x)\mathrm{d}x = \frac{1}{2}\,\Delta x_{n-1},$$
$$\int_0^1 B_i(x)\mathrm{d}x = \int_{x_{i-1}}^{x_{i+1}} B_i(x)\mathrm{d}x = \frac{1}{2}\,(\Delta x_{i-1} + \Delta x_i), \quad 1 \le i < n.$$

There follows

$$\int_0^1 s_1(f;x)\mathrm{d}x = \frac{1}{2}\,\Delta x_0 \cdot f(x_0) + \frac{1}{2}\sum_{i=1}^{n-1}(\Delta x_{i-1} + \Delta x_i)f(x_i) + \frac{1}{2}\,\Delta x_{n-1}\cdot f(x_n).$$

If $\Delta x_i = h$, $i = 0, 1, \ldots, n-1$, this becomes the composite trapezoidal formula.

(b) According to interpolation theory, for $x_i \le x \le x_{i+1}$, $i = 0, 1, \ldots, n-1$,

$$f(x) - s_1(f;x) = (x - x_i)(x - x_{i+1})\,\frac{f''(\xi_i(x))}{2}, \quad x_i < \xi_i(x) < x_{i+1}.$$

Thus,

$$\int_0^1 f(x)\mathrm{d}x - \int_0^1 s_1(f;x)\mathrm{d}x = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} [f(x) - s_1(f;x)]\mathrm{d}x$$
$$= \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1})\,\frac{f''(\xi_i(x))}{2}\,\mathrm{d}x.$$

With the change of variables $x = x_i + t\Delta x_i$, and applying the mean

value theorem of integration, this becomes

$$\begin{aligned}
&\sum_{i=0}^{n-1} \frac{1}{2} (\Delta x_i)^3 f''(\xi_i) \int_0^1 t(t-1)\mathrm{d}t \\
&= -\frac{1}{12} \sum_{i=0}^{n-1} (\Delta x_i)^3 f''(\xi_i) \\
&= -\frac{1}{12} \left(\sum_{j=0}^{n-1} (\Delta x_j)^3 \right) \sum_{i=0}^{n-1} \frac{(\Delta x_i)^3}{\sum_{j=0}^{n-1} (\Delta x_j)^3} f''(\xi_i) \\
&= -\frac{1}{12} \left(\sum_{j=0}^{n-1} (\Delta x_j)^3 \right) f''(\xi), \quad 0 < \xi < 1.
\end{aligned}$$

Therefore,

$$E(f) = \text{const} \cdot f''(\xi), \quad \text{const} = -\frac{1}{12} \sum_{j=0}^{n-1} (\Delta x_j)^3.$$

If $\Delta x_i = h$, $i = 0, 1, \ldots, n-1$, then

$$\text{const} = -\frac{1}{12}\, nh^3 = -\frac{1}{12}\, nh \cdot h^2 = -\frac{1}{12}\, h^2,$$

and we get the error term (3.25) for the composite trapezoidal rule.

16. (a) Exactness for $f(x) \equiv 1$ and $f(x) \equiv x$ yields

$$\begin{aligned}
a_0 + a_1 &= \tfrac{1}{\alpha+1}\,, \\
a_1 &= \tfrac{1}{\alpha+2}\,,
\end{aligned}$$

giving

$$a_0 = \frac{1}{(\alpha+1)(\alpha+2)}\,, \quad a_1 = \frac{1}{\alpha+2}\,.$$

The formula becomes

$$\int_0^1 f(x) x^\alpha \mathrm{d}x = \frac{1}{\alpha+2} \left[\frac{1}{\alpha+1}\, f(0) + f(1) \right] + E(f).$$

If $\alpha = 0$, this reduces to the trapezoidal rule, as it should. If $-1 < \alpha < 0$, the weight function x^α is unbounded at $x = 0$, which explains why $a_0 > a_1$ in this case. If $\alpha > 0$, then x^α vanishes at $x = 0$, and therefore, appropriately, $a_0 < a_1$.

(b) From interpolation theory, we have

$$\begin{aligned} E(f) &= \int_0^1 [f(x) - p_1(f;x)]x^\alpha \mathrm{d}x = \int_0^1 \frac{f''(\xi(x))}{2!} x(x-1)\cdot x^\alpha \mathrm{d}x \\ &= -\frac{f''(\xi)}{2}\int_0^1 x^{\alpha+1}(1-x)\mathrm{d}x = -\frac{f''(\xi)}{2}\left(\frac{1}{\alpha+2} - \frac{1}{\alpha+3}\right) \\ &= -\frac{1}{2(\alpha+2)(\alpha+3)} f''(\xi), \quad 0 < \xi < 1. \end{aligned}$$

(c) By (a) and (b) we have

$$\int_0^1 f(x)x^\alpha \mathrm{d}x = \frac{1}{\alpha+2}\left[\frac{1}{\alpha+1} f(0) + f(1)\right] - \frac{1}{2(\alpha+2)(\alpha+3)} f''(\xi),$$
$$0 < \xi < 1.$$

Therefore, using the change of variables $t = xh$, one gets

$$\begin{aligned} &\int_0^h g(t)t^\alpha \mathrm{d}t = \int_0^1 g(xh)(xh)^\alpha h \mathrm{d}x = h^{\alpha+1}\int_0^1 g(xh)x^\alpha \mathrm{d}x \\ &= h^{\alpha+1}\left\{\frac{1}{\alpha+2}\left[\frac{1}{\alpha+1} g(0) + g(h)\right] - \frac{h^2 g''(\xi h)}{2(\alpha+2)(\alpha+3)}\right\} \\ &= \frac{h^{\alpha+1}}{\alpha+2}\left[\frac{1}{\alpha+1} g(0) + g(h)\right] - \frac{h^{\alpha+3}}{2(\alpha+2)(\alpha+3)} g''(\overline{h}), \\ &\qquad 0 < \overline{h} < h. \end{aligned}$$

17. (a) Putting $f(x) \equiv 1$ and $f(x) \equiv x$ gives, respectively,

$$\begin{aligned} a_0 + a_1 &= \tfrac{1}{4}, \\ a_1 &= \tfrac{1}{9}. \end{aligned}$$

Hence, $a_0 = 5/36$ and $a_1 = 1/9$, giving

$$\int_0^1 f(x)\cdot x\ \ln(1/x)\mathrm{d}x \approx \frac{5}{36} f(0) + \frac{1}{9} f(1).$$

(b) Let $t = hx$, $\mathrm{d}t = h\mathrm{d}x$. Then

$$\begin{aligned} &\int_0^h g(t)\cdot t\ \ln(1/t)\mathrm{d}t = \int_0^1 g(hx)\cdot hx \cdot \ln\left(\frac{1}{hx}\right)\cdot h\mathrm{d}x \\ &= h^2\left\{\int_0^1 g(hx)\cdot x\ \ln\left(\frac{1}{x}\right)\mathrm{d}x + \int_0^1 g(hx)\cdot x\ln\left(\frac{1}{h}\right)\mathrm{d}x\right\}. \end{aligned}$$

Evaluating the first integral as in (a), and the second by the trapezoidal rule (for example), one gets

$$\int_0^h g(t)\cdot t\ \ln(1/t)\mathrm{d}t \approx h^2\left(\frac{5}{36}\,g(0)+\frac{1}{9}\,g(h)\right)+\frac{1}{2}\,h^2\ln\left(\frac{1}{h}\right)g(h).$$

18. (a) We have $s(x-0)=s(x+0)=1$, but $s'(x-0)=3\neq s'(x+0)=-3$, so that $s\in\mathbb{S}_3^0(\Delta)$.

(b) Since the composite trapezoidal rule, with n even, can be interpreted as the sum of two composite trapezoidal rules, one over the interval $[-1,0]$, the other over $[0,1]$, the error is

$$E_n^T(s)=-\frac{h^2}{12}[s''(\xi_{-1})+s''(\xi_1)]$$

for some $-1<\xi_{-1}<0$, $0<\xi_1<1$. Thus,

$$E_n^T(s)=-\frac{h^2}{12}\cdot 6(\xi_{-1}+1+1-\xi_1)=-\frac{h^2}{2}(2+\xi_{-1}-\xi_1).$$

Since $-2<\xi_{-1}-\xi_1<0$, we can write this also as

$$E_n^T(s)=-\frac{h^2}{2}\,\xi,\quad 0<\xi<2.$$

(c) If n is divisible by 4, the error is zero since, similarly as in (a), it is the error of two composite Simpson rules, each integrating cubics exactly. If n is not divisible by 4, then the error is

$$\begin{aligned}E_n^S(s)&=\int_{-h}^h s(x)\mathrm{d}x-\frac{h}{3}\,(s(-h)+4s(0)+s(h))\\&=2\int_0^h s(x)\mathrm{d}x-\frac{2h}{3}\,(2s(0)+s(h)),\end{aligned}$$

which by an elementary calculation becomes

$$E_n^S(s)=-h^2\left(1-\tfrac{1}{6}h^2\right)=O(h^2).$$

It is seen that Simpson's rule, in this case, has the same order of accuracy as the trapezoidal rule, the reason being the lack of differentiability of s at $x=0$.

(d) The error is zero since the Gauss–Legendre rule integrates cubics exactly.

19. Let $\pi_{n+1}^*(t)=\prod_{k=1}^{n+1}(t-t_k^*)$. Since $\prod_{k=1}^n(t-t_k^{(n)})=\pi_n(t;w)$, and the node polynomial is $\omega_n(t)=\pi_n(t;w)\pi_{n+1}^*(t)$, the quadrature rule in question, by

part (b) of Theorem 3.2.1, to have degree of exactness $d = 2n + 1 - 1 + k = 2n + k$, must satisfy

$$\int_a^b \pi_{n+1}^*(t)p(t)\pi_n(t;w)w(t)\mathrm{d}t = 0, \quad \text{all } p \in \mathbb{P}_{k-1}.$$

If we want $d = 3n + 1$, this must hold for $k = n + 1$, i.e., $\pi_{n+1}^*(t)$ must be orthogonal to all lower-degree polynomials relative to the (oscillating) weight function $\pi_n(t;w)w(t)$. By part (a) of the same theorem, the quadrature rule must also be interpolatory, which uniquely determines the weights w_k and w_k^* if the t_k^* are distinct among themselves and from the $t_k^{(n)}$.

20. *Necessity.* (a) is trivial. To show (b), observe that for any $p \in \mathbb{P}_{n-1}$ one has

$$f(t) := [\omega_n(t)]^3 p(t) \in \mathbb{P}_{4n-1},$$

and since $d = 4n - 1$,

$$\int_a^b f(t)w(t)\mathrm{d}t = \sum_{k=1}^n [w_k f(t_k) + w_k' f'(t_k) + w_k'' f''(t_k)].$$

But

$$\begin{aligned} f(t_k) &= 0, \\ f'(t_k) &= [\omega_n^3]'(t_k)p(t_k) + [\omega_n^3](t_k)p'(t_k) = 0, \\ f''(t_k) &= [\omega_n^3]''(t_k)p(t_k) + 2[\omega_n^3]'(t_k)p'(t_k) + [\omega_n^3](t_k)p''(t_k) = 0, \end{aligned}$$

so that

$$\int_a^b [\omega_n(t)]^3 p(t)w(t)\mathrm{d}t = 0.$$

Sufficiency. Let $p \in \mathbb{P}_{4n-1}$ and assume (a) and (b). Divide p by ω_n^3:

$$p(t) = q(t)[\omega_n(t)]^3 + r(t), \quad q \in \mathbb{P}_{n-1}, \quad r \in \mathbb{P}_{3n-1}.$$

Then

$$\begin{aligned} \int_a^b p(t)w(t)\mathrm{d}t &= \int_a^b [\omega_n(t)]^3 q(t)w(t)\mathrm{d}t + \int_a^b r(t)w(t)\mathrm{d}t \\ &\overset{\text{(b)}}{=} \int_a^b r(t)w(t)\mathrm{d}t \\ &\overset{\text{(a)}}{=} \sum_{k=1}^n [w_k r(t_k) + w_k' r'(t_k) + w_k'' r''(t_k)] \\ &= \sum_{k=1}^n \big\{ w_k \left(p(t_k) - q(t_k)[\omega_n]^3(t_k)\right) \\ &\qquad + w_k' \left(p'(t_k) - [q\omega_n^3]'(t_k)\right) + w_k'' \left(p''(t_k) - [q\omega_n^3]''(t_k)\right) \big\} \\ &= \sum_{k=1}^n [w_k p(t_k) + w_k' p'(t_k) + w_k'' p''(t_k)], \end{aligned}$$

i.e., the quadrature rule has degree of exactness $d = 4n - 1$.

21. The necessary and sufficient conditions, according to Theorem 3.2.1, are that (i) each quadrature rule Q_σ be interpolatory, and (ii) the node polynomial $\omega_n(t) = \Pi_{k=1}^n (t - t_k)$ satisfy

$$\int_a^b \omega_n(t) p(t) w_\sigma(t) \mathrm{d}t = 0, \quad \text{all} \ \ p \in \mathbb{P}_{m_\sigma - 1}, \quad \sigma = 1, 2, \ldots, s.$$

The requirement (i) involves ns conditions while (ii) involves $\sum_{\sigma=1}^s m_\sigma = n$ conditions. Altogether, these are $ns + n$ conditions for the same number of unknowns: the n nodes and the ns weights. Thus, one expects the formulae to be optimal.

22. See the text.

23. For the Gauss–Radau formula with $t_1 = a$,

$$\int_a^b f(t) w(t) \mathrm{d}t = w_1 f(a) + \sum_{k=2}^n w_k f(t_k) + E_n(f),$$
$$E_n(\mathbb{P}_{2n-2}) = 0, \quad n \geq 2,$$

let

$$p_1(t) = \prod_{k=2}^n \left(\frac{t - t_k}{a - t_k} \right)^2,$$
$$p_j(t) = \frac{t - a}{t_j - a} \prod_{\substack{k=2 \\ k \neq j}}^n \left(\frac{t - t_k}{t_j - t_k} \right)^2, \quad j = 2, \ldots, n.$$

Here, $p_1 \in \mathbb{P}_{2n-2}$ and $p_j \in \mathbb{P}_{2n-3}$ for $j = 2, \ldots, n$, and $p_1 \geq 0$, $p_j \geq 0$ on $[a, b]$. Applying the Gauss–Radau formula to $f = p_1$ and $f = p_j$ then gives

$$0 < \int_a^b p_1(t) w(t) \mathrm{d}t = w_1,$$
$$0 < \int_a^b p_j(t) w(t) \mathrm{d}t = w_j, \quad j = 2, \ldots, n.$$

The reasoning is the same for the Gauss–Radau formula with $t_n = b$.

For the Gauss–Lobatto formula

$$\int_a^b f(t) w(t) \mathrm{d}t = w_1 f(a) + \sum_{k=2}^{n-1} w_k f(t_k) + w_n f(b) + E_n(f),$$
$$E_n(\mathbb{P}_{2n-3}) = 0, \quad n \geq 3,$$

one argues similarly, taking

$$p_1(t) = \frac{b-t}{b-a}\prod_{k=2}^{n-1}\left(\frac{t-t_k}{a-t_k}\right)^2,$$
$$p_j(t) = \frac{t-a}{t_j-a}\,\frac{b-t}{b-t_j}\prod_{\substack{k=2\\k\neq j}}^{n-1}\left(\frac{t-t_k}{t_j-t_k}\right)^2, \quad j=2,\ldots,n-1,$$
$$p_n(t) = \frac{t-a}{b-a}\prod_{k=2}^{n-1}\left(\frac{t-t_k}{b-t_k}\right)^2.$$

Since both formulae have degrees of exactness that tend to infinity as $n\to\infty$, their convergence is proved in the same way as for Gaussian quadrature rules; cf. Sect. 3.2.3(c).

24. Clearly, $\Delta_k \in \mathbb{P}_{2n-2}$ and $\Delta_k(t_\ell) = 0$ for $\ell = 1, 2, \ldots, n$. Therefore,

$$\Delta_k(t) = q(t)\omega_n(t), \quad q \in \mathbb{P}_{n-2}.$$

Expand q in Legendre polynomials,

$$q(t) = c_0P_0(t) + c_1P_1(t) + \cdots + c_{n-2}P_{n-2}(t).$$

Since the leading coefficient of Δ_k, and hence of q, is positive, we have $c_{n-2} > 0$. (The leading coefficient of P_k is positive for each k.) We then have, by orthogonality of the Legendre polynomials,

$$\begin{aligned}\int_{-1}^{1}\Delta_k(t)\mathrm{d}t &= \int_{-1}^{1}\sum_{j=0}^{n-2}c_jP_j(t)[P_n(t)+\alpha P_{n-1}(t)+\beta P_{n-2}(t)]\mathrm{d}t\\ &= \beta c_{n-2}\int_{-1}^{1}P_{n-2}^2(t)\mathrm{d}t \le 0\end{aligned}$$

since $\beta \le 0$, that is, using (3.39),

$$\int_{-1}^{1}[\ell_k(t)]^2\mathrm{d}t \le \int_{-1}^{1}\ell_k(t)\mathrm{d}t = w_k.$$

Therefore, $w_k > 0$, all k.

25. (a) The appropriate table of divided differences, with $f_0 = f(0)$, $f_1 = f(1)$, $f_0' = f'(0)$, is

x	$f(x)$			
0	f_0			
0	f_0	f_0'		
1	f_1	f_1-f_0	$f_1-f_0-f_0'$	

Therefore, by Newton's formula (with remainder),

$$p(x) = f_0 + xf_0' + x^2(f_1 - f_0 - f_0') = (1-x^2)f_0 + x^2 f_1 + x(1-x)f_0',$$
$$f(x) = p(x) + R(x), \quad R(x) = \frac{x^2(x-1)}{3!}\, f'''(\xi(x)), \quad 0 < \xi(x) < 1.$$

(b) We have

$$\int_0^1 f(x)\mathrm{d}x = f_0 \int_0^1 (1-x^2)\mathrm{d}x + f_1 \int_0^1 x^2\mathrm{d}x + f_0' \int_0^1 x(1-x)\mathrm{d}x + \int_0^1 R(x)\mathrm{d}x = \frac{2}{3}\, f(0) + \frac{1}{3}\, f(1) + \frac{1}{6}\, f'(0) + E(f),$$

hence,

$$a_0 = \frac{2}{3},\ a_1 = \frac{1}{3},\ b_0 = \frac{1}{6}, \quad E(f) = \int_0^1 R(x)\mathrm{d}x.$$

By the mean value theorem,

$$E(f) = \int_0^1 \frac{x^2(x-1)}{6}\, f'''(\xi(x))\mathrm{d}x = -\frac{1}{6}\, f'''(\xi) \int_0^1 x^2(1-x)\mathrm{d}x = -\frac{1}{72}\, f'''(\xi), \quad 0 < \xi < 1.$$

(c) Putting $t = c + xh$, we have

$$\int_c^{c+h} y(t)\mathrm{d}t = h \int_0^1 y(c+xh)\mathrm{d}x,$$

and applying (b) to $f(x) = y(c+xh)$, with $f^{(k)}(x) = h^k y^{(k)}(c+xh)$, we get

$$\int_c^{c+h} y(t)\mathrm{d}t = h\left[\frac{2}{3}\, y(c) + \frac{1}{3}\, y(c+h) + \frac{h}{6}\, y'(c)\right] - \frac{h^4}{72}\, y'''(c+\xi h), \quad 0 < \xi < 1.$$

26. Orthogonality is shown by repeated integration by parts, exactly as in Ch. 2, Sect. 2.1.4(2).

From the recurrence relation (2.38) of Ch. 2 we get

$$\alpha_k = \frac{t\pi_k(t) - \pi_{k+1}(t)}{\pi_k(t)} - \beta_k \frac{\pi_{k-1}(t)}{\pi_k(t)},$$

which is valid for all t not equal to a zero of π_k. In particular, as $t \to \infty$,

$$\alpha_k = \lim_{t\to\infty} \frac{t\pi_k(t) - \pi_{k+1}(t)}{\pi_k(t)} = \lim_{t\to\infty} \frac{(\lambda_k - \lambda_{k+1})t^k + \cdots}{t^k + \cdots} = \lambda_k - \lambda_{k+1}, \ \ k \geq 0,$$

where $\lambda_k = 0$ if $k = 0$ and $\lambda_k - \lambda_{k+1} > 0$ (as will become apparent shortly). Similarly,

$$\begin{aligned}\beta_k &= \frac{(t-\alpha_k)\pi_k(t) - \pi_{k+1}(t)}{\pi_{k-1}(t)} \\ &= \frac{\lambda_k t^k + \mu_k t^{k-1} - \alpha_k(t^k + \lambda_k t^{k-1}) - \lambda_{k+1}t^k - \mu_{k+1}t^{k-1} + \cdots}{t^{k-1} + \cdots} \\ &= \frac{(\lambda_k - \lambda_{k+1} - \alpha_k)t^k + (\mu_k - \mu_{k+1} - \alpha_k\lambda_k)t^{k-1} + \cdots}{t^{k-1} + \cdots} \\ &= \mu_k - \mu_{k+1} - \alpha_k\lambda_k,\end{aligned}$$

where the above formula for α_k has been used and the right-hand side is positive since $\beta_k > 0$.

By Leibniz's rule of differentiation, we now have

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}(t^{\alpha+k}\mathrm{e}^{-t})$$

$$= (-1)^k t^\alpha \mathrm{e}^{-t}[t^k - k(\alpha+k)t^{k-1} + \binom{k}{2}(\alpha+k)(\alpha+k-1)t^{k-2} - \cdots\ ],$$

hence

$$\pi_k(t) = t^k - k(\alpha+k)t^{k-1} + \tfrac{1}{2}k(k-1)(\alpha+k)(\alpha+k-1)t^{k-2} - \cdots\ .$$

There follows

$$\lambda_k = -k(\alpha+k), \quad \mu_k = \tfrac{1}{2}k(k-1)(\alpha+k)(\alpha+k-1).$$

Therefore,

$$\begin{aligned}\alpha_k &= \lambda_k - \lambda_{k+1} \\ &= -k(\alpha+k) + (k+1)(\alpha+k+1) = 2k + \alpha + 1, \\ \beta_k &= \mu_k - \mu_{k+1} - \alpha_k\lambda_k \\ &= \tfrac{1}{2}k(k-1)(\alpha+k)(\alpha+k-1) - \tfrac{1}{2}(k+1)k(\alpha+k+1)(\alpha+k) \\ &\qquad + (2k+\alpha+1)k(\alpha+k) \\ &= \tfrac{1}{2}k(\alpha+k)[(k-1)(\alpha+k-1) - (k+1)(\alpha+k+1) + 2(2k+\alpha+1)] \\ &= k(\alpha+k).\end{aligned}$$

27. Orthogonality follows as in the Example of Ch. 2, Sect. 2.1.4(2), by repeated integration by parts.

From the Rodrigues formula given for π_k it follows directly that

$$\begin{aligned}\pi_{k+1}(t) &= \frac{(-1)^{k+1}}{2^{k+1}} e^{t^2} \frac{d^k}{dt^k}\left(\frac{d}{dt} e^{-t^2}\right) \\ &= \frac{(-1)^{k+1}}{2^{k+1}} e^{t^2} \frac{d^k}{dt^k}\left(-2te^{-t^2}\right) \\ &= \frac{(-1)^k}{2^k} e^{t^2}\left(t\frac{d^k}{dt^k} e^{-t^2} + k\frac{d^{k-1}}{dt^{k-1}} e^{-t^2}\right) \\ &= t\pi_k(t) - \tfrac{1}{2}k\pi_{k-1}(t).\end{aligned}$$

Since $\pi_0 = 1$, this also shows that all π_k are monic.

28. (a) Let $\pi_2(t) = t^2 + at + b$. Then orthogonality of π_2 to 1 and t gives the equations (use the *Hint*):

$$\begin{aligned} 2! + 1!a + \quad b &= 0, \\ 3! + 2!a + 1!b &= 0, \end{aligned}$$

hence $a + b = -2$, $2a + b = -6$, that is, $a = -4$, $b = 2$. Therefore,

$$\pi_2(t) = t^2 - 4t + 2.$$

(b) The zeros of π_2 are $t_1 = 2 + \sqrt{2}$, $t_2 = 2 - \sqrt{2}$. Therefore, the 2-point Gauss–Laguerre formula has the form

$$\int_0^\infty f(t)e^{-t}dt = w_1 f(t_1) + w_2 f(t_2) + E_2(f).$$

Since $E_2(f) = 0$ for $f(t) = 1$ and $f(t) = t$, we get

$$\begin{aligned} w_1 + \quad w_2 &= 1, \\ t_1 w_1 + t_2 w_2 &= 1, \end{aligned}$$

hence,

$$\begin{aligned} w_1 &= \frac{1 - t_2}{t_1 - t_2} = \frac{2 - \sqrt{2}}{4} \quad \left(= \frac{1}{4}\, t_2\right), \\ w_2 &= \frac{t_1 - 1}{t_1 - t_2} = \frac{2 + \sqrt{2}}{4} \quad \left(= \frac{1}{4}\, t_1\right). \end{aligned}$$

Furthermore, after an elementary calculation (using the *Hint*),

$$\begin{aligned} E_2(f) &= \frac{f^{(4)}(\tau)}{4!} \int_0^\infty [\pi_2(t)]^2 e^{-t}dt \\ = \frac{f^{(4)}(\tau)}{4!} \int_0^\infty (t^2 - 4t + 2)^2 e^{-t}dt &= \frac{f^{(4)}(\tau)}{4!} \cdot 4 = \frac{1}{6}\, f^{(4)}(\tau). \end{aligned}$$

(c) The Gauss–Laguerre approximation of I is

$$I \approx \frac{w_1}{t_1+1} + \frac{w_2}{t_2+1} = .03317\ldots + .53825\ldots = .57142\ldots.$$

The true error is $E_2 = .59634\ldots - .57142\ldots = .02491\ldots$, while the estimated error (from (b)) is

$$0 < E_2 = \frac{1}{6}\,\frac{4!}{(1+\tau)^5} < 4.$$

Given the true error $E_2 = .0249187895\ldots$, we have $(1+\tau)^5 = 4/E_2$, so that

$$\tau = \left(\frac{4}{E_2}\right)^{1/5} - 1 = 1.761255\ldots\ .$$

29. Let $w(t) = \mathrm{e}^{-t^2}$ and $\pi_2(t) = \pi_2(t; w)$. Since w is even, so is π_2:

$$\pi_2(t) = t^2 - a.$$

Orthogonality requires

$$\int_{-\infty}^{\infty} \pi_2(t)\mathrm{e}^{-t^2}\mathrm{d}t = 0,$$

that is,

$$\int_{0}^{\infty} (t^2 - a)\mathrm{e}^{-t^2}\mathrm{d}t = 0.$$

Here (by the *Hint* provided),

$$\int_0^\infty \mathrm{e}^{-t^2}\mathrm{d}t = \frac{1}{2}\sqrt{\pi}, \quad \int_0^\infty t^2\mathrm{e}^{-t^2}\mathrm{d}t = \frac{1}{4}\sqrt{\pi}.$$

Therefore,

$$a = \frac{1}{2}, \qquad \pi_2(t) = t^2 - \frac{1}{2},$$

and

$$t_1 = \frac{1}{\sqrt{2}} = .70710678\ldots, \qquad t_2 = -t_1.$$

Letting $f(t) = 1$ and $f(t) = t$ in $\int_{-\infty}^{\infty} f(t)\mathrm{e}^{-t^2}\mathrm{d}t = w_1 f(t_1) + w_2 f(t_2) + E_2(f)$, where $E_2(f) = 0$ in either case, we get for the weights

$$w_1 + w_2 = \sqrt{\pi},$$

$$\frac{1}{\sqrt{2}}\,w_1 - \frac{1}{\sqrt{2}}\,w_2 = 0,$$

that is,

$$w_1 = w_2 = \frac{1}{2}\sqrt{\pi} = .88622692\ldots\ .$$

For the remainder, we have

$$E_2(f) = \frac{f^{(4)}(\tau)}{4!} \int_{-\infty}^{\infty} [\pi_2(t)]^2 \mathrm{e}^{-t^2} \mathrm{d}t, \quad \tau \in \mathbb{R}.$$

Using again the information provided in the *Hint*, we get

$$\int_{-\infty}^{\infty} [\pi_2(t)]^2 \mathrm{e}^{-t^2} \mathrm{d}t = 2 \int_0^{\infty} \left(t^2 - \frac{1}{2}\right)^2 \mathrm{e}^{-t^2} \mathrm{d}t = 2 \int_0^{\infty} \left(t^4 - t^2 + \frac{1}{4}\right) \mathrm{e}^{-t^2} \mathrm{d}t$$

$$= 2 \left(\frac{3}{4} - \frac{1}{2} + \frac{1}{4}\right) \frac{\sqrt{\pi}}{2} = \frac{\sqrt{\pi}}{2},$$

hence

$$E_2(f) = \frac{\sqrt{\pi}}{48} f^{(4)}(\tau), \quad \tau \in \mathbb{R}.$$

30. (a) Use the Gauss quadrature rule and the fact that $\pi_k \cdot \pi_\ell \in \mathbb{P}_{2n-2}$ for $k, \ell = 0, 1, \ldots, n-1$, to obtain

$$0 = \int_a^b \pi_k(t)\pi_\ell(t)w(t)\mathrm{d}t = \sum_{\nu=1}^{n} w_\nu \pi_k(t_\nu)\pi_\ell(t_\nu), \quad k \neq \ell.$$

(b) Apply Gaussian quadrature: since $\ell_i \cdot \ell_k \in \mathbb{P}_{2n-2}$, we get

$$\int_a^b \ell_i(t)\ell_k(t)w(t)\mathrm{d}t = \sum_{\nu=1}^{n} w_\nu \ell_i(t_\nu)\ell_k(t_\nu)$$

$$= \sum_{\nu=1}^{n} w_\nu \delta_{i\nu}\delta_{k\nu} = w_k \delta_{ik} = 0, \quad i \neq k.$$

31. (a) Putting $f(x) = x^r$, $r = 0, 1, 2$, one obtains

$$\begin{aligned} a + b &= 1, \\ bc &= 1, \\ bc^2 &= 2. \end{aligned}$$

Clearly, $bc \neq 0$. Dividing the third equation by the second gives $c = 2$, hence $b = \frac{1}{2}$ and $a = \frac{1}{2}$. Thus,

$$\int_0^{\infty} \mathrm{e}^{-x} f(x)\mathrm{d}x = \frac{1}{2} \left[f(0) + f(2)\right] + E(f).$$

This is the 2-point Gauss–Radau formula for the weight function e^{-x} on $[0,\infty]$.

(b) From the table of divided differences

$$\begin{array}{cc|ccc} x & f & & & \\ \hline 0 & f_0 & & & \\ 2 & f_2 & \frac{1}{2}(f_2 - f_0) & & \\ 2 & f_2 & f_2' & \frac{1}{2}\left(f_2' - \frac{1}{2}(f_2 - f_0)\right) & \end{array}$$

one gets

$$\begin{aligned} p_2(x) &= f_0 + \tfrac{1}{2}(f_2 - f_0)x + \tfrac{1}{2}\left(f_2' - \tfrac{1}{2}(f_2 - f_0)\right)x(x-2) \\ &= \left(1 - x + \tfrac{1}{4}x^2\right) f_0 + \left(x - \tfrac{1}{4}x^2\right) f_2 + \tfrac{1}{2}x(x-2)f_2', \end{aligned}$$

hence

$$\begin{aligned} \int_0^\infty \mathrm{e}^{-x} p_2(x)\mathrm{d}x &= f_0 \int_0^\infty e^{-x}(1 - x + \frac{1}{4}x^2)dx \\ +f_2 \int_0^\infty \mathrm{e}^{-x}(x - \frac{1}{4}x^2)\mathrm{d}x &+ \tfrac{1}{2}f_2' \int_0^\infty \mathrm{e}^{-x}\, x(x-2)\mathrm{d}x \\ = \left(1 - 1 + \tfrac{1}{4}\cdot 2\right) f_0 + \left(1 - \tfrac{1}{4}\cdot 2\right) f_2 &+ \tfrac{1}{2}(2-2)f_2' = \tfrac{1}{2}\, f_0 + \tfrac{1}{2}\, f_2, \end{aligned}$$

which is the same formula as the one obtained in (a).

(c) Using the remainder term for Hermite interpolation, one gets

$$\begin{aligned} E(f) &= \int_0^\infty x(x-2)^2 \frac{f'''(\xi(x))}{6}\, \mathrm{e}^{-x}\mathrm{d}x \\ &= \tfrac{1}{6} f'''(\xi) \int_0^\infty (x^3 - 4x^2 + 4x)\mathrm{e}^{-x}\mathrm{d}x \\ &= \tfrac{1}{6} f'''(\xi)(6 - 4\cdot 2 + 4) = \tfrac{1}{3} f'''(\xi). \end{aligned}$$

32. (a) By contradiction: If there is no such zero, then $\pi_n(t; w)$ is of constant sign (in fact, nonnegative) on $a \le t \le b$, and $\pi_n(t; w) \not\equiv 0$ (since π_n is monic). By orthogonality of π_n to the constant function 1,

$$0 = \int_a^b \pi_n(t; w) w(t)\mathrm{d}t,$$

which is impossible, since the integrand is nonnegative and not identically zero.

(b) By part (a), we have $r_0 \ge 1$. Suppose $r_0 < n$. Then $\pi_n(t; w)\prod_{i=1}^{r_0}(t - t_{k_i}^{(n)})$ is nonnegative on $[a, b]$ and certainly $\not\equiv 0$. On the other hand, since $r_0 < n$, by orthogonality,

$$\int_a^b \pi_n(t; w) \prod_{i=1}^{r_0}(t - t_{k_i}^{(n)}) w(t)\mathrm{d}t = 0,$$

which again is impossible. Hence, $r_0 = n$, and all zeros are necessarily simple.

33. Let the two sets of zeros be respectively $\tau_\nu^{(n)}$ and $\tau_\mu^{(n+1)}$. From the recurrence relation

$$\pi_{n+1}(t) = (t - \alpha_n)\pi_n(t) - \beta_n \pi_{n-1}(t)$$

one gets

$$(*) \qquad \pi_{n+1}(\tau_\nu^{(n)}) = -\beta_n \pi_{n-1}(\tau_\nu^{(n)}), \quad \nu = 1, 2, \ldots, n.$$

If $n = 1$, this shows that $\pi_2(\tau_1^{(1)}) = -\beta_1 < 0$, hence π_2 must vanish at a point to the right of $\tau_1^{(1)}$ and at a point to the left (since $\pi_2(t)$ tends to $+\infty$ as $t \to \pm\infty$). The assertion, therefore, is true for $n = 1$. Using induction, assume it is true up to degree n. Then from a graph of π_{n-1} and π_n it is clear that

$$\operatorname{sgn} \pi_{n-1}(\tau_\nu^{(n)}) = (-1)^{\nu-1}, \quad \nu = 1, 2, \ldots, n.$$

Therefore, by (*) above, since $\beta_n > 0$,

$$\operatorname{sgn} \pi_{n+1}(\tau_\nu^{(n)}) = (-1)^{\nu}, \quad \nu = 1, 2, \ldots, n.$$

From this, interlacing follows readily.

34. (a) We must have

$$h_\nu(\tau_\nu) = 1, \qquad h_\nu'(\tau_\nu) = 0,$$

the other required equations $h_\nu(\tau_\mu) = h_\nu'(\tau_\mu) = 0$, $\mu \neq \nu$, being automatically satisfied in view of the proposed form of h_ν. Thus,

$$a_\nu + b_\nu \tau_\nu = 1, \qquad b_\nu + (a_\nu + b_\nu \tau_\nu) \cdot 2\ell_\nu(\tau_\nu)\ell_\nu'(\tau_\nu) = 0,$$

that is,

$$a_\nu + b_\nu \tau_\nu = 1,$$
$$b_\nu + 2\ell_\nu'(\tau_\nu) = 0.$$

Solving this for a_ν and b_ν and inserting the result in the proposed form of h_ν gives

$$h_\nu(t) = (1 - 2(t - \tau_\nu)\ell_\nu'(\tau_\nu))\ell_\nu^2(t), \quad \nu = 1, 2, \ldots, n.$$

Likewise, k_ν must satisfy

$$k_\nu(\tau_\nu) = 0, \qquad k_\nu'(\tau_\nu) = 1,$$

giving

$$c_\nu + d_\nu \tau_\nu = 0, \qquad d_\nu + (c_\nu + d_\nu \tau_\nu) \cdot 2\ell_\nu(\tau_\nu)\ell_\nu'(\tau_\nu) = 1,$$

that is,

$$c_\nu + d_\nu \tau_\nu = 0, \quad d_\nu = 1.$$

Thus, $c_\nu = -\tau_\nu$, and

$$k_\nu(t) = (t-\tau_\nu)\ell_\nu^2(t), \qquad \nu = 1, 2, \ldots, n.$$

Note that

$$\ell_\nu'(\tau_\nu) = \sum_{\mu\neq\nu} \frac{1}{\tau_\nu - \tau_\mu}\,.$$

(b) The quadrature rule in question is

$$\int_a^b f(t)w(t)\mathrm{d}t = \int_a^b p(t)w(t)\mathrm{d}t + E_n(f),$$

which clearly has degree of exactness $2n-1$. Using (a), we get

$$\begin{aligned}\int_a^b p(t)w(t)\mathrm{d}t &= \int_a^b \sum_{\nu=1}^n (h_\nu(t)f_\nu + k_\nu(t)f_\nu')w(t)\mathrm{d}t \\ &= \sum_{\nu=1}^n \left(f_\nu \int_a^b h_\nu(t)w(t)\mathrm{d}t + f_\nu' \int_a^b k_\nu(t)w(t)\mathrm{d}t \right).\end{aligned}$$

Thus,

$$\lambda_\nu = \int_a^b h_\nu(t)w(t)\mathrm{d}t, \quad \mu_\nu = \int_a^b k_\nu(t)w(t)\mathrm{d}t, \quad \nu = 1, 2, \ldots, n.$$

(c) For all μ_ν to be zero, we must have

$$\int_a^b k_\nu(t)w(t)\mathrm{d}t = 0, \qquad \nu = 1, 2, \ldots, n,$$

or, by the results in (a), noting that $\ell_\nu(t) = \frac{\omega_n(t)}{(t-\tau_\nu)\omega_n'(\tau_\nu)}$,

$$\frac{1}{\omega_n'(\tau_\nu)} \int_a^b \frac{\omega_n(t)}{(t-\tau_\nu)\,\omega_n'(\tau_\nu)}\omega_n(t)w(t)\mathrm{d}t = 0, \quad \nu = 1, 2, \ldots, n,$$

that is,

$$\int_a^b \ell_\nu(t)\omega_n(t)w(t)\mathrm{d}t = 0, \qquad \nu = 1, 2, \ldots, n.$$

Since $\{\ell_\nu(t)\}_{\nu=1}^n$ forms a basis of $\mathbb{P}_{n-1}$ (the ℓ_ν are linearly independent and span $\mathbb{P}_{n-1}$), ω_n must be orthogonal with respect to the weight function w to all polynomials of lower degree, i.e., $\omega_n(t) = \pi_n(t; w)$. We get the Gauss quadrature rule.

35. Following the *Hint*, put $1-t=x^2,\ \mathrm{d}t=-2x\mathrm{d}x$. Then

$$\int_0^1 (1-t)^{-1/2} f(t)\mathrm{d}t = -\int_1^0 \frac{1}{x}\, f(1-x^2)\cdot 2x\mathrm{d}x$$
$$= 2\int_0^1 f(1-x^2)\mathrm{d}x.$$

Here, the last integral can be accurately computed by Gauss–Legendre quadrature on $[0,1]$, since the integrand is smooth.

36. The area of the unit disk is

$$A = 2\int_{-1}^1 (1-t^2)^{1/2}\mathrm{d}t = 2\int_{-1}^1 (1-t^2)(1-t^2)^{-1/2}\mathrm{d}t,$$

hence can be evaluated exactly by the 2-point Gauss–Chebyshev quadrature rule applied to $f(t)=1-t^2$:

$$A = 2\cdot\frac{\pi}{2}\,(1-t_1^2+1-t_2^2) = \pi(2-t_1^2-t_2^2).$$

But

$$t_1 = \cos\frac{\pi}{4} = \frac{1}{\sqrt{2}}\,,\quad t_2 = \cos\frac{3\pi}{4} = -\frac{1}{\sqrt{2}}\,.$$

Thus,

$$A = \pi(2-\tfrac{1}{2}-\tfrac{1}{2}) = \pi.$$

37. (a) Put $x=\frac{1-t}{2}\,a+\frac{1+t}{2}\,b,\ \mathrm{d}x=\frac{b-a}{2}\,\mathrm{d}t$. Then

$$\int_a^b f(x)\mathrm{d}x = \frac{b-a}{2}\int_{-1}^1 f\left(\frac{1-t}{2}\,a+\frac{1+t}{2}\,b\right)\mathrm{d}t$$
$$\approx \frac{b-a}{2}\sum_{\nu=1}^n w_\nu^G f\left(\frac{1-t_\nu^G}{2}\,a+\frac{1+t_\nu^G}{2}\,b\right),$$

where t_ν^G and w_ν^G are the standard Gauss–Legendre nodes and weights. The new nodes and weights, therefore, are

$$t_\nu = \frac{1-t_\nu^G}{2}\,a+\frac{1+t_\nu^G}{2}\,b,\quad w_\nu = \frac{b-a}{2}\,w_\nu^G,\quad \nu=1,2,\ldots,n.$$

(b) Put $x=1+\frac{t}{a},\ \mathrm{d}x=\frac{1}{a}\,\mathrm{d}t$. Then

$$\int_1^\infty \mathrm{e}^{-ax} f(x)\mathrm{d}x = \frac{1}{a}\int_0^\infty \mathrm{e}^{-a-t} f\left(1+\frac{t}{a}\right)\mathrm{d}t$$
$$= \frac{\mathrm{e}^{-a}}{a}\int_0^\infty \mathrm{e}^{-t} f\left(1+\frac{t}{a}\right)\mathrm{d}t \approx \frac{\mathrm{e}^{-a}}{a}\sum_{\nu=1}^n w_\nu^L f\left(1+\frac{t_\nu^L}{a}\right),$$

where t_ν^L and w_ν^L are the standard Gauss–Laguerre nodes and weights. The new nodes and weights, therefore, are

$$t_\nu = 1 + \frac{t_\nu^L}{a}\,, \quad w_\nu = \frac{\mathrm{e}^{-a}}{a}\, w_\nu^L, \quad \nu = 1, 2, \ldots, n.$$

(c) Following the *Hint*, we write

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

Then

$$\int_{-\infty}^{\infty} \mathrm{e}^{-(ax^2+bx)} f(x)\mathrm{d}x = \mathrm{e}^{b^2/4a} \int_{-\infty}^{\infty} \mathrm{e}^{-a(x+b/2a)^2} f(x)\mathrm{d}x.$$

Now put $x + \frac{b}{2a} = \frac{t}{\sqrt{a}}$, $\mathrm{d}x = \frac{1}{\sqrt{a}}\,\mathrm{d}t$, to get

$$\begin{aligned}\int_{-\infty}^{\infty} \mathrm{e}^{-(ax^2+bx)} f(x)\mathrm{d}x &= \frac{\mathrm{e}^{b^2/4a}}{\sqrt{a}} \int_{-\infty}^{\infty} \mathrm{e}^{-t^2} f\left(\frac{t}{\sqrt{a}} - \frac{b}{2a}\right)\mathrm{d}t \\ &\approx \frac{\mathrm{e}^{b^2/4a}}{\sqrt{a}} \sum_{\nu=1}^{n} w_\nu^H f\left(\frac{t_\nu^H}{\sqrt{a}} - \frac{b}{2a}\right),\end{aligned}$$

where t_ν^H and w_ν^H are the standard Gauss–Hermite nodes and weights. The new nodes and weights, therefore, are

$$t_\nu = \frac{t_\nu^H}{\sqrt{a}} - \frac{b}{2a}\,, \quad w_\nu = \frac{\mathrm{e}^{b^2/4a}}{\sqrt{a}}\, w_\nu^H, \quad \nu = 1, 2, \ldots, n.$$

(d) With the transformation $xt = \tau$ one gets

$$\int_0^\infty \frac{\mathrm{e}^{-xt}}{y+t}\,\mathrm{d}t = \int_0^\infty \frac{\mathrm{e}^{-\tau}}{y + \tau/x}\,\frac{\mathrm{d}\tau}{x} = \int_0^\infty \frac{\mathrm{e}^{-\tau}}{xy+\tau}\,\mathrm{d}\tau \approx \sum_{\nu=1}^{n} \frac{w_\nu^L}{xy + t_\nu^L}\,,$$

where t_ν^L and w_ν^L are the Gauss–Laguerre nodes and weights. Letting $f(t) = (xy+t)^{-1}$, one has

$$f^{(2n)}(t) = (2n)!(xy+t)^{-(2n+1)} > 0 \ \text{ on } \ [0, \infty],$$

hence for the error (integral minus sum)

$$\frac{f^{(2n)}(\tau)}{(2n)!} \int_0^\infty \pi_n^2(t)\mathrm{e}^{-t}\mathrm{d}t > 0, \quad \tau > 0,$$

where π_n is the monic Laguerre polynomial of degree n. Thus, the approximation is too small.

38. (a) Let $\pi_n(t) = \pi_n(t; w)$ and define $p_n(t) := (-1)^n \pi_n(-t)$. Then, for each n, the polynomial p_n is monic of degree n, and

$$\begin{aligned}\int_{-1}^{1} p_k(t)p_\ell(t)w(t)\mathrm{d}t &= (-1)^{k+\ell}\int_{-1}^{1} \pi_k(-t)\pi_\ell(-t)w(t)\mathrm{d}t \\ &= (-1)^{k+\ell}\int_{-1}^{1} \pi_k(\tau)\pi_\ell(\tau)w(-\tau)\mathrm{d}\tau \\ &= (-1)^{k+\ell}\int_{-1}^{1} \pi_k(\tau)\pi_\ell(\tau)w(\tau)\mathrm{d}\tau \\ &= 0 \quad \text{if} \quad k \neq \ell.\end{aligned}$$

By uniqueness of monic orthogonal polynomials, we get $p_n(t) \equiv \pi_n(t; w)$, as was to be shown.

(b) The assertion for the nodes is an immediate consequence of (a). For the weights, we have, since $(-1)^{n+1}\pi_n'(-t) = \pi_n'(t)$,

$$\begin{aligned}w_{n+1-\nu} &= \int_a^b \frac{\pi_n(t)w(t)\mathrm{d}t}{(t - t_{n+1-\nu})\pi_n'(t_{n+1-\nu})} = \int_a^b \frac{\pi_n(-t)w(t)\mathrm{d}t}{(-t - t_{n+1-\nu})\pi_n'(-t_\nu)} \\ &= (-1)^{n+1}\int_a^b \frac{\pi_n(t)w(t)\mathrm{d}t}{(t - t_\nu)(-1)^{n+1}\pi_n'(t_\nu)} = w_\nu.\end{aligned}$$

(c) By (a), the orthogonal polynomial $\pi_2(t) = \pi_2(t; w)$ is even, say, $\pi_2(t) = t^2 - c$. Orthogonality to the constant 1 gives

$$\begin{aligned}0 &= \int_{-1}^{1} \pi_2(t)w(t)\mathrm{d}t = 2\int_0^1 \pi_2(t)w(t)\mathrm{d}t \\ &= 2\int_0^1 (t^2 - c)(1 - t)\mathrm{d}t = 2\left(\frac{1}{3} - \frac{1}{4} - \frac{1}{2}\,c\right) = \frac{1}{6} - c,\end{aligned}$$

that is, $c = \frac{1}{6}$. Therefore, $t_1 = -t_2 = \sqrt{\frac{1}{6}}$. By symmetry, $w_1 = w_2$, and putting $f \equiv 1$ in the quadrature formula yields $2w_1 = \int_{-1}^{1} w(t)\mathrm{d}t = 1$, i.e., $w_1 = w_2 = \frac{1}{2}$.

If we assume $f \in C^4[-1, 1]$, the error term is

$$E_2(f) = \frac{f^{(4)}(\tau)}{4!}\int_{-1}^{1} \pi_2^2(t)w(t)\mathrm{d}t, \quad \tau \in (-1, 1).$$

Here, the integral on the right is

$$2\int_0^1 \pi_2^2(t)w(t)\mathrm{d}t = 2\int_0^1 \left(t^2 - \frac{1}{6}\right)^2 (1 - t)\mathrm{d}t = \frac{7}{180},$$

giving

$$E_2(f) = \frac{7}{4320}\, f^{(4)}(\tau).$$

39. For any $P \in \mathbb{P}_{4n-1}$, one has

$$\int_{-1}^{1} P(\tau)\mathrm{d}\tau = \sum_{k=1}^{2n} w_k^{(2n)} P(t_k^{(2n)}).$$

In particular, taking $P(\tau) = p(\tau^2)$ with $p \in \mathbb{P}_{2n-1}$, and using the symmetry of the Gauss formula (cf. Ex. 38(b)), we have

$$\int_{-1}^{1} p(\tau^2)\mathrm{d}\tau = 2\sum_{k=1}^{n} w_k^{(2n)} p([t_k^{(2n)}]^2).$$

The assertion now follows by making the transformation of variables $\tau^2 = t$ in the integral on the left,

$$\int_{-1}^{1} p(\tau^2)\mathrm{d}\tau = 2\int_{0}^{1} p(\tau^2)\mathrm{d}\tau = 2\int_{0}^{1} p(t)\frac{\mathrm{d}t}{2\sqrt{t}} = \int_{0}^{1} t^{-1/2}p(t)\mathrm{d}t.$$

40. Put $\theta = \pi t$. Then

$$I_{\alpha,\beta}(f) = \pi\int_{0}^{1} f(\pi t)[\cos(\tfrac{1}{2}\pi t)]^{\alpha}\,[\sin(\tfrac{1}{2}\pi t)]^{\beta}\mathrm{d}t.$$

Since

$$\cos(\tfrac{1}{2}\pi t) \sim \tfrac{1}{2}\pi(1-t) \quad \text{as} \quad t \to 1,$$
$$\sin(\tfrac{1}{2}\pi t) \sim \tfrac{1}{2}\pi t \quad \text{as} \quad t \to 0,$$

write

$$I_{\alpha,\beta}(f) = \pi\int_{0}^{1} f(\pi t)\left[\frac{\cos(\frac{1}{2}\pi t)}{1-t}\right]^{\alpha}\left[\frac{\sin(\frac{1}{2}\pi t)}{t}\right]^{\beta}(1-t)^{\alpha}t^{\beta}\mathrm{d}t$$

and use Gauss–Jacobi quadrature on the interval $[0,1]$ with weight function $w(t) = (1-t)^{\alpha}t^{\beta}$. (The remaining part of the integrand is now smooth.)

41. We have, by the triangle inequality,

$$\begin{aligned}|Q_n f - Q_{n^*} f| &= |(Q_n f - If) - (Q_{n^*} f - If)| \\ &\geq |Q_n f - If| - |Q_{n^*} f - If|,\end{aligned}$$

which by assumption is

$$\begin{aligned}&\geq 2|Q_{n^*} f - If| - |Q_{n^*} f - If| \\ &= |Q_{n^*} f - If|.\end{aligned}$$

42. (a) We have

$$E_n\left(\frac{1}{t\pm x}\right) = \int_{-1}^{1} \frac{1}{t\pm x}\, w(t)\mathrm{d}t - \sum_{k=1}^{n}[x^2-(t_k^G)^2]w_k^G \frac{1}{t_k^G \pm x}$$
$$= \int_{-1}^{1} (\pm x - t)\, \frac{w(t)}{x^2-t^2}\,\mathrm{d}t - \sum_{k=1}^{n} w_k^G(\pm x - t_k^G).$$

Since $\pm x - t \in \mathbb{P}_1$ and (G) is certainly exact for linear functions, the above is equal to zero.

(b) We have, if $n \geq 2$,

$$E_n(g) = \int_{-1}^{1} g(t)w(t)\mathrm{d}t - \sum_{k=1}^{n} w_k g(t_k)$$
$$= \int_{-1}^{1} [g(t)(x^2-t^2)]\, \frac{w(t)}{x^2-t^2}\,\mathrm{d}t - \sum_{k=1}^{n} w_k^G[g(t_k^G)(x^2-(t_k^G)^2)],$$

which by (G) vanishes since $g(t)(x^2-t^2) \in \mathbb{P}_{2n-1}$.

43. One has, for $\nu = 1, 2, \ldots, 2n$,

$$\int_{-1}^{1} \frac{w(x)}{1+\xi_\nu x}\,\mathrm{d}x = \int_{-1}^{1} \frac{\prod_{\substack{\mu=1\\ \mu\neq\nu}}^{2n}(1+\xi_\mu x)}{\omega_{2n}(x)}\, w(x)\mathrm{d}x$$

$$= \int_{-1}^{1} \prod_{\substack{\mu=1\\ \mu\neq\nu}}^{2n} (1+\xi_\mu x)\, w^*(x)\mathrm{d}x \quad \text{(by definition of } w^*)$$
$$= \sum_{k=1}^{n} w_k^G \prod_{\substack{\mu=1\\ \mu\neq\nu}}^{2n} (1+\xi_\mu x_k^G) \quad \left(\text{since } \prod_{\substack{\mu=1\\ \mu\neq\nu}}^{2n} (1+\xi_\mu x) \in \mathbb{P}_{2n-1}\right)$$
$$= \sum_{k=1}^{n} \frac{w_k^G\, \omega_{2n}(x_k^G)}{1+\xi_\nu x_k^G}$$
$$= \sum_{k=1}^{n} \frac{w_k^*}{1+\xi_\nu x_k^*} \quad \text{(by definition of } x_k^* \text{ and } w_k^*),$$

as was to be shown.

44. (a) Let $\tilde{\pi}_k$ be the orthonormal polynomials relative to the weight function w and $\tilde{\boldsymbol{\pi}}(t) = [\tilde{\pi}_0(t), \tilde{\pi}_1(t), \ldots, \tilde{\pi}_{n-1}]^T$. Then the three-term recurrence relation for $\tilde{\pi}_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_{n-1}$ (cf. Chap. 2, Ex. 21(a)) can be written in matrix form as

$$t\tilde{\boldsymbol{\pi}}(t) = \boldsymbol{J}\tilde{\boldsymbol{\pi}}(t) + \sqrt{\beta_n}\tilde{\pi}_n(t)\boldsymbol{e}_n, \quad \boldsymbol{e}_n^T = [0, 0, \ldots, 1],$$

where $\boldsymbol{J} = \boldsymbol{J}_n(w)$ is the Jacobi matrix defined in Sect. 3.2.3(e). Since t_k is a zero of $\tilde{\pi}_n$, putting here $t = t_k$ yields

$$t_k \tilde{\boldsymbol{\pi}}(t_k) = \boldsymbol{J} \tilde{\boldsymbol{\pi}}(t_k),$$

showing not only that t_k is an eigenvalue of $\boldsymbol{J}$, but also that $\tilde{\boldsymbol{\pi}}(t_k)$ is a corresponding eigenvector. (Note that $\tilde{\boldsymbol{\pi}}(t_k) \neq \boldsymbol{0}$ since the first component is $\tilde{\pi}_0 = 1/\sqrt{\beta_0} \neq 0$.)

(b) If we put $x = t_k$ in the Christoffel–Darboux formula (see Chap. 2, Ex. 21(b), where n needs to be replaced by $n-1$), we get

$$(*) \qquad \sum_{\nu=0}^{n-1} \tilde{\pi}_\nu(t_k)\tilde{\pi}_\nu(t) = \sqrt{\beta_n}\, \frac{\tilde{\pi}_{n-1}(t_k)\tilde{\pi}_n(t)}{t - t_k}\,.$$

Integrating with respect to the weight function w and making use of the orthogonality of $\tilde{\pi}_\nu$, $\nu \geq 1$, to 1, and the orthonormality of $\tilde{\pi}_0$, we find

$$1 = \sqrt{\beta_n}\, \tilde{\pi}_{n-1}(t_k) \int_a^b \frac{\tilde{\pi}_n(t)}{t - t_k}\, w(t)\mathrm{d}t.$$

Therefore, since $w_k = \int_a^b [\tilde{\pi}_n(t)/(t-t_k)\tilde{\pi}_n'(t_k)]w(t)\mathrm{d}t$ (cf. (3.39), (3.40)),

$$w_k = \frac{1}{\sqrt{\beta_n}\, \tilde{\pi}_{n-1}(t_k)\tilde{\pi}_n'(t_k)}\,.$$

Letting $t \to t_k$ in (*), on the other hand, gives

$$\sum_{\nu=0}^{n-1} [\tilde{\pi}_\nu(t_k)]^2 = \sqrt{\beta_n}\, \tilde{\pi}_{n-1}(t_k)\tilde{\pi}_n'(t_k),$$

so that

$$w_k = \frac{1}{\sum\limits_{\nu=0}^{n-1} [\tilde{\pi}_\nu(t_k)]^2}\,.$$

Now the normalized eigenvectors $\boldsymbol{v}_k$ of the Jacobi matrix $\boldsymbol{J}$, according to (a), are

$$\boldsymbol{v}_k = \frac{\tilde{\boldsymbol{\pi}}(t_k)}{\sqrt{\sum\limits_{\nu=0}^{n-1} [\tilde{\pi}_\nu(t_k)]^2}} = \tilde{\boldsymbol{\pi}}(t_k)\sqrt{w_k},$$

the first component thereof giving

$$v_{k,1} = \tilde{\pi}_0(t_k)\sqrt{w_k} = \sqrt{w_k}/\sqrt{\beta_0}.$$

Squaring both sides yields (3.47).

45. We show that

$$(I^{p+1}g)(s) := \int_0^s \mathrm{d}s_1 \int_0^{s_1} \mathrm{d}s_2 \cdots \mathrm{d}s_p \int_0^{s_p} g(t)\mathrm{d}t = \int_0^s \frac{(s-t)^p}{p!}\, g(t)\mathrm{d}t.$$

It suffices to show that both sides satisfy the initial value problem

$$y^{(p+1)}(s) = g(s), \quad y(0) = y'(0) = \cdots = y^{(p)}(0) = 0.$$

By uniqueness of the solution, the assertion then follows. Putting $s = 1$ gives the stated result, (3.50).

For the left-hand side, the above initial value problem is trivially true. Denote the right-hand side by $y(s)$. Then the assertion is trivially true for $p = 0$, and for $p > 0$ one has

$$y^{(k)}(s) = \int_0^s \frac{(s-t)^{p-k}}{(p-k)!}\, g(t)\mathrm{d}t, \quad k = 0, 1, \ldots, p.$$

From this, the assertion follows immediately.

46. (a) Putting in turn $y(s) = 1$, $y(s) = s$, $y(s) = s^2$, one gets

$$\begin{aligned} a \;+\; b \qquad &= 1, \\ b \qquad &= \tfrac{1}{2}, \\ b \;-2c &= \tfrac{1}{3}, \end{aligned}$$

hence $a = \frac{1}{2},\ b = \frac{1}{2},\ c = \frac{1}{12}$, that is,

$$\int_0^1 y(s)\mathrm{d}s \approx \frac{1}{2}\,[y(0) + y(1)] - \frac{1}{12}\,[y'(1) - y'(0)].$$

(b) The transformation $t = x + hs$, $\mathrm{d}t = h\mathrm{d}s$ yields

$$\begin{aligned} \int_x^{x+h} f(t)\mathrm{d}t = h\int_0^1 f(x+hs)\mathrm{d}s \approx\ & \frac{h}{2}\,[f(x) + f(x+h)] \\ & -\frac{h^2}{12}\,[f'(x+h) - f'(x)]. \end{aligned}$$

(c) Letting $h = (b-a)/n$, $x_k = a + kh$, $f_k = f(x_k)$, $f_k' = f'(x_k)$, $k = 0, 1, \ldots, n$, one finds, using (b), that

$$\begin{aligned} \int_a^b f(t)\mathrm{d}t &= \sum_{k=0}^{n-1} \int_{x_k}^{x_k+h} f(t)\mathrm{d}t \approx \frac{h}{2}\,[(f_0 + f_1) + (f_1 + f_2) + \cdots \\ &+ (f_{n-1} + f_n)] - \frac{h^2}{12}\,[(f_1' - f_0') + (f_2' - f_1') + \cdots + (f_n' - f_{n-1}')] \\ &= h\left(\frac{1}{2}\,f_0 + f_1 + \cdots + f_{n-1} + \frac{1}{2}\,f_n\right) - \frac{h^2}{12}\,[f'(b) - f'(a)]. \end{aligned}$$

This may be interpreted as a composite trapezoidal rule with "end correction". The correction approximates the error term of the composite trapezoidal rule:

$$-\frac{b-a}{12}\,h^2 f''(\xi) \approx -\frac{b-a}{12}\,h^2\,\frac{f'(b)-f'(a)}{b-a} = -\frac{h^2}{12}\,[f'(b)-f'(a)].$$

47. Taking for $f(t)$ in turn $1,\ t,\ t^2$ gives

$$\begin{aligned} \alpha_{-1} \quad + \quad 2\alpha_0 \quad + \quad \alpha_1 &= 4, \\ -\alpha_{-1} \qquad\quad + \quad \alpha_1 &= 0, \\ \alpha_{-1} \qquad\quad + \quad \alpha_1 &= \tfrac{16}{7}\,. \end{aligned}$$

The solution is

$$\alpha_{-1} = \alpha_1 = \frac{8}{7}, \quad \alpha_0 = \frac{6}{7}\,.$$

Since $E(t^3) = 0$ and $E(t^4) = \frac{2}{5} - 2 \cdot \frac{8}{7} \cdot \frac{1}{5} \cdot \frac{31}{32} = -\frac{3}{70} \neq 0$, we have $d = 3$.

48. (a) For the piece $p(x) = s_2(x)|_{[0,1]}$ of the spline we must have $p(0) = p'(0) = 0$, $p(1) = 1$, which is a (polynomial) Hermite interpolation problem with solution $p(x) = x^2$. Thus,

$$s_2(x) = \begin{cases} 0 & \text{if} \quad -1 \le x \le 0, \\ x^2 & \text{if} \quad 0 \le x \le 1. \end{cases}$$

(b) Clearly, s is a quadratic spline on $[-1, 1]$ (with knot at $x = 0$). The conditions imposed on s amount to

$$\begin{aligned} c_0 \quad - \quad c_1 \quad + \quad c_2 \qquad\quad &= f_{-1}, \\ c_0 \qquad\quad\quad\quad\quad\quad &= f_0, \\ \qquad\quad c_1 \qquad\quad\quad &= f_0', \\ c_0 \quad + \quad c_1 \quad + \quad c_2 \quad + \quad c_3 &= f_1. \end{aligned}$$

This yields immediately

$$c_0 = f_0, \quad c_1 = f_0', \quad c_2 = f_{-1} - f_0 + f_0', \quad c_3 = f_1 - f_{-1} - 2f_0',$$

hence

$$s(x) = f_0 + f_0' x + (f_{-1} - f_0 + f_0')x^2 + (f_1 - f_{-1} - 2f_0')x_+^2.$$

(c) Integrating the spline $s(x)$ obtained in (b) gives

$$\begin{aligned}\int_{-1}^{1} s(x)\mathrm{d}x &= 2f_0 + \tfrac{2}{3}\,(f_{-1} - f_0 + f_0') + \tfrac{1}{3}\,(f_1 - f_{-1} - 2f_0') \\ &= \tfrac{1}{3}\,(f_{-1} + 4f_0 + f_1),\end{aligned}$$

that is, Simpson's rule.

49. Let

$$\hat{\varphi}_j = \sum_{k=1}^{n} c_{jk}\varphi_k, \quad j = 1, 2, \ldots, n$$

be another basis, hence $\det \boldsymbol{C} \neq 0$, where $\boldsymbol{C} = [c_{jk}]$. Then

$$\hat{\boldsymbol{G}} = [L_i\hat{\varphi}_j] = \left[\sum_{k=1}^{n} c_{jk}L_i\varphi_k\right] = \boldsymbol{C}\boldsymbol{G}^T,$$

hence

$$\det \hat{\boldsymbol{G}} = \det \boldsymbol{C} \cdot \det \boldsymbol{G}^T = \det \boldsymbol{C} \cdot \det \boldsymbol{G} \neq 0.$$

50. If $t > b$, then $(x-t)_+^r = 0$ for $x \in [a, b]$, hence $K_r(t) = 0$. If $t < a$, then $(x-t)_+^r = (x-t)^r$ for $x \in [a, b]$, and $K_r(t)$ vanishes because $r \leq d$ and E by assumption annihilates polynomials of degree d.

51. The Peano representation of E is

$$Ef = \int_a^b K_r(t) f^{(r+1)}(t)\mathrm{d}t,$$

where the Peano kernel $K_r(t)$ is continuous on $[a, b]$ by assumption. Suppose, for definiteness, that $e_{r+1} > 0$. We claim that $K_r(t) \geq 0$ on (a, b). Suppose not. Then there exists $t_0 \in (a, b)$ such that $K_r(t_0) < 0$, and, by continuity, an interval I around t_0 in which $K_r(t) < 0$. Let $h(t)$ be the hat function which is identically zero outside of I and has the value (say) 1 at t_0. Define f by $f^{(r+1)}(t) = h(t)$. Evidently, $f \in C^{r+1}[a, b]$. Then, for this particular f,

$$Ef = \int_I K_r(t)h(t)\mathrm{d}t < 0$$

by construction. This contradicts $Ef = e_{r+1}f^{(r+1)}(\overline{t}) \geq 0$ and establishes our claim. The reasoning in the case $e_{r+1} < 0$ is analogous.

52. By contradiction. Let $0 \leq \delta \leq d-1$, and suppose that K_δ is definite. Then, for any $f \in C^{\delta+1}[a, b]$,

$$E(f) = \int_a^b K_\delta(t) f^{(\delta+1)}(t)\mathrm{d}t = f^{(\delta+1)}(\tau)\int_a^b K_\delta \mathrm{d}t, \quad \tau \in (a, b),$$

where

$$\int_a^b K_\delta(t)\mathrm{d}t = E\left(\frac{x^{\delta+1}}{(\delta+1)!}\right) = 0,$$

since $\delta+1 \le d$ and the formula has degree of exactness d. This contradicts the definiteness of K_δ.

53. (a) Since by assumption $f^{(r+1)}$ is continuous only for $r=0$, we must compute the Peano kernel K_0 for the functional E. Assuming $0<t<1$, we have

$$\begin{aligned} K_0(t) &= \int_0^1 \sqrt{x}(x-t)_+^0\mathrm{d}x + \frac{2}{15}\cdot 0 - \frac{4}{5}\int_0^1 (x-t)_+^0\mathrm{d}x \\ &= \int_t^1 \sqrt{x}\,\mathrm{d}x - \frac{4}{5}\int_t^1 \mathrm{d}x = \frac{2}{3}\,x^{\frac{3}{2}}\Big|_t^1 - \frac{4}{5}(1-t) \\ &= \frac{2}{3}(1-t^{\frac{3}{2}}) - \frac{4}{5}(1-t) \\ &= -\frac{2}{15}(1-6t+5t^{\frac{3}{2}}). \end{aligned}$$

Therefore,

$$E(f) = -\frac{2}{15}\int_0^1 (1-6t+5t^{3/2})f'(t)\mathrm{d}t.$$

(b) Let $q(t) = 1-6t+5t^{\frac{3}{2}}$. Then, $q(0)=1$, $q(1)=0$, $q'(t) = -6+\frac{15}{2}t^{\frac{1}{2}}$. Since $q'(1)>0$, the function q changes sign on (0,1). In fact, q decreases from 1 to a negative value and then increases from there on to zero. Thus, q has exactly one zero, $t=t_0$, in (0,1). To compute it, let $t=u^2$, $0<u<1$. Then

$$q(u^2) = 1-6u^2+5u^3 = (u-1)(5u^2-u-1),$$

and this has exactly one zero in (0,1) given by $u_0 = (1+\sqrt{21})/10$. Therefore, $t_0 = u_0^2 = (11+\sqrt{21})/50 = .3116515138\ldots$. It now follows that

$$|Ef| \le \int_0^1 |K_0(t)|\,|f'(t)|\mathrm{d}t \le c_0\|f'\|_\infty,$$

where

$$\begin{aligned} c_0 &= \int_0^1 |K_0(t)|\mathrm{d}t = \frac{2}{15}\left(\int_0^{t_0} q(t)\mathrm{d}t - \int_{t_0}^1 q(t)\mathrm{d}t\right) \\ &= \frac{2}{15}\cdot 2(t-3t^2+2t^{\frac{5}{2}})_{t=t_0} = .03432397564\ldots . \end{aligned}$$

54. See the text.

55. Let, as usual, $x_k = a + kh$, $h = (b-a)/n$. Then

$$Ef = \sum_{k=0}^{n-1} \left\{ \int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x - \frac{h}{2}[f(x_k) + f(x_{k+1})] \right\}.$$

Let $K_{1,k}$ be the Peano kernel of

$$E_k f = \int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x - \frac{h}{2}[f(x_k) + f(x_{k+1})].$$

Then, if $x_k < t < x_{k+1}$,

$$\begin{aligned} K_{1,k}(t) = E_k(x-t)_+ &= \int_t^{x_{k+1}} (x-t)\mathrm{d}x - \frac{h}{2}(x_{k+1} - t) \\ &= \tfrac{1}{2}(x_{k+1}-t)^2 - \tfrac{1}{2}h(x_{k+1}-t) = \tfrac{1}{2}(x_{k+1}-t)(x_k - t) \\ &= -\tfrac{1}{2}h^2\tau(1-\tau) \quad \text{if } t = x_k + \tau h \end{aligned}$$

and

$$K_{1,k}(t) = 0 \quad \text{if } t \notin [x_k, x_{k+1}].$$

Since $Ef = \sum_{k=0}^{n-1} E_k f$, one gets

$$K_1(t) = \sum_{k=0}^{n-1} K_{1,k}(t) = -\frac{1}{2}h^2\tau(1-\tau) \quad \text{if } t = x_i + \tau h,\ 0 \le \tau \le 1.$$

Thus, $K_1(t)$ is made up of a sequence of identical parabolic arcs over the subintervals of $[a, b]$.

56. See the text.

57. (a) The formula is exact for quadratic polynomials if it is exact for the first three powers of x,

$$\begin{aligned} a \quad + \quad b \qquad\quad &= 1, \\ b \qquad\quad &= \tfrac{1}{2}, \\ b \quad + \quad 2c &= \tfrac{1}{3}, \end{aligned}$$

that is, if

$$a = \frac{1}{2}, \quad b = \frac{1}{2}, \quad c = -\frac{1}{12}.$$

To get exactness for $f(x) = x^3$ requires $b + 6c\gamma = \frac{1}{4}$, giving

$$\gamma = \frac{1}{2}.$$

Since, with these choices of the parameters,

$$E(x^4) = \frac{1}{5} - \frac{1}{2} + \frac{1}{4} = -\frac{1}{20} \neq 0,$$

the precise degree of exactness is $d = 3$.

(b) With

$$E(f) = \int_0^1 f(x)\mathrm{d}x - \frac{1}{2}\,f(0) - \frac{1}{2}\,f(1) + \frac{1}{12}\,f''(\frac{1}{2})$$

we have

$$6K_3(t) = E_{(x)}((x-t)_+^3) = \int_t^1 (x-t)^3\mathrm{d}x - \frac{1}{2}\,(1-t)^3 + \frac{1}{12}\cdot 6\,(\frac{1}{2}-t)_+.$$

Thus, if $0 \le t \le \frac{1}{2}$, then

$$6K_3(t) = \frac{1}{4}\,(1-t)^4 - \frac{1}{2}\,(1-t)^3 + \frac{1}{2}\,(\frac{1}{2}-t) = -\frac{1}{4}\,t^3(2-t),$$

and if $\frac{1}{2} \le t \le 1$,

$$6K_3(t) = -\frac{1}{4}\,(1-t)^3(2-(1-t)).$$

We see that K_3 is negative definite on $[0,1]$ and symmetric with respect to the midpoint. In particular,

$$E(f) = e_4 f^{(4)}(\tau), \quad 0 < \tau < 1,$$

with

$$e_4 = E(\frac{x^4}{4!}) = -\frac{1}{4!}\frac{1}{20} = -\frac{1}{480} \quad \text{(cf. (a))}.$$

58. (a) Requiring exactness for $f(x) \equiv 1$, $f(x) \equiv x$, $f(x) \equiv x^2$ gives, respectively,

$$\begin{aligned} \beta &= 1\,, \\ -\alpha + \tfrac{1}{2}\beta + \alpha &= \tfrac{1}{2}\,, \\ \tfrac{1}{4}\beta + 2\alpha &= \tfrac{1}{3}\,. \end{aligned}$$

Here, the second relation, in view of the first, is trivially satisfied, while the others yield

$$\beta = 1, \quad \alpha = \frac{1}{24}\,,$$

giving

$$\int_0^1 f(x)\mathrm{d}x = -\frac{1}{24}\,f'(0) + f\left(\frac{1}{2}\right) + \frac{1}{24}\,f'(1) + E(f).$$

(b) One checks that

$$E(x^3) = \frac{1}{4} - \frac{1}{8} - \frac{3}{24} = 0$$

and

$$E(x^4) = \frac{1}{5} - \frac{1}{16} - \frac{4}{24} = -\frac{7}{240}\,.$$

Therefore, the precise degree of exactness is $d = 3$.

(c) One has for the Peano kernel $K_d(t),\ 0 \le t \le 1$, with $d = 3$,

$$6K_3(t) = E_{(x)}((x-t)_+^3) = \int_0^1 (x-t)_+^3 \mathrm{d}x + \frac{1}{24}\,3\cdot(0-t)_+^2 - \left(\frac{1}{2}-t\right)_+^3$$

$$-\frac{1}{24}\cdot 3(1-t)_+^2 = \int_t^1 (x-t)^3 \mathrm{d}x - \left\{ \begin{array}{cc} \left(\frac{1}{2}-t\right)^3 & \text{if } \ 0 \le t \le \frac{1}{2} \\ 0 & \text{otherwise} \end{array} \right\} - \frac{1}{8}(1-t)^2.$$

Thus, when $0 \le t \le \frac{1}{2}$, then

$$6K_3(t) = \frac{1}{4}\,(1-t)^4 - \left(\frac{1}{2}-t\right)^3 - \frac{1}{8}\,(1-t)^2 = -\frac{t^2}{8}(1-2t^2) \le 0,$$

and when $\frac{1}{2} \le t \le 1$, then

$$6K_3(t) = \frac{1}{4}\,(1-t)^4 - \frac{1}{8}\,(1-t)^2 = -\frac{1}{8}\,(1-t)^2(1-2(1-t)^2) \le 0.$$

The Peano kernel is therefore negative definite (and symmetric with respect to the midpoint), hence

$$E(f) = e_4 f^{(4)}(\xi), \quad 0 < \xi < 1,$$

where, from (b),

$$e_4 = E\left(\frac{x^4}{4!}\right) = -\frac{1}{24}\,\cdot\,\frac{7}{240} = -\frac{7}{5760}\,.$$

(d) Making the change of variables $t = c + hx$, and noting that

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k}\,g(c+hx) = h^k g^{(k)}(c+hx), \quad k = 0, 1, 2, \ldots,$$

one gets from (a) and (c)

$$\int_c^{c+h} g(t)\mathrm{d}t = h\int_0^1 g(c+hx)\mathrm{d}x = -\frac{h^2}{24}\,g'(c)$$
$$+hg\left(c+\frac{1}{2}\,h\right) + \frac{h^2}{24}\,g'(c+h) + e_4 h^5 g^{(4)}(c+\xi h), \quad 0 < \xi < 1.$$

With $h = \frac{b-a}{n}$, $x_k = a + kh$, therefore

$$\int_{x_k}^{x_{k+1}} g(t)\mathrm{d}t = -\frac{h^2}{24}\, g'(x_k) + hg\left(x_k + \frac{1}{2}\, h\right) + \frac{h^2}{24}\, g'(x_{k+1}) + E_k,$$

$$E_k = e_4 h^5 g^{(4)}(\tau_k), \quad x_k < \tau_k < x_{k+1}.$$

Summing over k from 0 to $n-1$, and noting that

$$\sum_{k=0}^{n-1} E_k = e_4 h^4 \frac{b-a}{n} \sum_{k=0}^{n-1} g^{(4)}(\tau_k) = (b-a) e_4 h^4 g^{(4)}(\tau), \quad a < \tau < b,$$

we obtain by the "telescoping effect" on the derivative terms,

$$\int_a^b g(t)\mathrm{d}t = -\frac{h^2}{24}\, g'(a) + h\sum_{k=0}^{n-1} g\left(x_k + \frac{1}{2}\, h\right) + \frac{h^2}{24}\, g'(b) + E(g),$$

where, using the value of e_4 from (c), we have

$$E(g) = -\frac{7}{5760}\,(b-a) h^4 g^{(4)}(\tau), \quad a < \tau < b.$$

This is clearly an enhanced composite midpoint rule, with the correction terms approximating

$$\frac{h^2}{24}\,(b-a) g''(\xi) \approx \frac{h^2}{24}\,(b-a) \cdot \frac{g'(b) - g'(a)}{b-a} = \frac{h^2}{24}\,[g'(b) - g'(a)].$$

59. (a) Putting $f(x) \equiv 1$ and $f(x) \equiv x$ gives

$$\begin{aligned} A \quad + \quad B &= \frac{1}{\alpha+1}\,, \\ \frac{1}{2}\, B &= \frac{1}{\alpha+2}\,, \end{aligned}$$

hence

$$A = -\frac{\alpha}{(\alpha+1)(\alpha+2)}\,, \qquad B = \frac{2}{\alpha+2}\,.$$

(b) We have

$$E(f) = \int_0^1 x^\alpha f(x)\mathrm{d}x + \frac{\alpha}{(\alpha+1)(\alpha+2)}\, f(0) - \frac{2}{\alpha+2} \int_0^1 f(x)\mathrm{d}x,$$

so that, for $0 \le t \le 1$,

$$K_1(t) = E_{(x)}((x-t)_+) = \int_0^1 x^\alpha (x-t)_+ \mathrm{d}x - \frac{2}{\alpha+2} \int_0^1 (x-t)_+ \mathrm{d}x$$

$$= \int_t^1 x^\alpha (x-t)\mathrm{d}x - \frac{2}{\alpha+2}\int_t^1 (x-t)\mathrm{d}x$$
$$= \frac{1}{\alpha+2}\,x^{\alpha+2}\Big|_t^1 - \frac{t}{\alpha+1}\,x^{\alpha+1}\Big|_t^1 - \frac{1}{\alpha+2}(x-t)^2\Big|_t^1$$
$$= \frac{1}{\alpha+2}\,(1-t^{\alpha+2}) - \frac{t}{\alpha+1}\,(1-t^{\alpha+1}) - \frac{1}{\alpha+2}\,(1-t)^2$$
$$= \frac{t}{(\alpha+1)(\alpha+2)}\,q(t), \quad q(t) = t^{\alpha+1} - (\alpha+1)t + \alpha.$$

Here,

$$q(0) = \alpha, \quad q(1) = 0, \quad q'(t) = (\alpha+1)(t^\alpha - 1).$$

If $\alpha > 0$, then $q'(t) < 0$ on (0,1), and q decreases from α to 0, hence is positive on (0,1). Similarly, if $\alpha < 0$, then q increases from α to 0, hence is negative on (0,1). Therefore,

$$K_1(t)\begin{cases} > 0 \ \text{ on } \ (0,1) \ \text{ if } \ \alpha > 0, \\ < 0 \ \text{ on } \ (0,1) \ \text{ if } \ \alpha < 0, \end{cases}$$

as claimed.

(c) Since $e_2 = E\left(\frac{1}{2}\,x^2\right)$, we have

$$e_2 = \frac{1}{2}\int_0^1 x^{\alpha+2}\mathrm{d}x - \frac{2}{\alpha+2}\,\frac{1}{2}\int_0^1 x^2\mathrm{d}x$$
$$= \frac{1}{2}\,\frac{1}{\alpha+3} - \frac{1}{3(\alpha+2)} = \frac{\alpha}{6(\alpha+2)(\alpha+3)}\,.$$

60. See the text.

61. (a) Putting in turn $f(x) = 1$, $f(x) = x$ gives

$$a_1 + \frac{1}{2}\,a_2 = \frac{1}{\alpha+1}\,,$$
$$\frac{1}{2}\,a_1 + \frac{1}{3}\,a_2 = \frac{1}{\alpha+2}\,.$$

Solving for a_1, a_2, we get

$$a_1 = 2\,\frac{1-\alpha}{(\alpha+1)(\alpha+2)}\,, \quad a_2 = \frac{6\alpha}{(\alpha+1)(\alpha+2)}\,.$$

(b) Putting in turn $f(x) = x^{1/2}$, $f(x) = x^{3/2}$ gives

$$\frac{2}{3}\,a_1 + \frac{2}{5}\,a_2 = \frac{2}{2\alpha+3}\,,$$

$$\frac{2}{5}\,a_1 + \frac{2}{7}\,a_2 = \frac{2}{2\alpha+5}\,.$$

Solving for a_1, a_2, we get

$$a_1 = 15\,\frac{1-\alpha}{(2\alpha+3)(2\alpha+5)}\,, \quad a_2 = \frac{35\alpha}{(2\alpha+3)(2\alpha+5)}\,.$$

62. (a) From the table of divided differences for the data, initialized as

x	f	
x_k	f_k	
x_k	f_k	f'_k
x_{k+1}	f_{k+1}	
x_{k+1}	f_{k+1}	f'_{k+1}

one obtains

$$p(x) := p_3(f; x_k, x_k, x_{k+1}, x_{k+1}; x) = a_0 + a_1(x-x_k) + a_2(x-x_k)^2 \\ +a_3(x-x_k)^2(x-x_{k+1}),$$

where

$$a_0 = f_k, \quad a_1 = f'_k,$$
$$a_2 = \frac{1}{h}\left(\frac{f_{k+1}-f_k}{h} - f'_k\right),$$
$$a_3 = \frac{1}{h^2}\left(f'_{k+1} - 2\,\frac{f_{k+1}-f_k}{h} + f'_k\right).$$

Integrating, making the change of variable $x = x_k + th$, one gets

$$\int_{x_k}^{x_{k+1}} p(x)\mathrm{d}x = h\left\{a_0 + ha_1\int_0^1 t\mathrm{d}t + h^2a_2\int_0^1 t^2\mathrm{d}t + h^3a_3\int_0^1 t^2(t-1)\mathrm{d}t\right\}$$
$$= h\left\{a_0 + \frac{1}{2}\,ha_1 + \frac{1}{3}\,h^2a_2 - \frac{1}{12}\,h^3a_3\right\}$$
$$= h\left\{f_k + \frac{1}{2}\,hf'_k + \frac{1}{3}\,(f_{k+1}-f_k) - \frac{1}{3}\,hf'_k - \frac{1}{12}\,[hf'_{k+1} - 2(f_{k+1}-f_k) + hf'_k]\right\}$$
$$= h\left\{\frac{1}{2}\,f_k + \frac{1}{2}\,f_{k+1} - \frac{h}{12}\,(f'_{k+1} - f'_k)\right\}.$$

Integrating the remainder term of interpolation gives

$$\begin{aligned} E_k &= \int_{x_k}^{x_{k+1}} (x-x_k)^2(x-x_{k+1})^2 \ \frac{f^{(4)}(\xi(x))}{4!}\,\mathrm{d}x \\ &= \frac{f^{(4)}(\xi_k)}{4!}\int_{x_k}^{x_{k+1}} (x-x_k)^2(x-x_{k+1})^2\mathrm{d}x = \frac{h^5}{4!}\,f^{(4)}(\xi_k)\int_0^1 t^2(t-1)^2\mathrm{d}t \\ &= \frac{h^5}{720}\,f^{(4)}(\xi_k), \quad x_k < \xi_k < x_{k+1}, \end{aligned}$$

assuming $f \in C^4[a,b]$. Thus,

$$\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x = \frac{h}{2}\,(f_k + f_{k+1}) - \frac{h^2}{12}\,(f'_{k+1} - f'_k) + E_k, \quad E_k = \frac{h^5}{720}\,f^{(4)}(\xi_k).$$

This is a corrected trapezoidal rule, with the correction term approximating

$$-\frac{1}{12}\,h^3 f''(\xi) \approx -\frac{1}{12}\,h^3\ \frac{f'_{k+1} - f'_k}{h} = -\frac{h^2}{12}\,(f'_{k+1} - f'_k).$$

(b) Using the same procedure as in Sect. 3.2.1 for the composite trapezoidal rule gives

$$\int_a^b f(x)\mathrm{d}x = h\left(\frac{1}{2}\,f_0 + f_1 + \cdots + f_{n-1} + \frac{1}{2}\,f_n\right) - \frac{h^2}{12}\,(f'(b) - f'(a)) + E_n(f),$$
$$E_n(f) = \frac{b-a}{720}\,h^4 f^{(4)}(\xi), \quad a < \xi < b.$$

63. See the text.

64. (a) Assuming f analytic near the origin, we have for h sufficiently small, say $h \le h_0$,

$$f(h) = \sum_{k=0}^{\infty} f^{(k)}(0)\ \frac{h^k}{k!}.$$

Thus,

$$\begin{aligned} d_1(h) = \frac{f(h) - f(0)}{h} &= \sum_{k=1}^{\infty}\ \frac{f^{(k)}(0)}{k!}\,h^{k-1} \\ &= f'(0) + \sum_{k=2}^{\infty}\ \frac{f^{(k)}(0)}{k!}\,h^{k-1} \\ &= f'(0) + \sum_{k=1}^{\infty}\ \frac{f^{(k+1)}(0)}{(k+1)!}\,h^k, \end{aligned}$$

which shows that the error $d_1(h) - f'(0)$ admits a (convergent) expansion in powers of h. Therefore, $p_k = k$, $k = 1, 2, \ldots$, in the expansion

(3.85), and the extrapolation algorithm (3.90) takes the form

$$A_{m,k} = A_{m,k-1} + \frac{A_{m,k-1} - A_{m-1,k-1}}{q^{-k} - 1}\ , \quad m \geq k \geq 1,$$
$$A_{m,0} = d_1(q^m h_0),$$

where $0 < q < 1$.

(b) We now have

$$\begin{aligned} d_2(h) &= \frac{1}{h^2}\left(\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}\, h^k - 2f(0) + \sum_{k=0}^{\infty}(-1)^k\, \frac{f^{(k)}(0)}{k!}\, h^k\right) \\ &= \frac{2}{h^2} \sum_{\substack{k=2 \\ k \text{ even}}}^{\infty} \frac{f^{(k)}(0)}{k!}\, h^k \\ &= f''(0) + 2\sum_{\ell=1}^{\infty} \frac{f^{(2\ell+2)}(0)}{(2\ell+2)!}\, h^{2\ell}. \end{aligned}$$

Therefore,

$$d_2(h) - f''(0) = \sum_{k=1}^{\infty} a_k h^{p_k}, \quad a_k = 2\ \frac{f^{(2k+2)}(0)}{(2k+2)!}\ , \quad p_k = 2k.$$

The extrapolation algorithm therefore becomes, with $0 < q < 1$,

$$A_{m,k} = A_{m,k-1} + \frac{A_{m,k-1} - A_{m-1,k-1}}{q^{-2k} - 1}\,, \quad m \geq k \geq 1,$$
$$A_{m,0} = d_2(q^m h_0).$$

ANSWERS TO MACHINE ASSIGNMENTS

1. We have

$$e_1(h) = \pi - \tfrac{1}{h}\,(f_1 - f_0), \quad e_2(h) = 2\pi^2 - \tfrac{1}{h}\,(f_1 - 2f_0 + f_{-1}),$$
$$e_3(h) = 6\pi^3 - \tfrac{1}{h}\,(f_2 - 3f_1 + 3f_0 - f_{-1}),$$
$$e_4(h) = 24\pi^4 - \tfrac{1}{h}\,(f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2}).$$

PROGRAM

```
%MAIII_1
%
f0='%11.5f %8.4f %8.3f %8.2f\n';
```

```
f1=['%2.0f %8.5f %8.4f %8.3f %8.2f %7.3f %7.3f' ...
  '%7.3f %7.3f\n'];
disp('            exact derivatives')
df1=pi; df2=2*pi^2; df3=6*pi^3; df4=24*pi^4;
fprintf(f0,df1,df2,df3,df4)
fprintf('\n')
disp([' k       approximate derivatives          r1' ...
'      r2     r3      r4'])
i=(-2:2)'; j=(1:4)'; e=zeros(4,11); r=zeros(4,1);
for k=1:11
  h=2^(-k-1);
  f=1./(1-pi*h*i);
  diff1=(f(4)-f(3))/h;
  diff2=(f(4)-2*f(3)+f(2))/(h^2);
  diff3=(f(5)-3*f(4)+3*f(3)-f(2))/(h^3);
  diff4=(f(5)-4*f(4)+6*f(3)-4*f(2)+f(1))/(h^4);
  e(1,k)=df1-diff1; e(2,k)=df2-diff2;
  e(3,k)=df3-diff3; e(4,k)=df4-diff4;
  if k>1
    r(j)=abs(e(:,k)./e(:,k-1));
    fprintf(f1,k-1,diff1,diff2,diff3,diff4,r(1), ...
      r(2),r(3),r(4))
  end
end
```

OUTPUT

```
>> MAIII_1
            exact derivatives
   3.14159  19.7392  186.038  2337.82

 k       approximate derivatives          r1      r2      r3      r4
 1  5.17304  23.3383 1024.958  7214.08   0.177   0.113   0.809   0.751
 2  3.90915  20.5307  318.619  2874.91   0.378   0.220   0.158   0.110
 3  3.48359  19.9313  233.744  2455.23   0.446   0.243   0.360   0.219
 4  3.30377  19.7869  206.788  2366.27   0.474   0.248   0.435   0.242
 5  3.22064  19.7511  195.759  2344.88   0.487   0.250   0.468   0.248
 6  3.18062  19.7422  190.747  2339.58   0.494   0.250   0.484   0.250
 7  3.16099  19.7400  188.356  2338.26   0.497   0.250   0.492   0.250
 8  3.15126  19.7394  187.188  2337.93   0.498   0.250   0.496   0.253
 9  3.14642  19.7393  186.611  2337.84   0.499   0.250   0.498   0.194
10  3.14400  19.7392  186.324  2337.78   0.500   0.250   0.499   1.705
>>
```

It can be seen that the convergence order is $O(h)$ for $n = 1$ and $n = 3$, the

ratios r_k tending to $\frac{1}{2}$, but $O(h^2)$ for $n = 2$ and $n = 4$, the limit of the r_k being $\frac{1}{4}$. For $n = 4$ the limit is corrupted by rounding errors for the last few values of k.

2. (a) We have

$$\begin{aligned}\sin(x_0 + re^{i\theta}) &= \frac{1}{2i}\left(e^{i(x_0+r\exp(i\theta))} - e^{-i(x_0+r\exp(i\theta))}\right)\\ &= \frac{1}{2i}\left(e^{-r\sin\theta+i(x_0+r\cos\theta)} - e^{r\sin\theta-i(x_0+r\cos\theta)}\right)\\ &= \frac{1}{2i}\left[\left(e^{-r\sin\theta} - e^{r\sin\theta}\right)\cos(x_0 + r\cos\theta)\right.\\ &\qquad \left.+i\left(e^{-r\sin\theta} + e^{r\sin\theta}\right)\sin(x_0 + r\cos\theta)\right]\\ &= \cosh(r\sin\theta)\sin(x_0 + r\cos\theta) + i\sinh(r\sin\theta)\cos(x_0 + r\cos\theta)\end{aligned}$$

and similarly

$$\begin{aligned}\cos(x_0 + re^{i\theta}) &= \frac{1}{2}\left(e^{-r\sin\theta+i(x_0+r\cos\theta)} + e^{r\sin\theta-i(x_0+r\cos\theta)}\right)\\ &= \cosh(r\sin\theta)\cos(x_0 + r\cos\theta) - i\sinh(r\sin\theta)\sin(x_0 + r\cos\theta).\end{aligned}$$

Thus,

$$\begin{aligned}\tan(x_0 + re^{i\theta}) &= \frac{\frac{1}{2}\left[\sin(2x_0 + 2r\cos\theta) + i\sinh(2r\sin\theta)\right]}{\cos^2(x_0 + r\cos\theta) + \sinh^2(r\sin\theta)}\\ &= \frac{\sin(2x_0 + 2r\cos\theta) + i\sinh(2r\sin\theta)}{\cos(2x_0 + 2r\cos\theta) + \cosh(2r\sin\theta)},\end{aligned}$$

and there follows

$$y_m(\theta) = \mathrm{Re}\left\{e^{-im\theta}\tan(x_0+re^{i\theta})\right\} = \frac{\sin m\theta\sinh(2r\sin\theta) + \cos m\theta\sin(2x_0 + 2r\cos\theta)}{\cosh(2r\sin\theta) + \cos(2x_0 + 2r\cos\theta)}.$$

Note that

$$y_m(0) = y_m(2\pi) = \tan(x_0 + r).$$

(b) The analogue to (3.19) for the mth derivative is

$$f^{(m)}(x_0) = \frac{m!}{2\pi i}\oint_C \frac{f(z)dz}{(z - x_0)^{m+1}} = \frac{m!}{2\pi r^m}\int_0^{2\pi} e^{-im\theta} f(x_0 + re^{i\theta})d\theta.$$

(c) Applying the composite trapezoidal rule to the real part of the integral in (b), with $x_0 = 0$, gives

$$f^{(m)}(0) \approx t_n^{(m)}, \quad t_n^{(m)} = \frac{m!}{nr^m}\left(y_m(0) + \sum_{k=1}^{n-1} y_m(k\cdot 2\pi/n)\right).$$

The successive derivatives of $f(z) = \tan z$ are

$$f' = 1 + \tan^2 z, \quad f'' = 2\tan z(1 + \tan^2 z),$$
$$f''' = 2(1 + 4\tan^2 z + 3\tan^4 z), \quad f^{(4)} = 8\tan z(2 + 5\tan^2 z + 3\tan^4 z),$$
$$f^{(5)} = 8(2 + 17\tan^2 z + 30\tan^4 z + 15\tan^6 z),$$

so that the exact values of the derivatives at $z = 0$ are

$$f'(0) = 1,\ f''(0) = 0,\ f'''(0) = 2,\ f^{(4)}(0) = 0,\ f^{(5)}(0) = 16.$$

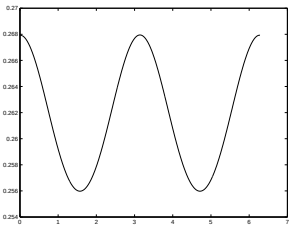
PROGRAM

```
%MAIII_2C
%
f0='%8.4f %5.1f %20.15f %12.4e\n';
f1='%14.1f %20.15f %12.4e\n';
m=1; mfac=1; dexact=1;
%m=2; mfac=2; dexact=0;
%m=3; mfac=6; dexact=2;
%m=4; mfac=24; dexcact=0;
%m=5; mfac=120; dexdact=16;
tp=(0:2*pi/500:2*pi)';
for rho=1:5
  r=rho*pi/12;
  r2=2*r;
  for n=5:5:50
    t=(2*pi/n:2*pi/n:2*pi*(1-1/n))';
    snum=sin(m*t).*sinh(r2*sin(t))+cos(m*t) ...
      .*sin(r2*cos(t));
    sden=cosh(r2*sin(t))+cos(r2*cos(t));
    s=snum./sden;
    if n==50
      figure(rho);
      yp=spline(t,s,tp);
      plot(tp,yp,'-');
    end
    d=mfac*(tan(r)+sum(s))/(n*(r^m));
    err=abs(dexact-d);
    if n==5
      fprintf(f0,r,n,d,err)
    else
      fprintf(f1,n,d,err)
    end
  end
  fprintf('\n')
```

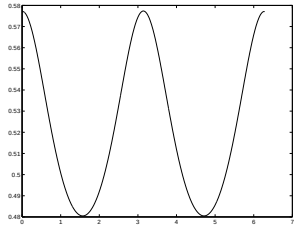
```
   end

    OUTPUT

>> MAIII_2C
     r     n             t_n             err
  0.2618   5.0    1.000000013405363   1.3405e-08   m=1
          10.0    1.000000013405362   1.3405e-08
          15.0    1.000000000000000   2.2204e-16
          20.0    1.000000000000000   2.2204e-16
          25.0    1.000000000000000   2.2204e-16
          30.0    1.000000000000000   2.2204e-16
          35.0    1.000000000000000   2.2204e-16
          40.0    1.000000000000000   2.2204e-16
          45.0    1.000000000000000   1.1102e-16
          50.0    1.000000000000000   2.2204e-16

  0.5236   5.0    1.000013727323495   1.3727e-05
          10.0    1.000013727323495   1.3727e-05
          15.0    1.000000000000004   3.9968e-15
          20.0    1.000000000232469   2.3247e-10
          25.0    1.000000000000000   2.2204e-16
          30.0    1.000000000000004   3.9968e-15
          35.0    1.000000000000000   2.2204e-16
          40.0    1.000000000000000   4.4409e-16
          45.0    1.000000000000000   1.1102e-16
          50.0    1.000000000000000   2.2204e-16

  0.7854   5.0    1.000792347014914   7.9235e-04
          10.0    1.000792347014914   7.9235e-04
          15.0    1.000000000754902   7.5490e-10
          20.0    1.000000773020022   7.7302e-07
          25.0    1.000000000000000   4.4409e-16
          30.0    1.000000000754902   7.5490e-10
          35.0    1.000000000000000   2.2204e-16
          40.0    1.000000000000737   7.3719e-13
          45.0    1.000000000000000   0.0000e+00
          50.0    1.000000000000001   6.6613e-16

  1.0472   5.0    1.014304604473103   1.4305e-02
          10.0    1.014304604473103   1.4305e-02
          15.0    1.000004227218872   4.2272e-06
          20.0    1.000243834798267   2.4383e-04
          25.0    1.000000001271239   1.2712e-09
          30.0    1.000004227218872   4.2272e-06
```

```
            35.0    1.000000000000382   3.8192e-13
            40.0    1.000000073306067   7.3306e-08
            45.0    1.000000000000000   0.0000e+00
            50.0    1.000000001271239   1.2712e-09

  1.3090     5.0    1.156127099861422   1.5613e-01
            10.0    1.156127099861423   1.5613e-01
            15.0    1.003429148446004   3.4291e-03
            20.0    1.021709201202098   2.1709e-02
            25.0    1.000089079067936   8.9079e-05
            30.0    1.003429148446004   3.4291e-03
            35.0    1.000002323294493   2.3233e-06
            40.0    1.000551868981410   5.5187e-04
            45.0    1.000000060600768   6.0601e-08
            50.0    1.000089079067936   8.9079e-05
>>
```

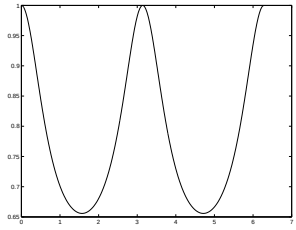
Interestingly, the results for $n = 5$ and $n = 10$ seem to be identical. Convergence is seen to be quite fast when r is relatively small and slows down as r becomes larger. This is so because with r increasing (toward $\frac{1}{2}\pi$) the contour of integration approaches a singularity of f. At the same time the integrand $y_m(\theta)$ develops a sharp peak at $\theta = \pi$, which the composite trapezoidal rule can handle well only if this point does not belong to the subdivision points of the rule, that is if n is odd. Otherwise, the peak value is not well conditioned and subject to relatively large error, more so the sharper the peak. This is clearly visible in the output, especially for the larger values of r, by comparing results for n even and n odd. The respective plots are shown below.



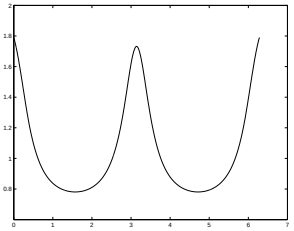
$r = \frac{1}{12}\pi$



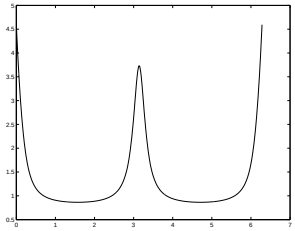
$r = \frac{1}{6}\pi$



$r = \frac{1}{4}\pi$



$r = \frac{1}{3}\pi$

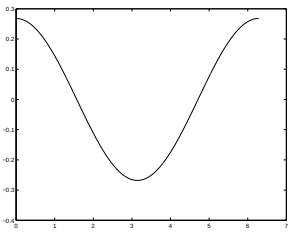


$r = \frac{5}{12}\pi$

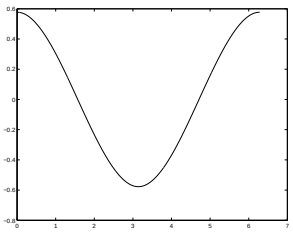
The results and plots for $m = 2, 3, 4, 5$ are displayed below in the same manner. Equality of the results for $n = 5$ and $n = 10$ is observed again when m is odd. When m is even, it is the even values of n that give better results, whereas for m odd, the odd values of n give more accurate results. A possible explanation is that for m even, the graphs of the integrand exhibit "reverse peaks" whose values are quite small and therefore less sensitive to inaccurate evaluation.

```
>> MAIII_2C
     r      n             t_n              err
  0.2618   5.0    0.000132742961844   1.3274e-04   m=2
          10.0   -0.000000000000000   4.8595e-16
          15.0    0.000000000002196   2.1959e-12
          20.0   -0.000000000000000   2.4298e-16
          25.0   -0.000000000000002   1.6846e-15
          30.0   -0.000000000000001   8.6392e-16
          35.0   -0.000000000000000   2.7769e-16
          40.0   -0.000000000000001   8.9091e-16
          45.0   -0.000000000000002   1.6198e-15
          50.0   -0.000000000000002   2.1382e-15

  0.5236   5.0    0.004247846635285   4.2478e-03
          10.0    0.000000000000000   0.0000e+00
          15.0    0.000000071925277   7.1925e-08
          20.0   -0.000000000000000   1.6198e-16
          25.0    0.000000000001217   1.2171e-12
          30.0   -0.000000000000001   5.6695e-16
          35.0   -0.000000000000000   2.0827e-16
          40.0   -0.000000000000000   1.4174e-16
          45.0   -0.000000000000001   5.3995e-16
          50.0   -0.000000000000001   8.9091e-16

  0.7854   5.0    0.032288065621579   3.2288e-02
          10.0   -0.000000000000000   3.5997e-17
          15.0    0.000031495639206   3.1496e-05
          20.0   -0.000000000000000   1.6198e-16
          25.0    0.000000030757459   3.0757e-08
          30.0   -0.000000000000000   4.1996e-16
          35.0    0.000000000030037   3.0037e-11
          40.0   -0.000000000000000   1.8898e-16
          45.0    0.000000000000029   2.8789e-14
          50.0   -0.000000000000001   8.1352e-16

  1.0472   5.0    0.138327230764456   1.3833e-01
          10.0   -0.000000000000000   1.6198e-16
```

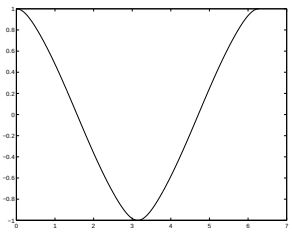
```
                15.0    0.002356859754063   2.3569e-03
                20.0   -0.000000000000000   2.0248e-16
                25.0    0.000040871340741   4.0871e-05
                30.0   -0.000000000000000   1.0799e-16
                35.0    0.000000708771577   7.0877e-07
                40.0   -0.000000000000000   1.2149e-16
                45.0    0.000000012291183   1.2291e-08
                50.0   -0.000000000000001   8.3422e-16

      1.3090    5.0    0.494709739051825   4.9471e-01
                10.0   -0.000000000000000   1.0367e-16
                15.0    0.067269019317941   6.7269e-02
                20.0   -0.000000000000000   1.5550e-16
                25.0    0.010819742723187   1.0820e-02
                30.0   -0.000000000000000   3.4557e-16
                35.0    0.001747261845200   1.7473e-03
                40.0   -0.000000000000000   1.5550e-16
                45.0    0.000282191755035   2.8219e-04
                50.0   -0.000000000000001   8.0863e-16
>>
```



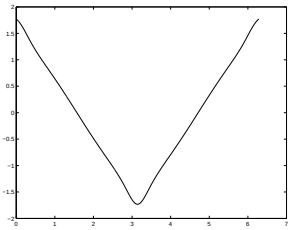
$r = \frac{1}{12}\pi$



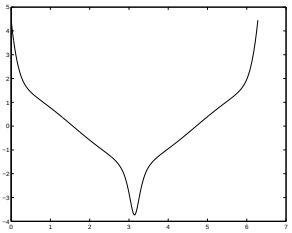
$r = \frac{1}{6}\pi$



$r = \frac{1}{4}\pi$



$r = \frac{1}{3}\pi$

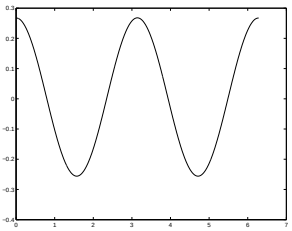


$r = \frac{5}{12}\pi$

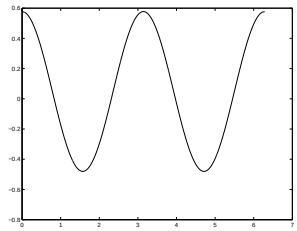
```
>> MAIII_2C
     r     n           t_n            err
  0.2618   5.0    2.000000032597886   3.2598e-08   m=3
           10.0   2.000000032597880   3.2598e-08
           15.0   1.999999999999985   1.5321e-14
```

```
         20.0    1.999999999999996   3.9968e-15
         25.0    2.000000000000001   1.3323e-15
         30.0    1.999999999999991   9.1038e-15
         35.0    1.999999999999997   2.8866e-15
         40.0    1.999999999999994   6.4393e-15
         45.0    2.000000000000008   8.4377e-15
         50.0    2.000000000000006   6.2172e-15

0.5236    5.0    2.000033380791990   3.3381e-05
         10.0    2.000033380791989   3.3381e-05
         15.0    2.000000000000005   4.8850e-15
         20.0    2.000000000565296   5.6530e-10
         25.0    2.000000000000001   1.3323e-15
         30.0    2.000000000000007   7.1054e-15
         35.0    2.000000000000000   4.4409e-16
         40.0    1.999999999999999   1.3323e-15
         45.0    2.000000000000002   1.7764e-15
         50.0    2.000000000000003   2.6645e-15

0.7854    5.0    2.001926753670502   1.9268e-03
         10.0    2.001926753670502   1.9268e-03
         15.0    2.000000001835699   1.8357e-09
         20.0    2.000001879759285   1.8798e-06
         25.0    2.000000000000002   2.2204e-15
         30.0    2.000000001835700   1.8357e-09
         35.0    2.000000000000000   4.4409e-16
         40.0    2.000000000001793   1.7928e-12
         45.0    2.000000000000000   4.4409e-16
         50.0    2.000000000000003   2.6645e-15

1.0472    5.0    2.034784569653270   3.4785e-02
         10.0    2.034784569653270   3.4785e-02
         15.0    2.000010279363670   1.0279e-05
         20.0    2.000592935128948   5.9294e-04
         25.0    2.000000003091284   3.0913e-09
         30.0    2.000010279363670   1.0279e-05
         35.0    2.000000000000929   9.2948e-13
         40.0    2.000000178258979   1.7826e-07
         45.0    2.000000000000000   4.4409e-16
         50.0    2.000000003091285   3.0913e-09

1.3090    5.0    2.379655047628025   3.7966e-01
         10.0    2.379655047628026   3.7966e-01
         15.0    2.008338689106426   8.3387e-03
         20.0    2.052790447079972   5.2790e-02
```

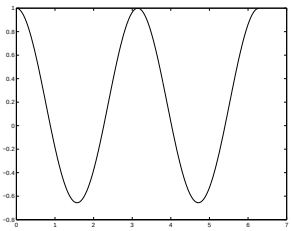
```
          25.0    2.000216614318425   2.1661e-04
          30.0    2.008338689106426   8.3387e-03
          35.0    2.000005649574752   5.6496e-06
          40.0    2.001341984441886   1.3420e-03
          45.0    2.000000147363397   1.4736e-07
          50.0    2.000216614318424   2.1661e-04
>>
```



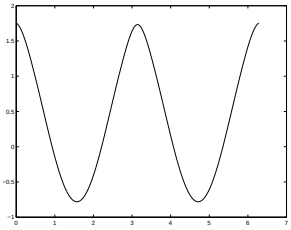
$r = \frac{1}{12}\pi$



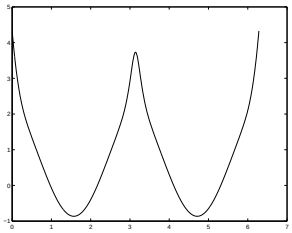
$r = \frac{1}{6}\pi$



$r = \frac{1}{4}\pi$



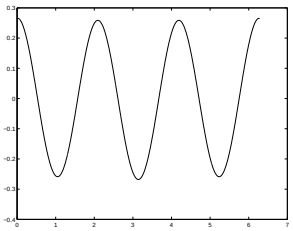
$r = \frac{1}{3}\pi$



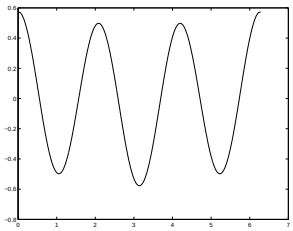
$r = \frac{5}{12}\pi$

```
>> MAIII_2C
     r      n            t_n             err
  0.2618   5.0    0.000645495187062   6.4550e-04   m=4
          10.0    0.000000000000000   0.0000e+00
          15.0    0.000000000010891   1.0891e-11
          20.0   -0.000000000000057   5.6721e-14
          25.0   -0.000000000000272   2.7226e-13
          30.0    0.000000000000076   7.5629e-14
          35.0   -0.000000000000049   4.8618e-14
          40.0   -0.000000000000043   4.2541e-14
          45.0   -0.000000000000202   2.0168e-13
          50.0   -0.000000000000261   2.6092e-13

  0.5236   5.0    0.020656195451272   2.0656e-02
          10.0    0.000000000000000   0.0000e+00
          15.0    0.000000349802623   3.4980e-07
          20.0   -0.000000000000004   3.5451e-15
          25.0    0.000000000005888   5.8877e-12
```

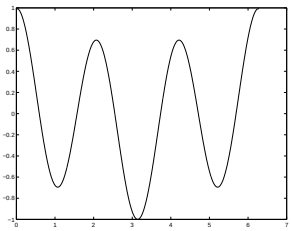
```
          30.0    0.000000000000007   7.0902e-15
          35.0   -0.000000000000006   6.0773e-15
          40.0   -0.000000000000005   5.3176e-15
          45.0   -0.000000000000029   2.9148e-14
          50.0   -0.000000000000038   3.8287e-14

 0.7854    5.0    0.157008654002984   1.5701e-01
          10.0   -0.000000000000001   7.0026e-16
          15.0    0.000153176420962   1.5318e-04
          20.0   -0.000000000000004   3.8515e-15
          25.0    0.000000149586336   1.4959e-07
          30.0   -0.000000000000000   2.3342e-16
          35.0    0.000000000146078   1.4608e-10
          40.0   -0.000000000000003   2.9761e-15
          45.0    0.000000000000134   1.3414e-13
          50.0   -0.000000000000012   1.1624e-14

 1.0472    5.0    0.672651674371379   6.7265e-01
          10.0    0.000000000000000   0.0000e+00
          15.0    0.011462391331814   1.1462e-02
          20.0   -0.000000000000001   1.3294e-15
          25.0    0.000198774365804   1.9877e-04
          30.0   -0.000000000000001   7.3856e-16
          35.0    0.000003447051603   3.4471e-06
          40.0   -0.000000000000001   9.9706e-16
          45.0    0.000000059777142   5.9777e-08
          50.0   -0.000000000000007   7.0902e-15

 1.3090    5.0    2.405701019883517   2.4057e+00
          10.0   -0.000000000000000   3.6302e-16
          15.0    0.327157278920604   3.2716e-01
          20.0   -0.000000000000001   7.2603e-16
          25.0    0.052620918692102   5.2621e-02
          30.0   -0.000000000000002   1.6941e-15
          35.0    0.008497662637858   8.4977e-03
          40.0   -0.000000000000001   1.4521e-15
          45.0    0.001372416126444   1.3724e-03
          50.0   -0.000000000000005   5.2274e-15
>>
```



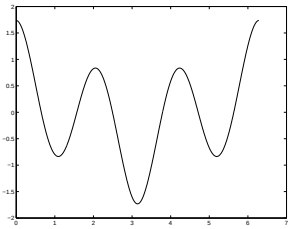
$r = \frac{1}{12}\pi$



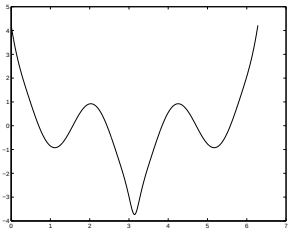
$r = \frac{1}{6}\pi$



$r = \frac{1}{4}\pi$



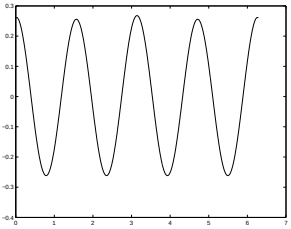
$r = \frac{1}{3}\pi$



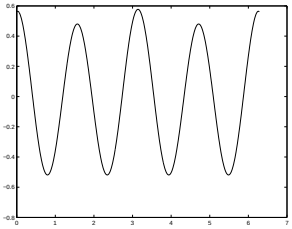
$r = \frac{5}{12}\pi$

```
>> MAIII_2C
     r      n            t_n             err
  0.2618   5.0   16.000000264231517   2.6423e-07   m=5
          10.0   16.000000264227726   2.6423e-07
          15.0   16.000000000006224   6.2244e-12
          20.0   15.999999999998913   1.0871e-12
          25.0   15.999999999992575   7.4252e-12
          30.0   15.999999999993042   6.9580e-12
          35.0   15.999999999998641   1.3589e-12
          40.0   15.999999999998641   1.3589e-12
          45.0   15.999999999998279   1.7213e-12
          50.0   16.000000000000373   3.7303e-13

  0.5236   5.0   16.000270574458330   2.7057e-04
          10.0   16.000270574458092   2.7057e-04
          15.0   16.000000000000458   4.5830e-13
          20.0   16.000000004582098   4.5821e-09
          25.0   15.999999999999575   4.2455e-13
          30.0   15.999999999999622   3.7836e-13
          35.0   15.999999999999950   4.9738e-14
          40.0   15.999999999999936   6.3949e-14
          45.0   15.999999999999874   1.2612e-13
          50.0   16.000000000000096   9.5923e-14

  0.7854   5.0   16.015617674096532   1.5618e-02
          10.0   16.015617674096486   1.5618e-02
```

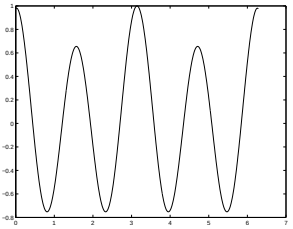
```
          15.0   16.000000014879696   1.4880e-08
          20.0   16.000015236754834   1.5237e-05
          25.0   15.999999999999929   7.1054e-14
          30.0   16.000000014879536   1.4880e-08
          35.0   15.999999999999989   1.0658e-14
          40.0   16.000000000014506   1.4506e-11
          45.0   15.999999999999970   3.0198e-14
          50.0   16.000000000000021   2.1316e-14

 1.0472    5.0   16.281953050045942   2.8195e-01
          10.0   16.281953050045928   2.8195e-01
          15.0   16.000083321383560   8.3321e-05
          20.0   16.004806151127035   4.8062e-03
          25.0   16.000000025056970   2.5057e-08
          30.0   16.000083321383514   8.3321e-05
          35.0   16.000000000007532   7.5318e-12
          40.0   16.000001444912858   1.4449e-06
          45.0   15.999999999999995   5.3291e-15
          50.0   16.000000025056998   2.5057e-08

 1.3090    5.0   19.077367424160691   3.0774e+00
          10.0   19.077367424160688   3.0774e+00
          15.0   16.067590868023096   6.7591e-02
          20.0   16.427903246650772   4.2790e-01
          25.0   16.001755809530920   1.7558e-03
          30.0   16.067590868023071   6.7591e-02
          35.0   16.000045793728077   4.5794e-05
          40.0   16.010877716166522   1.0878e-02
          45.0   16.000001194482703   1.1945e-06
          50.0   16.001755809530927   1.7558e-03
>>
```



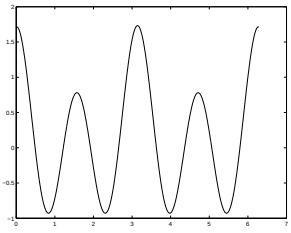
$r = \frac{1}{12}\pi$



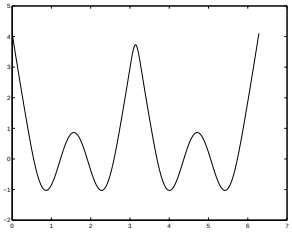
$r = \frac{1}{6}\pi$



$r = \frac{1}{4}\pi$



$r = \frac{1}{3}\pi$



$r = \frac{5}{12}\pi$

(d) The exact values of the derivatives of f at $z = x_0$ are now, with $t = \tan x_0$,

$$f'(x_0) = 1 + t^2,$$
$$f''(x_0) = 2t(1 + t^2).$$
$$f'''(x_0) = 2(1 + 4t^2 + 3t^4),$$
$$f^{(4)}(x_0) = 8(2t + 5t^3 + 3t^5),$$
$$f^{(5)}(x_0) = 8(2 + 17t^2 + 30t^4 + 15t^6),$$

giving, for $d_m = f^{(m)}(\frac{7}{16}\pi)$,

$$d_1 = 2.627414236908816 \times 10^1,$$
$$d_2 = 2.641780671077078 \times 10^2,$$
$$d_3 = 4.036886773910330 \times 10^3,$$
$$d_4 = 8.223591348331046 \times 10^4,$$
$$d_5 = 2.094121920500841 \times 10^6.$$

PROGRAM

```
%MAIII_2D
%
f0='%8.4f %5.1f %20.15f %12.4e\n';
f1='%14.1f %20.15f %12.4e\n';
disp('     r     n              t_n              err')
m=1; mfac=1; dexact=2.627414236908816e+01;
%m=2; mfac=2; dexact=2.641780671077078e+02;
%m=3; mfac=6; dexact=4.036886773910330e+03;
%m=4; mfac=24; dexact=8.223591348331046e+04;
%m=5; mfac=120; dexact=2.094121920500841e+06;
x0=7*pi/16;
tp=(0:2*pi/500:2*pi)';
for r=pi/32;
  r2=2*r;
```

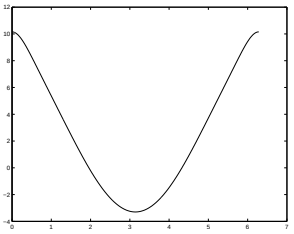
```
    for n=5:5:50
      t=(2*pi/n:2*pi/n:2*pi*(1-1/n))';
      snum=sin(m*t).*sinh(r2*sin(t))+cos(m*t) ...
        .*sin(2*x0+r2*cos(t));
      sden=cosh(r2*sin(t))+cos(2*x0+r2*cos(t));
      s=snum./sden;
      if n==50
        figure(1);
        yp=spline(t,s,tp);
        plot(tp,yp,'-');
      end
      d=mfac*(tan(x0+r)+sum(s))/(n*(r^m));
      err=abs((dexact-d)/dexact);
      if n==5
        fprintf(f0,r,n,d,err)
      else
        fprintf(f1,n,d,err)
      end
    end
    fprintf('\n')
  end
```

OUTPUT

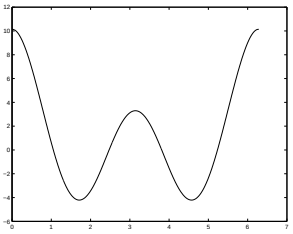
```
>> MAIII_2D
     r     n           t_n              err
  0.0982   5.0   27.11085923767755   3.1846e-02   m=1
          10.0   26.29949742579631   9.6502e-04
          15.0   26.27493396499295   3.0128e-05
          20.0   26.27416710572881   9.4148e-07
          25.0   26.27414314210744   2.9421e-08
          30.0   26.27414239324500   9.1941e-10
          35.0   26.27414236984307   2.8732e-11
          40.0   26.27414236911171   8.9622e-13
          45.0   26.27414236908884   2.5691e-14
          50.0   26.27414236908815   5.4087e-16
>>
>> MAIII_2D
     r     n           t_n              err
  0.0982   5.0  272.7007951791209   3.2261e-02   m=2
          10.0  264.4363315945859   9.7762e-04
          15.0  264.1861302373828   3.0522e-05
          20.0  264.1783190730600   9.5377e-07
          25.0  264.1780749816171   2.9805e-08
          30.0  264.1780673537671   9.3141e-10
```

```
          35.0  264.1780671153968   2.9105e-11
          40.0  264.1780671079470   9.0544e-13
          45.0  264.1780671077143   2.4529e-14
          50.0  264.1780671077073   1.9365e-15
>>
>> MAIII_2D
     r     n          t_n              err
  0.0982   5.0  4167.104466755103   3.2257e-02   m=3
          10.0  4040.832764602646   9.7748e-04
          15.0  4037.009969457004   3.0517e-05
          20.0  4036.890623657326   9.5364e-07
          25.0  4036.886894214783   2.9801e-08
          30.0  4036.886777669848   9.3129e-10
          35.0  4036.886774027802   2.9100e-11
          40.0  4036.886773913979   9.0400e-13
          45.0  4036.886773910416   2.1403e-14
          50.0  4036.886773910308   5.5198e-15
>>
>> MAIII_2D
     r     n          t_n              err
  0.0982   5.0  84888.68653862808   3.2258e-02   m=4
          10.0  82316.30054559265   9.7752e-04
          15.0  82238.42320237930   3.0519e-05
          20.0  82235.99190971252   9.5368e-07
          25.0  82235.91593413193   2.9802e-08
          30.0  82235.91355989916   9.3133e-10
          35.0  82235.91348570332   2.9097e-11
          40.0  82235.91348338491   9.0529e-13
          45.0  82235.91348331215   2.0527e-14
          50.0  82235.91348330962   1.0263e-14
>>
>> MAIII_2D
     r     n          t_n              err
  0.0982   5.0  68310172.32495427   3.1620e+01   m=5
          10.0  2096168.960232492   9.7752e-04
          15.0  2094185.829972144   3.0519e-05
          20.0  2094123.917612735   9.5368e-07
          25.0  2094121.982910499   2.9802e-08
          30.0  2094121.922451136   9.3132e-10
          35.0  2094121.920561763   2.9092e-11
          40.0  2094121.920502725   8.9958e-13
          45.0  2094121.920500869   1.3342e-14
          50.0  2094121.920500811   1.4454e-14
>>
```

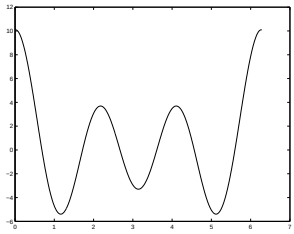
The integrand $y_m(\theta)$, as shown in the graphs below, is now uniformly smooth, with no unusual behavior in the neighborhood of $\theta = \pi$. Accordingly, convergence as $n \to \infty$ is essentially monotone.



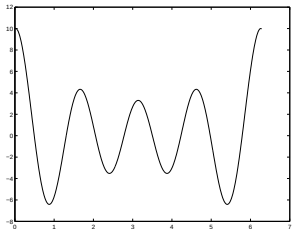
$m = 1$



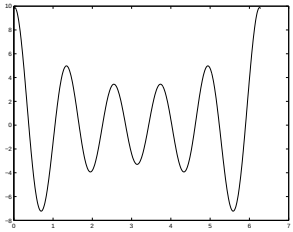
$m = 2$



$m = 3$



$m = 4$



$m = 5$

(e) In terms of the difference and differential operators

$$
\begin{aligned}
&(E^\alpha f)(x) = f(x + \alpha h),\\
&\delta f = (E^{\frac{1}{2}} - E^{-\frac{1}{2}})f, \quad \mu f = \tfrac{1}{2}(E^{\frac{1}{2}} + E^{-\frac{1}{2}})f,\\
&(D^m f)(x) = f^{(m)}(x),
\end{aligned}
$$

one has (see, e.g., F.B. Hildebrand, *Introduction to numerical analysis*, 2nd ed., McGraw-Hill, New York, 1974, Eqs. (5.3.11)–(5.3.16))

$$
\begin{aligned}
Df &= h^{-1}\mu(\delta - \tfrac{1}{6}\delta^3 + \cdots\ )f,\\
D^2 f &= h^{-2}(\delta^2 - \tfrac{1}{12}\delta^4 + \cdots\ )f,\\
D^3 f &= h^{-3}\mu(\delta^3 - \tfrac{1}{4}\delta^5 + \cdots\ )f,\\
D^4 f &= h^{-4}(\delta^4 - \tfrac{1}{6}\delta^6 + \cdots\ )f,\\
D^5 f &= h^{-5}\mu(\delta^5 - \tfrac{1}{3}\delta^7 + \cdots\ )f.
\end{aligned}
$$

Retaining only the first term in each expansion yields the central finite

difference approximations

$$
\begin{aligned}
Df &\approx \tfrac{1}{2}h^{-1}(E - E^{-1})f,\\
D^2f &\approx h^{-2}(E - 2I + E^{-1})f,\\
D^3f &\approx \tfrac{1}{2}h^{-3}(E^2 - 2E + 2E^{-1} - E^{-2})f,\\
D^4f &\approx h^{-4}(E^2 - 4E + 6I - 4E^{-1} + E^{-2})f,\\
D^5f &\approx \tfrac{1}{2}h^{-5}(E^3 - 4E^2 + 5E - 5E^{-1} + 4E^{-2} - E^{-3})f,
\end{aligned}
$$

which are exact, respectively, for polynomials of degree 2, 2, 4, 4, 6, 6. (They could also have been derived by the method of undetermined coefficients.) For functions f that are odd (like $f = \tan$), the derivatives of even order vanish at the origin, and the others simplify to

$$
\begin{aligned}
(Df)(0) &\approx h^{-1}f(h),\\
(D^3f)(0) &\approx h^{-3}[f(2h) - 2f(h)],\\
(D^5f)(0) &\approx h^{-5}[f(3h) - 4f(2h) + 5f(h)].
\end{aligned}
$$

The matlab script below implements the last three formulae for $h = 1, \frac{1}{5}, \frac{1}{25}, \frac{1}{125}, \frac{1}{625}$.

```
 PROGRAM

%MAIII_2E
%
f0='%6.4f %18.13f %18.13f %18.13f\n';
h=5;
for ih=1:5
  h=h/5;
  d1=tan(h)/h;
  d3=(tan(2*h)-2*tan(h))/(h^3);
  d5=(tan(3*h)-4*tan(2*h)+5*tan(h))/(h^5);
  fprintf(f0,h,d1,d3,d5)
end
```

The results, shown below, as expected are less accurate than those in (c), the errors increasing with the order of the derivative.

```
 OUTPUT
>> MAIII_2E
   h            d1                  d3                  d5
1.0000    1.5574077246549   -5.2998553125713   16.3846515332463
```

```
0.2000    1.0135501775434    2.1716434651021   20.3565966637745
0.0400    1.0005336748879    2.0064174538037   16.1460604285982
0.0080    1.0000213338795    2.0002560278558   16.0058043513110
0.0016    1.0000008533342    2.0000102400483   16.0003200247357
>>
```

(f) At $x_0 = \frac{7}{16}\pi$ the central difference approximations are even less accurate than at $x_0 = 0$.

```
 PROGRAM

%MAIII_2F
%
f0='%8.6f %15.8e %15.8e\n';
f1='%8.6f %15.8e %15.8e %15.8e\n';
disp('    h           d1              d2')
x0=7*pi/16; t0=tan(x0);
h=5*pi/32;
for ih=1:5
  h=h/5;
  t1=tan(x0+h); tm1=tan(x0-h);
  d1=(t1-tm1)/(2*h); d2=(t1-2*t0+tm1)/(h^2);
  fprintf(f0,h,d1,d2)
end
fprintf('\n')
disp('    h           d3              d4              d5')
h=5*pi/32;
for ih=1:5
  h=h/5;
  t1=tan(x0+h); t2=tan(x0+2*h); t3=tan(x0+3*h);
  tm1=tan(x0-h); tm2=tan(x0-2*h); tm3=tan(x0-3*h);
  d3=(t2-2*t1+2*tm1-tm2)/(2*h^3);
  d4=(t2-4*t1+6*t0-4*tm1+tm2)/(h^4);
  d5=(t3-4*t2+5*t1-5*tm1+4*tm2-tm3)/(2*h^5);
  fprintf(f1,h,d3,d4,d5)
end

 OUTPUT

>> MAIII_2F
    h           d1              d2
0.098175  3.49204391e+01  3.52246216e+02
0.019635  2.65361535e+01  2.66846798e+02
0.003927  2.62845222e+01  2.64283791e+02
0.000785  2.62745574e+01  2.64182294e+02
```

```
0.000157  2.62741590e+01  2.64178236e+02

    h           d3              d4              d5
0.098175  8.62957614e+18  1.75800282e+20 -3.58137393e+21
0.019635  4.24755886e+03  8.65276894e+04  2.42133085e+06
0.003927  4.04497386e+03  8.24006622e+04  2.10589844e+06
0.000785  4.03720974e+03  8.22424771e+04  2.09460840e+06
0.000157  4.03689983e+03  8.22285278e+04  2.06647829e+06
>>
```

3. See the text.

4. Let $h = \frac{1}{n},\ x_k = kh,\ k = 0, 1, \ldots, n.$

   (a) Ignoring the singularity, the integrals are approximated by

$$I_c \approx h\left(\sum_{k=1}^{n-1} \frac{\cos x_k}{\sqrt{x_k}} + \frac{1}{2}\cos 1\right), \quad I_s \approx h\left(\sum_{k=1}^{n-1} \frac{\sin x_k}{\sqrt{x_k}} + \frac{1}{2}\sin 1\right).$$

PROGRAM

```
%MAIII_4A
%
f0='%8.0f %12.8f %12.8f\n';
f1='     inf %12.8f %12.8f\n';
disp('      n        Ic           Is')
Icexact=1.80904848; Isexact=.62053660;
for n=100:100:1000
  h=1/n; k=(1:n-1)'; x=k*h;
  Ic=h*(sum(cos(x)./sqrt(x))+cos(1)/2);
  Is=h*(sum(sin(x)./sqrt(x))+sin(1)/2);
  fprintf(f0,n,Ic,Is)
end
fprintf('\n')
fprintf(f1,Icexact,Isexact)
```

OUTPUT

```
>> MAIII_4A
      n        Ic           Is
     100   1.66300389   0.62032971
     200   1.70578352   0.62046335
     300   1.72473385   0.62049671
     400   1.73603018   0.62051068
     500   1.74373907   0.62051805
     600   1.74942950   0.62052249
```

```
      700   1.75385208   0.62052540
      800   1.75741700   0.62052743
      900   1.76036988   0.62052892
     1000   1.76286792   0.62053004

      inf   1.80904848   0.62053660
>>
```

It can be seen that convergence is extremely slow for the cos-integral, and somewhat faster (but still slow) for the sin-integral. The reason is that for the former integral the integrand is singular at $x = 0$, whereas for the latter the first derivative is singular while the integrand itself is continuous at $x = 0$.

(b) One now approximates (cf. (3.43))

$$I_c \approx \frac{2}{3}h^{1/2}(2+\cos h) + h\left(\frac{1}{2}h^{-1/2}\cos h + \sum_{k=2}^{n-1}\frac{\cos x_k}{\sqrt{x_k}} + \frac{1}{2}\cos 1\right)$$
$$= \frac{h^{1/2}}{6}(8+7\cos h) + h\left(\sum_{k=2}^{n-1}\frac{\cos x_k}{\sqrt{x_k}} + \frac{1}{2}\cos 1\right),$$

and

$$I_s \approx \frac{2}{3}h^{1/2}\sin h + h\left(\frac{1}{2}h^{-1/2}\sin h + \sum_{k=2}^{n-1}\frac{\sin x_k}{\sqrt{x_k}} + \frac{1}{2}\sin 1\right)$$
$$= \frac{7}{6}h^{1/2}\sin h + h\left(\sum_{k=2}^{n-1}\frac{\sin x_k}{\sqrt{x_k}} + \frac{1}{2}\sin 1\right).$$

PROGRAM

```
%MAIII_4B
%
f0='%8.0f %12.8f %12.8f\n';
f1='     inf %12.8f %12.8f\n';
disp('      n        Ic           Is')
Icexact=1.80904848; Isexact=.62053660;
for n=100:100:1000
  h=1/n; k=(2:n-1)'; x=k*h;
  Ic=sqrt(h)*(8+7*cos(h))/6+h*(sum(cos(x) ...
    ./sqrt(x))+cos(1)/2);
  Is=7*sqrt(h)*sin(h)/6+h*(sum(sin(x) ...
    ./sqrt(x))+sin(1)/2);
  fprintf(f0,n,Ic,Is)
end
fprintf('\n')
```

```
fprintf(f1,Icexact,Isexact)

 OUTPUT

>> MAIII_4B
      n        Ic           Is
    100   1.81300306   0.62049638
    200   1.81184939   0.62052228
    300   1.81133633   0.62052878
    400   1.81103015   0.62053151
    500   1.81082109   0.62053296
    600   1.81066673   0.62053383
    700   1.81054674   0.62053440
    800   1.81045001   0.62053480
    900   1.81036987   0.62053509
   1000   1.81030208   0.62053531

    inf   1.80904848   0.62053660
>>
```

The accuracy is improved compared to the results in (a), but convergence is still very slow.

(c) The change of variables $x = t^2$ yields

$$I_c = 2\int_0^1 \cos t^2 \mathrm{d}t \approx 2h\left(\frac{1}{2} + \sum_{k=1}^{n-1} \cos x_k^2 + \frac{1}{2}\cos 1\right)$$

and

$$I_s = 2\int_0^1 \sin t^2 \mathrm{d}t \approx 2h\left(\sum_{k=1}^{n-1} \sin x_k^2 + \frac{1}{2}\sin 1\right).$$

```
 PROGRAM

%MAIII_4C
%
f0='%8.0f %12.8f %12.8f\n';
f1='     inf %12.8f %12.8f\n';
disp('      n        Ic           Is')
Icexact=1.80904848; Isexact=.62053660;
for n=20:20:200
  h=1/n; k=(1:n-1)'; x=k*h;
  Ic=2*h*(1/2+sum(cos(x.^2))+cos(1)/2);
  Is=2*h*(sum(sin(x.^2))+sin(1)/2);
  fprintf(f0,n,Ic,Is)
end
```

```
fprintf('\n')
fprintf(f1,Icexact,Isexact)
```

OUTPUT

```
>> MAIII_4C
      n        Ic           Is
     20   1.80834725   0.62098711
     40   1.80887317   0.62064918
     60   1.80897056   0.62058663
     80   1.80900465   0.62056475
    100   1.80902043   0.62055461
    120   1.80902900   0.62054911
    140   1.80903417   0.62054579
    160   1.80903752   0.62054364
    180   1.80903982   0.62054216
    200   1.80904146   0.62054111

    inf   1.80904848   0.62053660
>>
```

Convergence is much improved, because the integrand is now analytic on the interval $[0,1]$ (though not periodic).

(d) With $x_k^{(n)}$, $w_k^{(n)}$ denoting the nodes and weights of the n-point Gauss–Legendre quadrature rule on the interval $[0,1]$, we now approximate by

$$I_c \approx 2\sum_{k=1}^{n} w_k^{(n)} \cos([x_k^{(n)}]^2), \quad I_s \approx 2\sum_{k=1}^{n} w_k^{(n)} \sin([x_k^{(n)}]^2).$$

PROGRAM

```
%MAIII_4D
%
f0='%8.0f %20.15f %20.15f\n';
f1='     inf %20.15f %20.15f\n';
disp('       n            Ic                    Is')
Icexact=1.809048475800544; Isexact=0.620536603446762;
n=10;
ab=r_jacobi01(n);
for k=1:n
  xw=gauss(k,ab);
  Ic=2*sum(xw(:,2).*cos(xw(:,1).^2));
  Is=2*sum(xw(:,2).*sin(xw(:,1).^2));
  fprintf(f0,k,Ic,Is)
end
```

```
fprintf('\n')
fprintf(f1,Icexact,Isexact)

 OUTPUT

>> MAIII_4D
      n          Ic                  Is
      1    1.937824843421289    0.494807918509046
      2    1.811712813751607    0.627311992455286
      3    1.808856161750068    0.620553770242084
      4    1.809048965080315    0.620532893432708
      5    1.809048529661320    0.620536620796030
      6    1.809048475486131    0.620536604070278
      7    1.809048475794572    0.620536603442626
      8    1.809048475800590    0.620536603446715
      9    1.809048475800544    0.620536603446763
     10    1.809048475800542    0.620536603446761

    inf    1.809048475800544    0.620536603446762
>>
```

We note a dramatic increase in the speed of convergence, reflecting the rapid convergence of Gaussian quadrature rules for entire functions.

(e) We now let $x_k^{(n)}$, $w_k^{(n)}$ denote the nodes and weights of the n-point Gauss–Jacobi quadrature rule on $[0,1]$ with parameters $\alpha = 0$, $\beta = -\frac{1}{2}$, and approximate

$$I_c = \int_0^1 \frac{\cos x}{\sqrt{x}}\,\mathrm{d}x \approx \sum_{k=1}^{n} w_k^{(n)} \cos(x_k^{(n)})$$

and

$$I_s = \int_0^1 \frac{\sin x}{\sqrt{x}}\,\mathrm{d}x \approx \sum_{k=1}^{n} w_k^{(n)} \sin(x_k^{(n)}).$$

```
 PROGRAM

%MAIII_4E
%
f0='%8.0f %20.15f %20.15f\n';
f1='     inf %20.15f %20.15f\n';
disp('       n           Ic                   Is')
Icexact=1.809048475800544; Isexact=0.620536603446762;
n=10;
ab=r_jacobi01(n,0,-1/2);
for k=1:n
```

```
  xw=gauss(k,ab);
  Ic=sum(xw(:,2).*cos(xw(:,1)));
  Is=sum(xw(:,2).*sin(xw(:,1)));
  fprintf(f0,k,Ic,Is)
end
fprintf('\n')
fprintf(f1,Icexact,Isexact)

 OUTPUT

>> MAIII_4E
      n           Ic                  Is
      1    1.889913892629475    0.654389393592305
      2    1.808616395377709    0.620331018130819
      3    1.809049386208496    0.620537056150696
      4    1.809048474778223    0.620536602926746
      5    1.809048475801256    0.620536603447130
      6    1.809048475800544    0.620536603446762
      7    1.809048475800544    0.620536603446762
      8    1.809048475800543    0.620536603446762
      9    1.809048475800543    0.620536603446762
     10    1.809048475800545    0.620536603446763

    inf    1.809048475800544    0.620536603446762
>>
```

We have the same rapid convergence as in (d), for similar reasons.

5. The composite trapezoidal rule is

$$I_p \approx h\{\tfrac{1}{2}f(0) + \sum_{k=1}^{n-1}(1-t_k)^p f(t_k)\},$$

where $h = \frac{1}{n},\ t_k = kh$. The Gauss–Jacobi formula on $[0,1]$ is

$$I_p \approx \textstyle\sum_{k=1}^{n} w_k^J f(t_k^J),$$

where $t_k^J,\ w_k^J$ are the nodes and weights of the n-point Gauss–Jacobi formula on $[0,1]$ with parameters $\alpha = p,\ \beta = 0$.

```
 PROGRAM

%MAIII_5
%
f0='%6.0f %13.6e %4.0f %22.15e   p=%1.0f\n';
```

```
f1='%6.0f %13.6e %4.0f %22.15e\n';
disp('     n     trapez       n         jacobi')
for p=5:5:20
  ab=r_jacobi01(50,p,0);
  for in=1:5
    nt=10*in; nj=2*in;
    h=1/nt; k=(1:nt-1)'; t=k*h;
    trap=h*sum((1-t).^p.*tan(t));
    xw=gauss(nj,ab);
    jac=sum(xw(:,2).*tan(xw(:,1)));
    if in==1
      fprintf(f0,nt,trap,nj,jac,p)
    else
      fprintf(f1,nt,trap,nj,jac)
    end
  end
  fprintf('\n')
end
```

OUTPUT

```
>> MAIII_5
     n     trapez       n         jacobi
    10  2.370035e-02    2  2.450805768844407e-02   p=5
    20  2.431732e-02    4  2.452509337127442e-02
    30  2.443263e-02    6  2.452511656963370e-02
    40  2.447307e-02    8  2.452511661572217e-02
    50  2.449180e-02   10  2.452511661583684e-02

    10  6.865513e-03    2  7.660916611254523e-03   p=10
    20  7.455964e-03    4  7.661949509972521e-03
    30  7.569823e-03    6  7.661949958329403e-03
    40  7.610014e-03    8  7.661949958700314e-03
    50  7.628677e-03   10  7.661949958700792e-03

    10  2.947505e-03    2  3.698237636934166e-03   p=15
    20  3.495461e-03    4  3.698394382887037e-03
    30  3.606878e-03    6  3.698394410534388e-03
    40  3.646653e-03    8  3.698394410545075e-03
    50  3.665201e-03   10  3.698394410545083e-03

    10  1.479825e-03    2  2.172405672264421e-03   p=20
    20  1.973725e-03    4  2.172443329013678e-03
    30  2.081782e-03    6  2.172443332206555e-03
    40  2.120975e-03    8  2.172443332207202e-03
```

```
        50  2.139363e-03   10  2.172443332207203e-03
>>
```

The Gauss rule is incomparably faster than the trapezoidal rule and, unlike the trapezoidal rule, gains in speed as p increases.

6. See the text.

EXERCISES AND MACHINE ASSIGNMENTS TO CHAPTER 4

EXERCISES

1. The following sequences all converge to zero as $n \to \infty$:

$$v_n = n^{-10}, \quad w_n = 10^{-n}, \quad x_n = 10^{-n^2}, \quad y_n = n^{10} \cdot 3^{-n}, \quad z_n = 10^{-3 \cdot 2^n}.$$

Indicate the type of convergence by placing a check mark in the appropriate position in the following table.

Type of Convergence	v	w	x	y	z
sublinear					
linear					
superlinear					
quadratic					
cubic					
none of the above					

2. Suppose a positive sequence $\{\varepsilon_n\}$ converges to zero with order $p > 0$. Does it then also converge to zero with order p' for any $0 < p' < p$?

3. The sequence $\varepsilon_n = \mathrm{e}^{-\mathrm{e}^n}$, $n = 0, 1, \ldots$, clearly converges to zero as $n \to \infty$. What is the order of convergence?

4. Give an example of a positive sequence $\{\varepsilon_n\}$ converging to zero in such a way that $\lim\limits_{n\to\infty} \dfrac{\varepsilon_{n+1}}{\varepsilon_n^p} = 0$ for some $p > 1$, but not converging (to zero) with any order $p' > p$.

5. Suppose $\{x_n\}$ converges linearly to α, in the sense that $\lim_{n\to\infty} \frac{x_{n+1}-\alpha}{x_n-\alpha} = c,\ 0 < |c| < 1$.

   (a) Define $x_n^* = \frac{1}{2}(x_n + x_{n-1}),\ n = 1, 2, 3, \ldots$. Clearly, $x_n^* \to \alpha$. Does $\{x_n^*\}$ converge appreciably faster than $\{x_n\}$? Explain by determining the asymptotic error constant.

   (b) Do the same for $x_n^* = \sqrt{x_n x_{n-1}}$, assuming $x_n > 0$ for all n, and $\alpha > 0$.

6. Let $\{x_n\}$ be a sequence converging to α linearly with asymptotic error constant c,

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = c, \quad |c| < 1,$$

and assume that $x_n \neq \alpha$ for all n.

   (a) Derive Aitken's Δ^2-process (4.21) by assuming two consecutive ratios in the above limit relation (say, for n and $n+1$) to be equal to c.

   (b) Show that the sequence $\{x_n'\}$ in Aitken's Δ^2-process is well defined for n sufficiently large.

(c) Prove (4.22).

7. Given an iterative method of order p and asymptotic error constant $c \neq 0$, define a new iterative method consisting of m consecutive steps of the given method. Determine the order of this new iterative method and its asymptotic error constant. Hence justify the definition of the efficiency index given near the end of Sect. 4.2.

8. Consider the equation

$$\frac{1}{x-1} + \frac{2}{x+3} + \frac{4}{x-5} - 1 = 0.$$

   (a) Discuss graphically the number of real roots and their approximate location.

   (b) Are there any complex roots?

9. Consider the equation $x \tan x = 1$.

   (a) Discuss the real roots of this equation: their number, approximate location, and symmetry properties. Use appropriate graphs.

   (b) How many bisections would be required to find the smallest positive root to within an error of $\frac{1}{2} \times 10^{-8}$? (Indicate the initial approximations.) Is your answer valid for all roots?

   (c) Are there any complex roots? Explain!

10. Consider the quadratic equation $x^2 - p = 0$, $p > 0$. Suppose its positive root $\alpha = \sqrt{p}$ is computed by the method of false position starting with two numbers a, b satisfying $0 < a < \alpha < b$. Determine the asymptotic error constant c as a function of b and α. What are the conditions on b for $0 < c < \frac{1}{2}$ to hold, that is, for the method of false position to be (asymptotically) faster than the bisection method?

11. The equation $x^2 - a = 0$ (for the square root $\alpha = \sqrt{a}$) can be written equivalently in the form

$$x = \varphi(x)$$

in many different ways, for example:

$$\varphi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right); \qquad \varphi(x) = \frac{a}{x}; \qquad \varphi(x) = 2x - \frac{a}{x}.$$

Discuss the convergence (or nonconvergence) behavior of the iteration $x_{n+1} = \varphi(x_n)$, $n = 0, 1, 2, \ldots$, for each of these three iteration functions. In case of convergence, determine the order of convergence.

12. Under the assumptions of Theorem 4.5.1, show that the secant method cannot converge faster than with order $p = \frac{1}{2}(1 + \sqrt{5})$ if the (simple) root α of $f(x) = 0$ satisfies $f''(\alpha) \neq 0$.

13. Let $\{x_n\}$ be a sequence converging to α. Suppose the errors $e_n = |x_n - \alpha|$ satisfy $e_{n+1} \leq M e_n^2 e_{n-1}$ for some constant $M > 0$. What can be said about the order of convergence?

14. Suppose the equation $f(x) = 0$ has a simple root α, and $f''(\alpha) = 0$, $f'''(\alpha) \neq 0$. Provide heuristics in the manner of the text preceding Theorem 4.5.1 showing that the secant method in this case converges quadratically.

15. (a) Consider the iteration $x_{n+1} = x_n^3$. Give a detailed discussion of the behavior of the sequence $\{x_n\}$ in dependence of x_0.

    (b) Do the same as (a), but for $x_{n+1} = x_n^{1/3}$, $x_0 > 0$.

16. Consider the iteration

$$x_{n+1} = \varphi(x_n), \qquad \varphi(x) = \sqrt{2 + x}.$$

    (a) Show that for any positive x_0 the iterates x_n remain on the same side of $\alpha = 2$ as x_0 and converge monotonically to α.

    (b) Show that the iteration converges globally, that is, for any $x_0 > 0$, and not faster than linearly (unless $x_0 = 2$).

    (c) If $0 < x_0 < 2$, how many iteration steps are required to obtain α with an error less than 10^{-10}?

17. Consider the equation $x = \cos x$.

    (a) Show graphically that there exists a unique positive root α. Indicate, approximately, where it is located.

    (b) Prove local convergence of the iteration $x_{n+1} = \cos x_n$.

    (c) For the iteration in (b) prove: if $x_n \in [0, \frac{\pi}{2}]$, then

$$|x_{n+1} - \alpha| < \left(\sin \frac{\alpha + \pi/2}{2}\right) |x_n - \alpha|.$$

    In particular, one has global convergence on $[0, \frac{\pi}{2}]$.

    (d) Show that Newton's method applied to $f(x) = 0$, $f(x) = x - \cos x$, also converges globally on $[0, \frac{\pi}{2}]$.

18. Consider the equation

$$x = \mathrm{e}^{-x}.$$

    (a) Show that there is a unique real root α and determine an interval containing it.

    (b) Show that the fixed point iteration $x_{n+1} = \mathrm{e}^{-x_n}$, $n = 0, 1, 2, \ldots$, converges locally to α and determine the asymptotic error constant.

    (c) Illustrate graphically that the iteration in (b) actually converges globally, that is, for arbitrary $x_0 > 0$. Then prove it.

(d) An equivalent equation is

$$x = \ln \tfrac{1}{x}\,.$$

Does the iteration $x_{n+1} = \ln \frac{1}{x_n}$ also converge locally? Explain.

19. Consider the equation

$$\tan x + \lambda x = 0, \qquad 0 < \lambda < 1.$$

(a) Show graphically, as simply as possible, that in the interval $[\frac{1}{2}\pi, \pi]$ there is exactly one root α.

(b) Does Newton's method converge to the root $\alpha \in [\frac{1}{2}\pi, \pi]$ if the initial approximation is taken to be $x_0 = \pi$? Justify your answer.

20. Consider the equation

$$f(x) = 0, \qquad f(x) = \ln^2 x - x - 1, \quad x > 0.$$

(a) Graphical considerations suggest that there is exactly one positive root α, and that $0 < \alpha < 1$. Prove this.

(b) What is the largest positive $b \leq 1$ such that Newton's method, started with $x_0 = b$, converges to α?

21. Consider "Kepler's equation"

$$f(x) = 0, \qquad f(x) = x - \varepsilon \sin x - \eta, \quad 0 < |\varepsilon| < 1, \quad \eta \in \mathbb{R},$$

where ε, η are parameters constrained as indicated.

(a) Show that for each ε, η there is exactly one real root $\alpha = \alpha(\varepsilon, \eta)$. Furthermore, $\eta - |\varepsilon| \leq \alpha(\varepsilon, \eta) \leq \eta + |\varepsilon|$.

(b) Writing the equation in fixed point form

$$x = \varphi(x), \qquad \varphi(x) = \varepsilon \sin x + \eta,$$

show that the fixed point iteration $x_{n+1} = \varphi(x_n)$ converges for arbitrary starting value x_0.

(c) Let m be an integer such that $m\pi < \eta < (m+1)\pi$. Show that Newton's method with starting value

$$x_0 = \begin{cases} (m+1)\pi & \text{if } (-1)^m \varepsilon > 0, \\ m\pi & \text{otherwise} \end{cases}$$

is guaranteed to converge (monotonically) to $\alpha(\varepsilon, \eta)$.

(d) Estimate the asymptotic error constant c of Newton's method.

22. (a) Devise an iterative scheme, using only addition and multiplication, for computing the reciprocal $\frac{1}{a}$ of some positive number a. {*Hint*: use Newton's method. For a cubically convergent scheme, see Ex. 40.}

    (b) For what positive starting values x_0 does the algorithm in (a) converge? What happens if $x_0 < 0$?

    (c) Since in (binary) floating-point arithmetic it suffices to find the reciprocal of the mantissa, assume $\frac{1}{2} \le a < 1$. Show, in this case, that the iterates x_n satisfy

$$\left| x_{n+1} - \frac{1}{a} \right| < \left| x_n - \frac{1}{a} \right|^2, \quad \text{all } n \ge 0.$$

    (d) Using the result of (c), estimate how many iterations are required, at most, to obtain $1/a$ with an error less than 2^{-48}, if one takes $x_0 = \frac{3}{2}$.

23. (a) If $A > 0$, then $\alpha = \sqrt{A}$ is a root of either equation

$$x^2 - A = 0, \qquad \frac{A}{x^2} - 1 = 0.$$

    Explain why Newton's method applied to the first equation converges for arbitrary starting value $x_0 > 0$, whereas the same method applied to the second equation produces positive iterates x_n converging to α only if x_0 is in some interval $0 < x_0 < b$. Determine b.

    (b) Do the same as (a), but for the cube root $\sqrt[3]{A}$ and the equations

$$x^3 - A = 0, \qquad \frac{A}{x^3} - 1 = 0.$$

24. (a) Show that Newton's iteration

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \qquad a > 0,$$

    for computing the square root $\alpha = \sqrt{a}$ satisfies

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

    Hence, directly obtain the asymptotic error constant.

    (b) What is the analogous formula for the cube root?

25. Consider Newton's method

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \qquad a > 0,$$

    for computing the square root $\alpha = \sqrt{a}$. Let $d_n = x_{n+1} - x_n$.

(a) Show that

$$x_n = \frac{a}{d_n + \sqrt{d_n^2 + a}} \ .$$

(b) Use (a) to show that

$$|d_n| = \frac{d_{n-1}^2}{2\sqrt{d_{n-1}^2 + a}} \ , \quad n = 1, 2, \ldots \ .$$

Discuss the significance of this result with regard to the overall behavior of Newton's iteration.

26. (a) Derive the iteration that results by applying Newton's method to $f(x) := x^3 - a = 0$ to compute the cube root $\alpha = a^{\frac{1}{3}}$ of $a > 0$.

(b) Consider the equivalent equation $f_\lambda(x) = 0$, where $f_\lambda(x) = x^{3-\lambda} - ax^{-\lambda}$, and determine λ so that Newton's method converges cubically. Write down the resulting iteration in its simplest form.

27. Consider the two (equivalent) equations

$$\text{(A)} \quad x \ln x - 1 = 0 \qquad \text{(B)} \quad \ln x - \frac{1}{x} = 0.$$

(a) Show that there is exactly one positive root and find a rough interval containing it.

(b) For both (A) and (B), determine the largest interval on which Newton's method converges. {*Hint*: investigate the convexity of the functions involved.}

(c) Which of the two Newton iterations converges asymptotically faster?

28. Prove Theorem 4.6.1.

29. Consider the equation

$$f(x) = 0, \quad \text{where} \quad f(x) = \tan x - cx, \quad 0 < c < 1.$$

(a) Show that the smallest positive root α is in the interval $(\pi, \frac{3}{2}\pi)$.

(b) Show that Newton's method started at $x_0 = \pi$ is guaranteed to converge to α if c is small enough. Exactly how small does c have to be?

30. We saw in Sect. 4.1.1 that the equation

$$(*) \qquad \cos x \cosh x - 1 = 0$$

has exactly two roots $\alpha_n < \beta_n$ in each interval $[-\frac{\pi}{2} + 2n\pi, \frac{\pi}{2} + 2n\pi]$, $n = 1, 2, 3 \ldots$. Show that Newton's method applied to (*) converges to α_n when initialized by $x_0 = -\frac{\pi}{2} + 2n\pi$, and to β_n when initialized by $x_0 = \frac{\pi}{2} + 2n\pi$.

31. In the engineering of circular shafts the following equation is important for determining critical angular velocities:

$$f(x) = 0, \quad f(x) = \tan x + \tanh x, \quad x > 0.$$

(a) Show that there are infinitely many positive roots, exactly one, α_n, in each interval $[(n - \frac{1}{2})\pi, n\pi]$, $n = 1, 2, 3, \ldots$.

(b) Determine $\lim_{n\to\infty}(n\pi - \alpha_n)$.

(c) Discuss the convergence of Newton's method when started at $x_0 = n\pi$.

32. The equation

$$f(x) := x \tan x - 1 = 0,$$

if written as $\tan x = 1/x$ and each side plotted separately, can be seen to have infinitely many positive roots, one, α_n, in each interval $[n\pi, (n + \frac{1}{2})\pi]$, $n = 0, 1, 2, \ldots$.

(a) Show that the smallest positive root α_0 can be obtained by Newton's method started at $x_0 = \frac{\pi}{4}$.

(b) Show that Newton's method started with $x_0 = (n + \frac{1}{4})\pi$ converges monotonically decreasing to α_n if $n \geq 1$.

(c) Expanding α_n (formally) in inverse powers of πn,

$$\alpha_n = \pi n + c_0 + c_1(\pi n)^{-1} + c_2(\pi n)^{-2} + c_3(\pi n)^{-3} + \cdots ,$$

determine $c_0, c_1, c_2, \ldots, c_9$. {*Hint*: use the Maple `series` command.}

(d) Use the Matlab function `fzero` to compute α_n for $n = 1 : 10$ and compare the results with the approximation furnished by the expansion in (c).

33. Consider the equation

$$f(x) = 0, \quad f(x) = x \sin x - 1, \quad 0 \leq x \leq \pi.$$

(a) Show graphically (as simply as possible) that there are exactly two roots in the interval $[0, \pi]$ and determine their approximate locations.

(b) What happens with Newton's method when it is started with $x_0 = \frac{1}{2}\pi$? Does it converge, and if so, to which root? Where do you need to start Newton's method to get the other root?

34. (Gregory, 1672) For an integer $n \geq 1$, consider the equation

$$f(x) = 0, \quad f(x) = x^{n+1} - b^n x + ab^n, \quad a > 0, \quad b > 0.$$

(a) Prove that the equation has exactly two distinct positive roots if and only if

$$a < \frac{n}{(n+1)^{1+\frac{1}{n}}}\, b.$$

{*Hint*: analyze the convexity of f.}

(b) Assuming that the condition in (a) holds, show that Newton's method converges to the smaller positive root, when started at $x_0 = a$, and to the larger one, when started at $x_0 = b$.

35. Suppose the equation $f(x) = 0$ has the root α with exact multiplicity $m \geq 2$, and Newton's method converges to this root. Show that convergence is linear, and determine the asymptotic error constant.

36. (a) Let α be a double root of the equation $f(x) = 0$, where f is sufficiently smooth near α. Show that the "doubly-relaxed" Newton's method

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

if it converges to α, does so at least quadratically. Obtain the condition under which the order of convergence is exactly 2, and determine the asymptotic error constant c in this case.

(b) What are the analogous statements in the case of an m-fold root?

37. Consider the equation $x \ln x = a$.

(a) Show that for each $a > 0$ the equation has a unique positive root, $x = x(a)$.

(b) Prove that

$$x(a) \sim \frac{a}{\ln a} \quad \text{as} \quad a \to \infty$$

(i.e., $\lim_{a\to\infty} \frac{x(a)\ln a}{a} = 1$). {*Hint*: use the rule of Bernoulli-L'Hospital.}

(c) For large a improve the approximation given in (b) by applying one step of Newton's method.

38. The equation $x^2 - 2 = 0$ can be written as a fixed point problem in different ways, for example,

$$\text{(a)} \quad x = \frac{2}{x} \qquad \text{(b)} \quad x = x^2 + x - 2 \qquad \text{(c)} \quad x = \frac{x+2}{x+1}\ .$$

How does the fixed point iteration perform in each of these three cases? Be as specific as you can.

39. Show that

$$x_{n+1} = \frac{x_n(x_n^2 + 3a)}{3x_n^2 + a}\ , \quad n = 0, 1, 2, \ldots\ ,$$

is a method for computing $\alpha = \sqrt{a}$, $a > 0$, which converges cubically to α (for suitable x_0). Determine the asymptotic error constant. (Cf. also Ex. 43(e) with $\lambda = 2$.)

40. Consider the fixed point iteration

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, 2, \dots ,$$

where

$$\varphi(x) = Ax + Bx^2 + Cx^3.$$

(a) Given a positive number α, determine the constants A, B, C such that the iteration converges locally to $1/\alpha$ with order $p = 3$. {This will give a cubically convergent method for computing the reciprocal $1/\alpha$ of α using only addition, subtraction, and multiplication.}

(b) Determine the precise condition on the initial error $\varepsilon_0 = x_0 - \frac{1}{\alpha}$ for the iteration to converge.

41. The equation $f(x) := x^2 - 3x + 2 = 0$ has the roots 1 and 2. Written in fixed point form $x = \frac{1}{\omega}(x^2 - (3 - \omega)x + 2)$, $\omega \neq 0$, it suggests the iteration

$$x_{n+1} = \frac{1}{\omega}(x_n^2 - (3 - \omega)x_n + 2), \quad n = 0, 1, 2, \dots \quad (\omega \neq 0).$$

(a) Identify as large an ω-interval as possible such that for any ω in this interval the iteration converges to 1 (when $x_0 \neq 1$ is suitably chosen).

(b) Do the same as (a), but for the root 2 (and $x_0 \neq 2$).

(c) For what value(s) of ω does the iteration converge quadratically to 1?

(d) Interpret the algorithm produced in (c) as a Newton iteration for some equation $F(x) = 0$, and exhibit F. Hence discuss for what initial values x_0 the method converges.

42. Let α be a simple zero of f and $f \in C^p$ near α, where $p \geq 3$. Show: if $f''(\alpha) = \cdots = f^{(p-1)}(\alpha) = 0$, $f^{(p)}(\alpha) \neq 0$, then Newton's method applied to $f(x) = 0$ converges to α locally with order p. Determine the asymptotic error constant.

43. The iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n) - \frac{1}{2}f''(x_n)\frac{f(x_n)}{f'(x_n)}} , \quad n = 0, 1, 2, \dots ,$$

for solving the equation $f(x) = 0$ is known as *Halley's method*.

(a) Interpret Halley's method geometrically as the intersection with the x-axis of a hyperbola with asymptotes parallel to the x- and y-axes that is osculatory to the curve $y = f(x)$ at $x = x_n$ (i.e., is tangent to the curve at this point and has the same curvature there).

(b) Show that the method can, alternatively, be interpreted as applying Newton's method to the equation $g(x) = 0$, $g(x) := f(x)/\sqrt{f'(x)}$.

(c) Assuming α is a simple root of the equation, and $x_n \to \alpha$ as $n \to \infty$, show that convergence is exactly cubic, unless the "Schwarzian derivative"

$$(Sf)(x) := \frac{f'''(x)}{f'(x)} - \frac{3}{2}\left(\frac{f''(x)}{f'(x)}\right)^2$$

vanishes at $x = \alpha$, in which case the order of convergence is larger than three.

(d) Is Halley's method more efficient than Newton's method as measured in terms of the efficiency index?

(e) How does Halley's method look in the case $f(x) = x^\lambda - a$, $a > 0$? (Compare with Ex. 39.)

44. Let $f(x) = x^d + a_{d-1}x^{d-1} + \cdots + a_0$ be a polynomial of degree $d \geq 2$ with real coefficients a_i.

(a) In analogy to (4.75), let

$$f(x) = (x^2 - tx - s)(x^{d-2} + b_{d-1}x^{d-3} + \cdots + b_2) + b_1(x - t) + b_0.$$

Derive a recursive algorithm for computing $b_{d-1}, b_{d-2}, \ldots, b_1, b_0$ in this order.

(b) Suppose α is a complex zero of f. How can f be deflated to remove the pair of zeros α, $\overline{\alpha}$?

(c) (Bairstow's method) Devise a method based on the division algorithm of (a) to compute a quadratic factor of f. Use Newton's method for a system of two equations in the two unknowns t and s, and exhibit recurrence formulae for computing the elements of the 2×2 Jacobian matrix of the system.

45. Let $p(t)$ be a monic polynomial of degree n. Let $\boldsymbol{x} \in \mathbb{C}^n$ and define

$$f_\nu(\boldsymbol{x}) = [x_1, x_2, \ldots, x_\nu]\, p, \qquad \nu = 1, 2, \ldots, n,$$

to be the divided differences of p relative to the coordinates x_μ of $\boldsymbol{x}$. Consider the system of equations

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}, \qquad [\boldsymbol{f}(\boldsymbol{x})]^T = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})].$$

(a) Let $\boldsymbol{\alpha}^T = [\alpha_1, \alpha_2, \ldots, \alpha_n]$ be the zeros of p, assumed mutually distinct. Show that $\boldsymbol{\alpha}$ is, except for a permutation of the components, the unique solution of $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$. {*Hint*: use Newton's formula of interpolation.}

(b) Describe the application of Newton's iterative method to the preceding system of nonlinear equations, $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$. {*Hint*: use Ch. 2, Ex. 58.}

(c) Discuss to what extent the procedure in (a) and (b) is valid for nonpolynomial functions p.

46. For the equation $f(x) = 0$ define

$$\begin{aligned} &y^{[0]}(x) = x, \\ &y^{[1]}(x) = \frac{1}{f'(x)}\ , \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &y^{[m]}(x) = \frac{1}{f'(x)}\ \frac{\mathrm{d}}{\mathrm{d}x}\, y^{[m-1]}(x), \quad m = 2, 3, \ldots\ . \end{aligned}$$

Consider the iteration function

$$\varphi_r(x) := \sum_{m=0}^{r} (-1)^m\ \frac{y^{[m]}(x)}{m!} [f(x)]^m.$$

(When $r = 1$ this is the iteration function for Newton's method.) Show that $\varphi_r(x)$ defines an iteration $x_{n+1} = \varphi_r(x_n),\ n = 0, 1, 2, \ldots$, converging locally with exact order $p = r + 1$ to a root α of the equation if $y^{[r+1]}(\alpha) f'(\alpha) \neq 0$.

MACHINE ASSIGNMENTS

1. (a) Write a Matlab program that computes (in Matlab double precision) the expanded form $p(x) = x^5 - 5x^4 + 10x^3 - 10x^2 + 5x - 1$ of the polynomial $(x - 1)^5$. Run the program to print $p(x)$/prec for 200 equally spaced x-values in a small neighborhood of $x = 1$ (say, $.9986 \leq x \leq 1.0014$), where prec = eps is the Matlab (double-precision) machine precision. Prepare a piecewise linear plot of the results. Explain what you observe. What is the "uncertainty interval" for the numerical root corresponding to the mathematical root $x = 1$?

   (b) Do the same as (a), but for the polynomial $p(x) = x^5 - 100x^4 + 3995x^3 - 79700x^2 + 794004x - 3160080$, the expanded form of $(x-18)(x-19)(x-20)(x-21)(x-22)$. Do the computation in Matlab single precision and take for prec the respective machine precision. Examine a small interval around $x = 22$ (say, $21.675 \leq x \leq 22.2$).

2. Consider the equation

$$\tfrac{1}{2}x - \sin x = 0.$$

   (a) Show that the only positive root is located in the interval $[\frac{1}{2}\pi, \pi]$.

   (b) Compute the root to 7, 15, and 33 decimal places

      (b1) by the method of bisection, using the Matlab function `sbisec` of Sect. 4.3.1 with starting values $a = \frac{1}{2}\pi$, $b = \pi$;

      (b2) by the method of false position, using the Matlab function `sfalsepos` of Sect. 4.4 with the same starting values as in (b1);

(b3) by the secant method, using the Matlab function `ssecant` of Sect. 4.5 with the same starting values as in (b1);

(b4) by Newton's method, using the Matlab function `snewton` of Sect. 4.6 with an appropriate starting value a.

In all cases print the number of iterations required.

3. For an integer $n \geq 2$, consider the equation

$$\frac{x + x^{-1}}{x^n + x^{-n}} = \frac{1}{n} \ .$$

(a) Write the equation equivalently as a polynomial equation, $p_n(x) = 0$.

(b) Use Descartes' rule of sign[1] (applied to $p_n(x) = 0$) to show that there are exactly two positive roots, one in (0,1), the other in (1,∞). How are they related? Denote the larger of the two roots by α_n (> 1). It is known (you do not have to prove this) that

$$1 < \alpha_{n+1} < \alpha_n < 3, \quad n = 2, 3, 4, \ldots \ .$$

(c) Write and run a program applying the bisection method to compute α_n, $n = 2, 3, \ldots, 20$, to six correct decimal places after the decimal point, using $[1, 3]$ as initial interval for α_2, and $[1, \alpha_n]$ as initial interval for α_{n+1} ($n \geq 2$). For each n count the number of iterations required. Similarly, apply Newton's method (to the equation $p_n(x) = 0$) to compute α_n to the same accuracy, using the initial value 3 for α_2 and the initial value α_n for α_{n+1} ($n \geq 2$). (Justify these choices.) Again, for each n, count the number of iterations required. In both cases, print n, α_n, and the number of iterations. Use a `do while` loop to program either method.

4. Consider the equation

$$x = \mathrm{e}^{-x}.$$

(a) Implement the fixed-point iteration $x_{n+1} = \mathrm{e}^{-x_n}$ on the computer, starting with $x_0 = 1$ and stopping at the first n for which x_{n+1} agrees with x_n to within the machine precision. Print this value of n and the corresponding x_{n+1}.

(b) If the equation is multiplied by ω ($\neq 0$ and $\neq -1$) and x is added on both sides, one gets the equivalent equation

$$x = \frac{\omega \mathrm{e}^{-x} + x}{1 + \omega} \ .$$

Under what condition on ω does the fixed-point iteration for this equation converge faster (ultimately) than the iteration in (a)? {This condition involves the root α of the equation.}

[1] Descartes' rule of sign says that if a real polynomial has s sign changes in the sequence of its nonzero coefficients, then it has s positive zeros or a (nonnegative) even number less.

(c) What is the optimal choice of ω? Verify it by a machine computation in a manner analogous to (a).

5. Consider the boundary value problem

$$y'' + \sin y = 0, \quad 0 \le x \le \tfrac{1}{4}\pi,$$
$$y(0) = 0, \quad y(\tfrac{1}{4}\pi) = 1,$$

which describes the angular motion of a pendulum.

(a) Use the Matlab integrator `ode45.m` to compute and plot the solution $u(x; s)$ of the associated initial value problem

$$u'' + \sin u = 0, \quad 0 \le x \le \tfrac{1}{4}\pi,$$
$$u(0) = 0, \quad u'(0) = s$$

for $s = .2(.2)2$.

(b) Write and run a Matlab program that applies the method of bisection, with tolerance .5e−12, to the equation $f(s) = 0$, $f(s) = u(1; s) - 1$. Use the plots of (a) to choose starting values s_0, s_1 such that $f(s_0) < 0$, $f(s_1) > 0$. Print the number of bisections and the value of s so obtained. {*Suggestion*: use a nonsymbolic version `bisec.m` of the program `sbisec.m` of Sect. 4.3.1.}

(c) Plot the solution curve $y(x) = u(x; s)$ of the boundary value problem, with s as obtained in (b).

6. The boundary value problem

$$y'' = g(x, y, y'), \quad 0 \le x \le 1,$$
$$y(0) = y_0, \quad y(1) = y_1$$

may be discretized by replacing the first and second derivatives by centered difference quotients relative to a grid of equally spaced points $x_k = \frac{k}{n+1}$, $k = 0, 1, \ldots, n, n + 1$.

(a) Interpret the resulting equations as a fixed point problem in $\mathbb{R}^n$ and formulate the respective fixed point iteration.

(b) Write a Matlab program that applies the fixed point iteration of (a) to the problem

$$y'' - y = 0, \quad y(0) = 0, \quad y(1) = 1$$

(cf. the first Example of Ch. 7, Sect. 7.1.1). Run the program for $n = 10, 100, 1\,000, 10\,000$, stopping the iteration the first time two successive iterates differ by less than .5e−14 in the ∞-norm. Print the number of iterations required and the maximum error of the final iterate as

an approximation to the exact solution vector. Assuming this error is $O(h^p)$, determine numerically the values of p and of the constant implied in the order term. {*Suggestion*: for solving tridiagonal systems of equations, use the Matlab routine `tridiag.m` of Ch. 2, MA 8(a).}

(c) Apply the fixed point iteration of (a) to the boundary value problem of MA 5. Show that the iteration function is contractive. {*Hint*: use the fact that the symmetric $n \times n$ tridiagonal matrix $\boldsymbol{A}$ with elements -2 on the diagonal and 1 on the two side diagonals has an inverse satisfying $\|\boldsymbol{A}^{-1}\|_\infty \leq (n+1)^2/8$.}

7. (a) Solve the finite difference equations obtained in MA 6(c) by Newton's method, using $n = 10, 100, 1000$ and an error tolerance of .5e−14. Print the number of iterations in each case and plot the respective solution curves. {*Suggestion*: same as in MA 6(b).}

(b) Do the same as in (a) for the boundary value problem

$$y'' = yy', \quad y(0) = 0, \quad y(1) = 1,$$

but with $n = 10, 50, 100$ and error tolerance .5e−6. How would you check your program for correctness?

(c) Show that the fixed point iteration applied to the finite difference equations for the boundary value problem of (b) does not converge. {*Hint*: use $n^2/8 \leq \|\boldsymbol{A}^{-1}\|_\infty \leq (n+1)^2/8$ for the $n \times n$ tridiagonal matrix $\boldsymbol{A}$ of MA 6(c).}

8. (a) The *Littlewood–Salem–Izumi constant* α_0, defined as the unique solution in $0 < \alpha < 1$ of

$$\int_0^{3\pi/2} \frac{\cos t}{t^\alpha}\, \mathrm{d}t = 0,$$

is of interest in the theory of positive trigonometric sums (cf., e.g., Koumandos[2011, Theorem 2]). Use Newton's method in conjunction with Gaussian quadrature to compute α_0. {*Hint*: you need the Matlab routine `gauss.m` along with the routines `r_jacobi01.m`, `r_jacobi.m`, `r_jaclog.m`, and `mm_log.m` to do the integrations required. All these routines are available on the web at `http://www.cs.purdue.edu/archives/2002/wxg/codes/SOPQ.html`. }

(b) Do the same as in (a) for the constant α_1, the unique solution in $0 < \alpha < 1$ of

$$\int_0^{5\pi/4} \frac{\cos(t + \pi/4)}{t^\alpha}\, \mathrm{d}t = 0$$

(cf. *ibid*, Theorem 4).

9. (a) Discuss how to simplify the system of nonlinear equations (4.17) for the recurrence coefficients of the polynomials $\{\pi_{k,n}\}_{k=0}^n$, generating the s-orthogonal polynomials $\pi_n = \pi_{n,n}$, when the measure $\mathrm{d}\lambda(t) = w(t)\mathrm{d}t$

is symmetric, i.e., the support of $\mathrm{d}\lambda$ is an interval $[-a, a]$, $0 < a \leq \infty$, and $w(-t) = w(t)$ for all t with $0 < t \leq a$. {*Hint*: first show that the respective monic s-orthogonal polynomial π_n satisfies $\pi_n(-t) = (-1)^n \pi_n(t)$, $t \in [-a, a]$ and similarly $\pi_{k,n}(-t) = (-1)^k \pi_{k,n}(t)$.}

(b) For $n = 2$, $s = 1$ and $s = 2$, explain how the recurrence coefficients β_0, β_1 can be obtained analytically in terms of the moments $\mu_k = \int_{\mathbb{R}} t^k \mathrm{d}\lambda(t)$, $k = 0, 1, 2, \ldots$, of the measure. Provide numerical answers in the case of the Legendre measure $\mathrm{d}\lambda(t) = \mathrm{d}t$, $t \in [-1, 1]$.

(c) Write a Matlab program for solving the system of nonlinear equations in (a), using the program `fsolve` of the Matlab Optimization Toolbox. Run the program for the Legendre measure and for $n = 2 : 10$ and $s = 1$ and $s = 2$ for each n. Choose initial approximations as deemed useful and apply appropriate $(s+1)n$-point Gaussian quadrature rules to do the necessary integrations. Print $\beta_0, \beta_1, \ldots, \beta_{n-1}$ and the zeros of π_n.

ANSWERS TO THE EXERCISES AND MACHINE ASSIGNMENTS OF CHAPTER 4

ANSWERS TO EXERCISES

1. As $n \to \infty$, we have

$$\frac{v_{n+1}}{v_n} = \left(\frac{n}{n+1}\right)^{10} \sim 1,$$

$$\frac{w_{n+1}}{w_n} = \frac{1}{10},$$

$$\frac{x_{n+1}}{x_n} = 10^{n^2-(n+1)^2} = 10^{-(2n+1)} \to 0$$

$$\text{and} \quad \frac{x_{n+1}}{x_n^p} = 10^{(p-1)n^2-2n-1} \to \infty \ \text{ for any } p > 1,$$

$$\frac{y_{n+1}}{y_n} = \left(\frac{n+1}{n}\right)^{10} \cdot 3^{-1} \sim \frac{1}{3},$$

$$\frac{z_{n+1}}{z_n^2} = \frac{10^{-3\cdot 2^{n+1}}}{(10^{-3\cdot 2^n})^2} = \frac{10^{-3\cdot 2^{n+1}}}{10^{-3\cdot 2^{n+1}}} = 1.$$

Therefore, the completed table should look as follows.

Type of Convergence	v	w	x	y	z
sublinear	×				
linear		×		×	
superlinear			×		
quadratic					×
cubic					
none of the above					

2. No. By assumption,

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^p} = c, \quad c \neq 0,$$

which implies

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^{p'}} = \frac{\varepsilon_{n+1}}{\varepsilon_n^p}\,\varepsilon_n^{p-p'} \sim c\,\varepsilon_n^{p-p'} \to 0 \ \text{ as } \ n \to \infty \ \text{ when } p' < p.$$

3. We have

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^p} = \frac{\mathrm{e}^{-\mathrm{e}^{n+1}}}{\mathrm{e}^{-p\mathrm{e}^n}} = \mathrm{e}^{-\mathrm{e}^n(\mathrm{e}-p)}.$$

As $n \to \infty$, this tends to a nonzero constant if and only if $p = \mathrm{e}$. Hence, the order of convergence is $\mathrm{e} = 2.71828\ldots$.

4. Take, for example, $\varepsilon_n = \exp(-np^n)$. Then

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^p} = \frac{\exp(-(n+1)p^{n+1})}{\exp(-np^{n+1})} = \exp(-p^{n+1}) \to 0 \ \text{ as } n \to \infty,$$

but

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^{p'}} = \exp(-(n+1)p^{n+1} + np'p^n) = \exp(p^n[(p'-p)n - p]) \to \infty$$

for any $p' > p$.

5. Let $x_n = \alpha + \varepsilon_n$, where

$$\varepsilon_n \to 0, \quad \frac{\varepsilon_{n+1}}{\varepsilon_n} \to c \ \text{ as } n \to \infty.$$

(a) Let $\varepsilon_n^* = \frac{1}{2}(x_n + x_{n-1}) - \alpha$. Then

$$\varepsilon_n^* = \tfrac{1}{2}(\alpha + \varepsilon_n + \alpha + \varepsilon_{n-1}) - \alpha = \tfrac{1}{2}(\varepsilon_n + \varepsilon_{n-1}) = \tfrac{1}{2}\varepsilon_n\left(1 + \tfrac{\varepsilon_{n-1}}{\varepsilon_n}\right),$$

that is,

$$\varepsilon_n^* \sim \tfrac{1}{2}\,\varepsilon_n\left(1 + \frac{1}{c}\right).$$

There follows

$$\frac{\varepsilon_{n+1}^*}{\varepsilon_n^*} \sim \frac{\varepsilon_{n+1}}{\varepsilon_n} \to c.$$

Thus, $\{x_n^*\}$ converges at the same (linear) rate as $\{x_n\}$.

(b) Similarly as in (a), let $\varepsilon_n^* = \sqrt{x_n x_{n-1}} - \alpha$. Then

$$\begin{aligned}
\varepsilon_n^* &= \sqrt{(\alpha+\varepsilon_n)(\alpha+\varepsilon_{n-1})} - \alpha \\
&= \frac{\alpha(\varepsilon_n + \varepsilon_{n-1}) + \varepsilon_n\varepsilon_{n-1}}{\sqrt{(\alpha+\varepsilon_n)(\alpha+\varepsilon_{n-1})} + \alpha} \\
&= \varepsilon_n\,\frac{\alpha\left(1 + \frac{\varepsilon_{n-1}}{\varepsilon_n}\right) + \varepsilon_{n-1}}{\sqrt{(\alpha+\varepsilon_n)(\alpha+\varepsilon_{n-1})} + \alpha} \\
&\sim \tfrac{1}{2}\,\varepsilon_n\left(1 + \frac{1}{c}\right).
\end{aligned}$$

Therefore, as in (a),

$$\frac{\varepsilon_{n+1}^*}{\varepsilon_n^*} \sim \frac{\varepsilon_{n+1}}{\varepsilon_n} \to c.$$

6. See the text.

7. Denote by e_n the error of the given method after n steps. Since the method has order p and asymptotic error constant $c \neq 0$, we have

$$\frac{e_{n+1}}{e_n^p} \to c \quad \text{as} \quad n \to \infty.$$

Now,

$$\begin{aligned}\frac{e_{n+m}}{e_n^{p^m}} &= \prod_{k=1}^{m} \left(\frac{e_{n+k}}{e_{n+k-1}^p} \right)^{p^{m-k}} \\ &= \left(\frac{e_{n+1}}{e_n^p} \right)^{p^{m-1}} \left(\frac{e_{n+2}}{e_{n+1}^p} \right)^{p^{m-2}} \left(\frac{e_{n+3}}{e_{n+2}^p} \right)^{p^{m-3}} \cdots \frac{e_{n+m}}{e_{n+m-1}^p},\end{aligned}$$

since each numerator on the far right cancels against the next denominator, leaving precisely the ratio on the far left. Therefore, when $n \to \infty$,

$$\frac{e_{n+m}}{e_n^{p^m}} \to c^{p^{m-1}+p^{m-2}+\cdots+1} = c^{\frac{p^m-1}{p-1}},$$

so that the new method has order p^m and asymptotic error constant $c^{\frac{p^m-1}{p-1}}$ ($= c^m$ if $p = 1$). If we consider one step of the original method as 1 "operation", then its efficiency index is p. The "new method" then requires m operations per step, hence its efficiency index is $(p^m)^{\frac{1}{m}} = p$, the same as the original method, as it should be.

8. (a) There are three vertical asymptotes at $x = -3$, $x = 1$ and $x = 5$. Near each, the function $f(x) = \frac{1}{x-1} + \frac{2}{x+3} + \frac{4}{x-5} - 1$ approaches $+\infty$ from the right and $-\infty$ from the left. Furthermore, $f'(x) < 0$ and $\lim_{x \to \pm\infty} f(x) = -1$. Therefore, there is exactly one real root in each of the intervals $(-3, 1)$, $(1, 5)$ and $(5, \infty)$.

   (b) If all denominators are cleared, there results a cubic equation. Since three distinct real roots are already accounted for, there can be no additional roots, real or complex.

9. See the text.

10. According to (4.44), when $f(x) = x^2 - p$, we have

$$\begin{aligned}c &= 1 - (b - \alpha)\frac{f'(\alpha)}{f(b)} = 1 - (b - \alpha)\frac{2\alpha}{b^2 - p} \\ &= 1 - \frac{(b - \alpha) \cdot 2\alpha}{(b - \alpha)(b + \alpha)} = 1 - \frac{2\alpha}{b + \alpha} \\ &= \frac{b - \alpha}{b + \alpha}.\end{aligned}$$

Thus, there holds $0 < c < \frac{1}{2}$ precisely if

$$\alpha < b < 3\alpha.$$

11. In the first case, we have

$$\varphi'(\alpha) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right)\Big|_{x=\alpha} = 0,$$
$$\varphi''(\alpha) = a\alpha^{-3} \neq 0.$$

Thus, the fixed point iteration converges quadratically (in fact, it is Newton's method).

In the second case,

$$\varphi'(\alpha) = -\frac{a}{x^2}\Big|_{x=\alpha} = -1.$$

Here, convergence is in doubt. In fact, for any initial value $x_0 \neq 0$, one finds

$$x_1 = \frac{a}{x_0},$$
$$x_2 = \frac{a}{x_1} = \frac{a}{a/x_0} = x_0,$$

that is, the iteration cycles.

In the third case,

$$\varphi'(\alpha) = \left(2 + \frac{a}{x^2}\right)\Big|_{x=\alpha} = 3,$$

and the iteration diverges unless $x_0 = \alpha$.

12. In addition to $M(\varepsilon)$, introduce

$$m(\varepsilon) = \min_{\substack{s\in I_\varepsilon \\ t\in I_\varepsilon}} \left|\frac{f''(s)}{2f'(t)}\right|,$$

which by assumption is positive for ε small enough. Eq. (4.46) then implies

$$|x_{n+1} - \alpha| \geq m|x_n - \alpha||x_{n-1} - \alpha|, \quad m = m(\varepsilon).$$

We can now mimic the arguments in the proof of Theorem 4.5.2: with $E_n = m|x_n - \alpha|$, we have

$$E_{n+1} \geq E_n E_{n-1},$$

hence, by induction,

$$E_n \geq E^{p^n}, \quad E = \min(E_0, E_1^{1/p}), \quad p = \frac{1}{2}(1 + \sqrt{5}).$$

Therefore,

$$|x_n - \alpha| = \frac{1}{m} E_n \geq \frac{1}{m} E^{p^n} =: \varepsilon_n,$$

where ε_n converges to zero with order p,

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^p} = \frac{\frac{1}{m} E^{p^{n+1}}}{\left(\frac{1}{m}\right)^p E^{p^{n+1}}} = \left(\frac{1}{m}\right)^{1-p} \quad \text{as} \quad n \to \infty.$$

13. Define $E_n = M^{\frac{1}{2}} e_n$. Then

$$(*) \qquad E_{n+1} \le E_n^2 E_{n-1}.$$

To guess the order of convergence, assume we have equality in (*). Taking logarithms,

$$y_n = \ln \frac{1}{E_n},$$

then gives $y_{n+1} = 2y_n + y_{n-1}$, a constant-coefficient difference equation whose characteristic equation is $t^2 - 2t - 1 = 0$. There are two solutions $t_i^n,\ i = 1, 2$, corresponding to the two roots $t_{1,2} = 1 \pm \sqrt{2}$. Only the first tends to ∞ as $n \to \infty$. Therefore,

$$y_n = c\,(1 + \sqrt{2})^n$$

for some constant $c > 0$. There follows

$$\ln \frac{1}{E_n} = c\,(1 + \sqrt{2})^n,$$

$$E_n = \mathrm{e}^{-c\,(1+\sqrt{2})^n},$$

$$\frac{E_{n+1}}{E_n^{1+\sqrt{2}}} = \frac{\mathrm{e}^{-c\,(1+\sqrt{2})^{n+1}}}{[\mathrm{e}^{-c\,(1+\sqrt{2})^n}]^{1+\sqrt{2}}} = 1.$$

Thus, E_n, and hence also e_n, converges to zero with order $p = 1 + \sqrt{2}$. Assume now (*) with inequality as shown, and let $p = 1 + \sqrt{2}$. Since

$$p^2 = 2p + 1,$$

one proves by induction, as in the proof of Theorem 4.5.2, that

$$E_n \le E^{p^n}, \qquad E = \max\ (E_0,\, E_1^{\frac{1}{p}}).$$

Therefore,

$$e_n = M^{-1/2} E_n \le M^{-\frac{1}{2}} E^{p^n} =: \varepsilon_n,$$

and

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^p} = \frac{M^{-\frac{1}{2}} E^{p^{n+1}}}{M^{-\frac{p}{2}} (E^{p^n})^p} = M^{\frac{p-1}{2}},$$

showing that ε_n converges to zero with order p, hence e_n at least with order p.

14. Since $f''(\alpha) = 0$, we have $[\alpha, \alpha, \alpha] f = 0$, and therefore

$$\begin{aligned}
[x_{n-1}, x_n, \alpha] f &= [x_{n-1}, x_n, \alpha] f - [\alpha, \alpha, \alpha] f \\
&= [x_{n-1}, x_n, \alpha] f - [\alpha, x_n, \alpha] f + [\alpha, x_n, \alpha] f - [\alpha, \alpha, \alpha] f \\
&= (x_{n-1} - \alpha)[x_{n-1}, x_n, \alpha, \alpha] f + (x_n - \alpha)[x_n, \alpha, \alpha, \alpha] f \\
&= \tfrac{1}{6}(x_{n-1} - \alpha) f'''(\xi_n) + \tfrac{1}{6}(x_n - \alpha) f'''(\eta_n), \quad \xi_n, \eta_n \to \alpha.
\end{aligned}$$

Therefore, by (4.46),

$$x_{n+1}-\alpha=\frac{1}{6}(x_n-\alpha)(x_{n-1}-\alpha)^2\frac{f'''(\xi_n)}{f'(\xi_n')}+\frac{1}{6}(x_n-\alpha)^2(x_{n-1}-\alpha)\frac{f'''(\eta_n)}{f'(\xi_n')}\,.$$

The quotients on the right tend to the same nonzero constant as $n\to\infty$. Letting $e_n=|x_n-\alpha|$ and assuming constancy of these quotients, we get

$$e_{n+1}=C(e_ne_{n-1}^2+e_n^2e_{n-1}),\quad C>0.$$

Let $E_n=\sqrt{C}e_n$. Then, multiplying by $\sqrt{C}$, we obtain

$$E_{n+1}=E_nE_{n-1}^2+E_n^2E_{n-1}.$$

Taking logarithms, and letting $y_n=\ln\frac{1}{E_n}=-\ln E_n$, we get

$$y_{n+1}=y_n+2y_{n-1}-\ln\left(1+\frac{E_n}{E_{n-1}}\right).$$

Since E_n converges to 0 faster than linearly, the y_n tend to ∞ and the log-term above tends to 0. It appears safe, therefore, to consider

$$y_{n+1}=y_n+2y_{n-1}.$$

The characteristic equation $t^2-t-2=0$ has roots -1 and 2, hence $y_n\sim\gamma\cdot 2^n$ for some constant $\gamma>0$ as $n\to\infty$. It follows that $E_n\sim \mathrm{e}^{-\gamma\cdot 2^n}$, and

$$\frac{E_{n+1}}{E_n^2}\sim\frac{\mathrm{e}^{-\gamma\cdot 2^{n+1}}}{\mathrm{e}^{-2\gamma\cdot 2^n}}=1,\quad n\to\infty,$$

that is, E_n converges to zero quadratically, hence also $e_n=\frac{1}{\sqrt{C}}\,E_n$.

15. (a) If $|x_0|>1$, then $x_n\to\operatorname{sgn}(x_0)\cdot\infty$. If $|x_0|=1$, then trivially $x_n=\operatorname{sgn}(x_0)$ for all n. If $|x_0|<1$, then $x_n\to 0$ monotonically decreasing, if $x_0>0$, and monotonically increasing, if $x_0<0$, the order of convergence being $p=3$ in either case.

    (b) Now, x_n converges to $\alpha=1$, monotonically increasing if $0<x_0<1$, monotonically decreasing if $x_0>1$, and trivially if $x_0=1$. Since for the iteration function $\varphi(x)=x^{1/3}$ we have $\varphi'(\alpha)=1/3$, convergence is linear with asymptotic error constant equal to $1/3$.

16. (a) If $0<x_n<2$, then $\sqrt{2}<x_{n+1}=\sqrt{2+x_n}<2$, and if $x_n>2$, then $x_{n+1}>2$, $n\geq 0$. Since for $t>0$ there holds $t^2>t+2$ if and only if $t>2$, and $t^2<t+2$ if and only if $t<2$, we have $x_{n+1}>x_n$, i.e., $x_n^2<2+x_n$, in case $x_n<2$, and $x_{n+1}<x_n$ in case $x_n>2$.

(b) Global convergence follows from the monotonicity of the x_n and the fact that there is only one positive limit, namely $\alpha = 2$. Moreover,

$$x_{n+1} - 2 = \sqrt{2 + x_n} - 2 = \frac{x_n - 2}{\sqrt{2 + x_n} + 2}\,,$$

so that, since $x_n > 0$ for all n,

$$|x_{n+1} - 2| \le \frac{|x_n - 2|}{\sqrt{2} + 2}, \quad n = 0, 1, 2, \dots\ .$$

Repeated application of this inequality gives

$$|x_n - 2| \le \frac{1}{(\sqrt{2} + 2)^n}\,|x_0 - 2|.$$

The speed of convergence cannot be faster than linear, since in the case $0 < x_0 < 2$,

$$|x_{n+1} - 2| \ge \left|\frac{x_n - 2}{\sqrt{2 + 2} + 2}\right| = \tfrac{1}{4}\,|x_n - 2|,$$

and in the case $x_0 > 2$, by the monotonicity of the x_n,

$$|x_{n+1} - 2| \ge \frac{1}{\sqrt{2 + x_0} + 2}\,|x_n - 2|.$$

(c) From the result in (b), if $0 < x_0 < 2$,

$$|x_n - 2| \le \frac{|x_0 - 2|}{(\sqrt{2} + 2)^n} \le \frac{2}{(\sqrt{2} + 2)^n}\,.$$

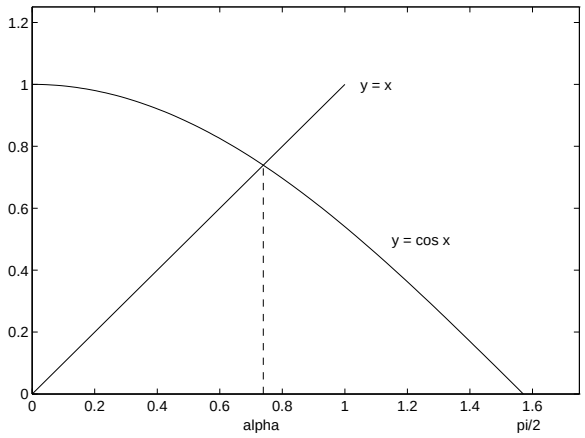
Thus, $|x_n - 2| \le 10^{-10}$ certainly if

$$\frac{2}{(\sqrt{2} + 2)^n} \le 10^{-10}, \quad 2 \times 10^{10} \le (\sqrt{2} + 2)^n,$$

that is, if

$$n \ge \left\lceil \frac{10 + \log 2}{\log(\sqrt{2} + 2)} \right\rceil = 20.$$

17. (a) From the graph below one sees that $\alpha \approx \frac{\pi}{4}$.



(b) The iteration function is $\varphi(x) = \cos x$. Since $\varphi'(x) = -\sin x$, we have $|\varphi'(\alpha)| = \sin\alpha < 1$, implying local convergence.

(c) If $x_0 \in [0, \frac{\pi}{2}]$, then clearly $x_n \in [0, 1]$ for all $n \geq 1$. Furthermore,

$$\begin{aligned} |x_{n+1} - \alpha| &= |\cos x_n - \cos\alpha| = 2\,|\sin\tfrac{1}{2}(x_n + \alpha)\sin\tfrac{1}{2}(x_n - \alpha)| \\ &< \sin\tfrac{1}{2}(x_n + \alpha) \cdot |x_n - \alpha| < \sin\left(\tfrac{\alpha+\pi/2}{2}\right)|x_n - \alpha|. \end{aligned}$$

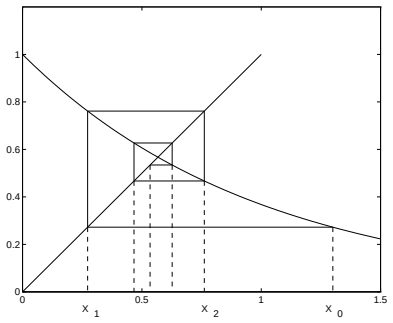
This implies global convergence on $[0, \frac{\pi}{2}]$, since $0 < \sin\left(\frac{\alpha+\pi/2}{2}\right) < 1$.

(d) Since $f(x) = x - \cos x$, $f'(x) = 1 + \sin x$, $f''(x) = \cos x > 0$, the function f is convex on $[0, \frac{\pi}{2}]$ and increases from -1 to $\pi/2$. If $\alpha < x_0 \leq \pi/2$, Newton's method converges monotonically decreasing. If $0 \leq x_0 < \alpha$, the same is true after the first Newton step, since $x_1 > \alpha$ and $x_1 = \cos x_0 \leq 1 < \frac{\pi}{2}$.

18. (a) Letting $f(x) = x - \mathrm{e}^{-x}$, we have $f'(x) = 1 + \mathrm{e}^{-x} \geq 1$ for all real x. Consequently, f increases monotonically on $\mathbb{R}$ from $-\infty$ to $+\infty$, hence has exactly one real zero, α. Given that $f(0) = -1$ and $f(1) = 1 - \mathrm{e}^{-1} > 0$, we have $0 < \alpha < 1$.

(b) The fixed point iteration is $x_{n+1} = \varphi(x_n)$ with $\varphi(x) = \mathrm{e}^{-x}$. Clearly, $\alpha = \varphi(\alpha)$, and $\varphi'(\alpha) = -\mathrm{e}^{-\alpha}$, hence $0 < |\varphi'(\alpha)| = \mathrm{e}^{-\alpha} < 1$. Therefore, we have local convergence, the asymptotic error constant being $c = -\mathrm{e}^{-\alpha}$.

(c) For definiteness, assume $x_0 > \alpha$. Then the fixed point iteration behaves as indicated in the figure below: the iterates "spiral" clockwise around, and into, the fixed point α. The same spiraling takes place if $0 < x_0 < \alpha$ (simply relabel x_1 in the figure as x_0).



Proof of global convergence. From the mean value theorem of calculus, applied to the function e^{-x}, one has

$$|x_{n+1} - \alpha| = |\mathrm{e}^{-x_n} - \mathrm{e}^{-\alpha}| = \mathrm{e}^{-\xi_n}|x_n - \alpha|,$$

where ξ_n is strictly between α and x_n. Letting $\mu = \min(x_0, x_1)$, it is clear from the graph above that $\mu > 0$ and

$$x_n \geq \mu \quad (\text{all } n \geq 0), \quad \alpha > \mu.$$

Therefore, ξ_n being stricly between α and x_n, we have that $\xi_n > \mu$ for all n, hence

$$|x_{n+1} - \alpha| < \mathrm{e}^{-\mu}|x_n - \alpha|.$$

Applying this repeatedly gives $|x_n - \alpha| < \mathrm{e}^{-\mu n}|x_0 - \alpha| \to 0$ as $n \to \infty$.

(d) Here, $x_{n+1} = \psi(x_n)$, where $\psi(x) = \ln \frac{1}{x}$. Since $|\psi'(\alpha)| = \frac{1}{\alpha} > 1$, we cannot have local convergence to α if $x_0 \neq \alpha$.

19. (a) The graphs of $y = \tan x$ and $y = -\lambda x$ for $x > 0$ intersect for the first time at a point whose abscissa, a root of the equation, is between $\frac{1}{2}\pi$ and π. This is the only root in that interval.

(b) With $f(x) = \tan x + \lambda x$, we have, on $[\frac{1}{2}\pi, \pi]$,

$$\begin{aligned} &f(\tfrac{1}{2}\pi + 0) = -\infty, \quad f(\pi) = \lambda\pi, \\ &f'(x) = 1 + \lambda + \tan^2 x > 0, \\ &f''(x) = 2\,\tan x\,(1 + \tan^2 x) < 0, \end{aligned}$$

so that on this interval f is monotonically increasing and concave. If Newton's method is started at the right endpoint, $x_0 = \pi$, then it will converge monotonically increasing if $x_1 > \frac{1}{2}\pi$. This is true, since

$$x_1 = \pi - \frac{f(\pi)}{f'(\pi)} = \pi - \frac{\lambda\pi}{1+\lambda} = \frac{\pi}{1+\lambda} > \frac{\pi}{2},$$

since $0 < \lambda < 1$.

20. (a) Plotting the graphs of $y = \ln^2 x$ and $y = x + 1$ shows that they intersect exactly once, between $x = 0$ and $x = 1$. A formal proof goes as follows. Clearly, $f(0) = +\infty$, $f(1) = -2$, and $f'(x) = \frac{2}{x} \ln x - 1 < 0$ on $(0, 1]$. This shows that there is a unique root between 0 and 1. The proof is complete if we can show that $f'(x) < 0$ also for $x \geq 1$. To do this, we differentiate once more:

$$f''(x) = 2\,\frac{1 - \ln x}{x^2}\,.$$

This is positive for $1 \leq x < \mathrm{e}$ (also, incidentally, for $0 < x < 1$) and negative for $x > \mathrm{e}$. Thus, f' increases on $[1, \mathrm{e})$ and decreases on (e, ∞), that is, has a maximum at $x = \mathrm{e}$, where $f'(\mathrm{e}) = \frac{2}{\mathrm{e}} - 1 < 0$. It follows that f' is negative everywhere on the positive real line.

(b) We have seen in (a) that f is convex on [0,1] and monotonically decreasing from $+\infty$ to -2. We can have $b > \alpha$, but the tangent to f at b must intersect the real line inside the interval $[0, 1]$. The largest b for which this is the case satisfies

$$b - \frac{f(b)}{f'(b)} = 0.$$

This gives

$$b - b\,\frac{\ln^2 b - b - 1}{2 \ln b - b} = 0,$$

that is,

$$1 = \frac{\ln^2 b - b - 1}{2 \ln b - b}\,, \quad 2 \ln b - b = \ln^2 b - b - 1,$$

and thus, finally,

$$\ln^2 b - 2 \ln b - 1 = 0.$$

This is a quadratic equation in $\ln b$, with the only negative root being $\ln b = 1 - \sqrt{2}$. Therefore,

$$b = \mathrm{e}^{1-\sqrt{2}} = .6608598\ldots\,.$$

21. (a) We have $f'(x) = 1 - \varepsilon \cos x$, and therefore $f'(x) > 1 - |\varepsilon| > 0$ for all $x \in \mathbb{R}$. It follows that f monotonically increases on $\mathbb{R}$ from $-\infty$ to $+\infty$, hence has exactly one real root. Since $f(\eta + |\varepsilon|) = |\varepsilon| - \varepsilon \sin(\eta + |\varepsilon|) \geq 0$ and $f(\eta - |\varepsilon|) = -|\varepsilon| - \varepsilon \sin(\eta - |\varepsilon|) \leq 0$, the assertion follows.

(b) We have $\varphi'(x) = \varepsilon \cos x$, hence $|\varphi'(x)| \leq |\varepsilon| < 1$. Since we know that $|x_{n+1} - \alpha| = |\varphi'(\xi_n)||x_n - \alpha|$ for some intermediate ξ_n, there follows $|x_{n+1} - \alpha| \leq |\varepsilon||x_n - \alpha|$, hence $|x_n - \alpha| \leq |\varepsilon|^n |x_0 - \alpha| \to 0$ as $n \to \infty$.

(c) We have $f(m\pi) = m\pi - \eta < 0$ and $f((m+1)\pi) = (m+1)\pi - \eta > 0$. Furthermore, $f''(x) = \varepsilon \sin x$. Therefore, on $(m\pi, (m+1)\pi)$, since $\sin x$

has the sign $(-1)^m$, we have $f''(x) > 0$ if $(-1)^m\varepsilon > 0$, and $f''(x) < 0$ if $(-1)^m\varepsilon < 0$. In the former case, Newton's method converges monotonically decreasing if started at $x_0 = (m+1)\pi$, and in the latter case, monotonically increasing if started at $x_0 = m\pi$.

(d) We have

$$c = \frac{1}{2}\,\frac{f''(\alpha)}{f'(\alpha)} = \frac{1}{2}\,\frac{\varepsilon\sin\alpha}{1-\varepsilon\cos\alpha}\,,$$

thus

$$|c| \le \frac{1}{2}\,\frac{|\varepsilon|}{1-|\varepsilon|}\,.$$

22. (a) Clearly, $\alpha = \frac{1}{a}$ is the solution of the equation $f(x) = 0,\ f(x) = \frac{1}{x} - a$. Newton's method applied to it gives

$$x_{n+1} = x_n - \frac{\frac{1}{x_n} - a}{-\frac{1}{x_n^2}} = x_n + x_n(1-ax_n) = x_n(2-ax_n), \quad n = 0, 1, 2, \ldots\,.$$

(b) From the convexity of f it follows that the method converges to α quadratically and monotonically increasing (except, possibly, for the first step), provided that the initial approximation satisfies $0 < ax_0 < 2$. If $x_0 < 0$, then $x_n \to -\infty$. The same is true if $ax_0 > 2$.

(c) We have

$$x_{n+1} - \tfrac{1}{a} = x_n(2-ax_n) - \tfrac{1}{a} = \tfrac{1}{a}(2ax_n - a^2x_n^2 - 1) = -\tfrac{1}{a}(ax_n - 1)^2,$$

so that

$$x_{n+1} - \frac{1}{a} = -a\left(x_n - \frac{1}{a}\right)^2.$$

Therefore, if $a < 1$ (in particular, if $\frac{1}{2} \le a < 1$), it follows that $|x_{n+1} - \frac{1}{a}| < |x_n - \frac{1}{a}|^2$.

(d) By (c), one gets

$$\left|x_n - \frac{1}{a}\right| < \left|x_{n-1} - \frac{1}{a}\right|^2 < \left|x_{n-2} - \frac{1}{a}\right|^{2^2} < \cdots < \left|x_0 - \frac{1}{a}\right|^{2^n}.$$

Since $1 < \frac{1}{a} \le 2$ and $x_0 = \frac{3}{2}$ (which satisfies the condition in (b)), we have $|x_0 - \frac{1}{a}| \le \frac{1}{2}$, so that $|x_n - \frac{1}{a}| < 2^{-2^n}$. Thus, the error is less than 2^{-48} if $2^{-2^n} < 2^{-48}$, that is, $n > \log_2 48 = 5.5849\ldots$, so $n \ge 6$. Therefore, no more than six iterations are required.

23. (a) In the first case, the function $f(x) = x^2 - A$ is convex on $\mathbb{R}_+$ and increases monotonically from $-A$ to ∞. Therefore, if $x_0 > \alpha$, then x_n converges monotonically decreasing to α. If $0 < x_0 < \alpha$, then $x_1 > \alpha$, and the same conclusion holds for $n \ge 1$.

In the second case, $f(x) = \frac{A}{x^2} - 1$ is again convex on $\mathbb{R}_+$, but decreases from ∞ to -1. If $0 < x_0 < \alpha$, then x_n converges monotonically increasing to α. If $x_0 > \alpha$, we must make sure that $x_1 > 0$, which means that

$$x_1 = x_0 - \frac{\frac{A}{x_0^2} - 1}{-2\,\frac{A}{x_0^3}} > 0, \quad x_0 + x_0\,\frac{A - x_0^2}{2A} > 0,$$

$$x_0(3A - x_0^2) > 0, \qquad x_0 < \sqrt{3A} =: b.$$

(b) For the first equation, the reasoning is the same as in (a), and similar for the second equation, the condition $x_1 > 0$ now becoming

$$x_0(4A - x_0^3) > 0, \qquad x_0 < \sqrt[3]{4A} =: b.$$

24. (a) We have

$$\begin{aligned} x_{n+1} - \alpha &= \frac{1}{2}\left(x_n - \alpha + \frac{\alpha^2}{x_n} - \alpha\right) \\ &= \frac{1}{2}\,\frac{x_n^2 - 2\alpha x_n + \alpha^2}{x_n} \\ &= \frac{1}{2}\,\frac{(x_n - \alpha)^2}{x_n}\,. \end{aligned}$$

Since $x_n \to \alpha$ as $n \to \infty$, we find directly

$$\lim_{n\to\infty}\ \frac{x_{n+1} - \alpha}{(x_n - \alpha)^2} = \frac{1}{2\alpha}\,,$$

in agreement with the general formula $c = \frac{1}{2}\,\frac{f''(\alpha)}{f'(\alpha)}$, $f(x) = x^2 - \alpha^2$, for the error constant.

(b) The iteration for the cube root $\alpha = a^{\frac{1}{3}}$ is

$$x_{n+1} = \frac{2}{3}\,x_n + \frac{\alpha^3}{3x_n^2}\,.$$

Therefore,

$$x_{n+1} - \alpha = \frac{2x_n^3 + \alpha^3 - 3\alpha x_n^2}{3x_n^2}\,.$$

By repeatedly trying to find factors $x_n - \alpha$ in the numerator on the right, or else, by synthetic division, one finds after some algebra

$$\frac{x_{n+1} - \alpha}{(x_n - \alpha)^2} = \frac{2x_n + \alpha}{3x_n^2} \to \frac{1}{\alpha} \quad \text{as} \quad n \to \infty.$$

25. (a) We have

$$d_n = x_{n+1} - x_n = \frac{1}{2}\left(\frac{a}{x_n} - x_n\right).$$

Thus,

$$x_n^2 + 2d_n x_n - a = 0,$$

and therefore, solving this quadratic equation for x_n and recalling $x_n > 0$,

$$x_n = -d_n + \sqrt{d_n^2 + a} = \frac{a}{d_n + \sqrt{d_n^2 + a}}.$$

(b) Substituting $x_n = \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}}\right)$ in the formula $d_n = \frac{1}{2}\left(\frac{a}{x_n} - x_n\right)$ of (a), one gets

$$d_n = \frac{a}{x_{n-1} + \frac{a}{x_{n-1}}} - \frac{1}{4}\left(x_{n-1} + \frac{a}{x_{n-1}}\right).$$

The result of (a), with n replaced by $n-1$, then yields

$$\begin{aligned}
x_{n-1} + \frac{a}{x_{n-1}} &= \frac{a}{d_{n-1} + \sqrt{d_{n-1}^2 + a}} + d_{n-1} + \sqrt{d_{n-1}^2 + a} \\
&= \frac{a + d_{n-1}\left(d_{n-1} + \sqrt{d_{n-1}^2 + a}\right) + \sqrt{d_{n-1}^2 + a}\left(d_{n-1} + \sqrt{d_{n-1}^2 + a}\right)}{d_{n-1} + \sqrt{d_{n-1}^2 + a}} \\
&= \frac{a + d_{n-1}^2 + 2d_{n-1}\sqrt{d_{n-1}^2 + a} + d_{n-1}^2 + a}{d_{n-1} + \sqrt{d_{n-1}^2 + a}} \\
&= 2\,\frac{\sqrt{d_{n-1}^2 + a}\left(\sqrt{d_{n-1}^2 + a} + d_{n-1}\right)}{\sqrt{d_{n-1}^2 + a} + d_{n-1}} \\
&= 2\,\sqrt{d_{n-1}^2 + a},
\end{aligned}$$

hence

$$d_n = \frac{a}{2\sqrt{d_{n-1}^2 + a}} - \frac{1}{2}\sqrt{d_{n-1}^2 + a} = -\frac{d_{n-1}^2}{2\sqrt{d_{n-1}^2 + a}},$$

from which the assertion follows.

Since $x_{n+1} = x_n + d_n$ we may think of the d_n as successive "correction terms". They clearly converge to zero, and the result obtained above

shows quadratic convergence (with asymptotic error constant $-1/2\alpha$). But if d_0 is very large (for example if $x_0 > 0$ is close to zero), then our result shows that initially $|d_n| \approx \frac{1}{2}\,|d_{n-1}|$, and the convergence behavior resembles that of linear convergence with error constant $1/2$.

26. (a) Since

$$x - \frac{f(x)}{f'(x)} = x - \frac{x^3 - a}{3x^2} = \frac{2}{3}\,x + \frac{a}{3x^2}\,,$$

Newton's iteration is

$$x_{n+1} = \frac{2x_n}{3} + \frac{a}{3x_n^2}\,.$$

(b) According to (4.63), we need to determine λ such that

$$f'_\lambda(\alpha) \neq 0, \quad f''_\lambda(\alpha) = 0.$$

We have

$$f'_\lambda(x) = (3-\lambda)x^{2-\lambda} + \lambda a x^{-\lambda-1},$$
$$f''_\lambda(x) = (3-\lambda)(2-\lambda)x^{1-\lambda} - \lambda(\lambda+1)ax^{-\lambda-2}.$$

Thus,

$$f'_\lambda(\alpha) = \alpha^{-\lambda-1}[(3-\lambda)\alpha^3 + \lambda\alpha^3] = 3\alpha^{2-\lambda} \neq 0,$$
$$f''_\lambda(\alpha) = (3-\lambda)(2-\lambda)\alpha^{1-\lambda} - \lambda(\lambda+1)\alpha^{1-\lambda} = 6(1-\lambda)\alpha^{1-\lambda},$$

and the condition on λ amounts to $\lambda = 1$. Thus, $f_1(x) = x^2 - \frac{a}{x}$, and Newton's iteration becomes

$$x_{n+1} = x_n - \frac{f_1(x_n)}{f_1'(x_n)} = x_n - \frac{x_n^2 - ax_n^{-1}}{2x_n + ax_n^{-2}} = x_n - x_n\,\frac{x_n^3 - a}{2x_n^3 + a}\,,$$

that is,

$$x_{n+1} = x_n\,\frac{x_n^3 + 2a}{2x_n^3 + a}\,, \quad n = 0, 1, 2, \ldots\ .$$

27. (a) The graphs of $y = \ln x$ and $y = \frac{1}{x}$ clearly intersect at exactly one point, whose abscissa is larger than 1 (obviously) and less than 2 (since $\ln 2 > \frac{1}{2}$).

(b) Let first $f(x) = x\ln x - 1$. Then $f'(x) = \ln x + 1$, $f''(x) = \frac{1}{x}$, so that f is convex on $\mathbb{R}_+$. For any x_0 in the interval $(0, \mathrm{e}^{-1})$, where f is monotonically decreasing, Newton's method produces a negative x_1, which is unacceptable. On the other hand, Newton's method, by convexity of f, converges monotonically decreasing (except, possibly, for the first step) for any x_0 in $(\mathrm{e}^{-1}, \infty)$.

Let now $g(x) = \ln x - \frac{1}{x}$. Here, $g'(x) = x^{-2}(x+1)$, $g''(x) = -x^{-3}(x+2)$, so that g increases monotonically from $-\infty$ to $+\infty$ and is concave on $\mathbb{R}_+$. For any $x_0 < \alpha$, therefore, Newton's method converges monotonically increasing. If $x_0 > \alpha$, one must ensure that $x_1 > 0$. Since

$$x_1 = x_0 - \frac{\ln x_0 - x_0^{-1}}{x_0^{-2}(x_0+1)} = x_0 \, \frac{x_0 + 2 - x_0 \ln x_0}{x_0 + 1} \,,$$

we thus must have $x_0 + 2 - x_0 \ln x_0 > 0$, i.e., $x_0 < x_*$ where

$$x_* \ln x_* - x_* - 2 = 0.$$

This has a unique solution between 4 and 5, which can be obtained in turn by Newton's method. The result is $x_* = 4.319136566\ldots$.

(c) The respective asymptotic error constants are

$$c_f = \left. \frac{f''(x)}{2f'(x)} \right|_{x=\alpha} = \left. \frac{1}{2x(\ln x + 1)} \right|_{x=\alpha} = \frac{1}{2(\alpha+1)}$$

and

$$c_g = \left. \frac{g''(x)}{2g'(x)} \right|_{x=\alpha} = -\frac{\alpha+2}{2\alpha(\alpha+1)} \,.$$

We have

$$\frac{c_f}{|c_g|} = \frac{1}{1 + \frac{2}{\alpha}} \,.$$

Since $1 + \frac{2}{\alpha} > 2$, we have $c_f < \frac{1}{2} |c_g|$, so that Newton's method for (A) converges asymptotically faster by more than a factor of 2.

28. By Taylor's theorem applied to f and f',

$$\begin{aligned} f(x) &= f(\alpha) + (x-\alpha) f'(\alpha) + \frac{1}{2} (x-\alpha)^2 f''(\xi) \\ &= (x-\alpha) f'(\alpha) \left(1 + (x-\alpha) \, \frac{f''(\xi)}{2f'(\alpha)} \right), \\ f'(x) &= f'(\alpha) + (x-\alpha) f''(\xi_1) \\ &= f'(\alpha) \left(1 + (x-\alpha) \, \frac{f''(\xi_1)}{f'(\alpha)} \right), \end{aligned}$$

where ξ and ξ_1 lie between α and x. If $x \in I_\varepsilon$ and $2\varepsilon M(\varepsilon) < 1$, then

$$\left| (x-\alpha) \frac{f''(\xi)}{2f'(\alpha)} \right| \le \varepsilon M(\varepsilon) < \frac{1}{2} \,,$$

so that $f(x) \neq 0$ for $x \neq \alpha$. Likewise,

$$\left| (x-\alpha) \frac{f''(\xi_1)}{f'(\alpha)} \right| \le 2\varepsilon M(\varepsilon) < 1,$$

so that also $f'(x) \neq 0$. Furthermore, $x_n \in I_\varepsilon$ implies by (4.62) that $|x_{n+1} - \alpha| \leq \varepsilon \cdot \varepsilon M(\varepsilon) < \frac{1}{2}\varepsilon$, in particular, $x_{n+1} \in I_\varepsilon$. Since $x_0 \in I_\varepsilon$ it follows that all $x_n \in I_\varepsilon$, and since $f'(x_n) \neq 0$, Newton's method is well defined. Finally, again by (4.62),

$$|x_{n+1} - \alpha| \leq |x_n - \alpha|\, \varepsilon M(\varepsilon), \qquad n = 0, 1, 2, \ldots,$$

giving

$$|x_n - \alpha| \leq [\varepsilon M(\varepsilon)]^n |x_0 - \alpha| \to 0 \quad \text{as} \quad n \to \infty,$$

since $\varepsilon M(\varepsilon) < \frac{1}{2}$.

29. (a) This is readily seen by plotting the graphs of $y = \tan x$ and $y = cx$ for $x > 0$, and observing that they intersect for the first time in the interval $(\pi, \frac{3\pi}{2})$.

(b) Note that

$$f'(x) = 1 + \tan^2 x - c > 0,$$
$$f''(x) = 2 \tan x (1 + \tan^2 x),$$

so that f on the interval $(\pi, \frac{3\pi}{2})$ is convex and monotonically increasing from $-c\pi$ to ∞. Therefore, Newton's method converges with $x_0 = \pi$ if $x_1 < \frac{3\pi}{2}$, which translates to

$$\pi + \frac{c\pi}{1 - c} < \frac{3\pi}{2},$$

that is, $c < \frac{1}{3}$.

30. We have

$$f(x) = \cos x \cosh x - 1,$$
$$f'(x) = -\sin x \cosh x + \cos x \sinh x,$$
$$f''(x) = -2 \sin x \sinh x.$$

Clearly, $f''(x) > 0$ on $[-\frac{\pi}{2} + 2n\pi, 2n\pi]$ and $f''(x) < 0$ on $[2n\pi, \frac{\pi}{2} + 2n\pi]$. Furthermore, $f(-\frac{\pi}{2} + 2n\pi) = f(\frac{\pi}{2} + 2n\pi) = -1$ and $f(2n\pi) = \cosh(2n\pi) > 1$. Since f is convex on the first half of the interval $[-\frac{\pi}{2} + 2n\pi, \frac{\pi}{2} + 2n\pi]$, Newton's method started at the left endpoint converges monotonically decreasing (except for the first step) to α_n, provided the first iterate is to the left of the midpoint. This is the case since, with $x_0 = -\frac{\pi}{2} + 2n\pi$, we have, for $n \geq 1$,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = -\frac{\pi}{2} + 2n\pi + \frac{1}{\cosh(-\frac{\pi}{2} + 2n\pi)}$$
$$< -\frac{\pi}{2} + 2n\pi + \frac{1}{\cosh(\frac{3\pi}{2})} = 2n\pi - 1.55283\ldots < 2n\pi.$$

Since f is concave on the second half of the interval, Newton's method started at the right endpoint converges monotonically decreasing to β_n.

31. (a) The graphs of $y = \tan x$ and $y = -\tanh x$ on $\mathbb{R}_+$ intersect infinitely often, exactly once in each interval $[(n-\frac{1}{2})\pi, n\pi]$, $n = 1, 2, 3, \ldots$. The respective abscissae α_n are the positive roots of the equation.

    (b) Since $\tanh x \to 1$ as $x \to \infty$, the geometric discussion in (a) shows that $\alpha_n - n\pi \sim \tan^{-1}(-1) = -\tan^{-1}(1)$, hence $n\pi - \alpha_n \sim \tan^{-1}(1) = \pi/4 = .785398\ldots$ as $n \to \infty$.

    (c) On the interval $I_n = [(n-\frac{1}{2})\pi, n\pi]$ we have $f((n-\frac{1}{2})\pi) = -\infty$, $f(n\pi) = \tanh n\pi > 0$ and

$$f'(x) = 2 + \tan^2 x - \tanh^2 x > 0,$$
$$f''(x) = 2\tan x(1 + \tan^2 x) - 2\tanh x(1 - \tanh^2 x) < 0.$$

    Thus, f is monotonically increasing and concave on I_n. Newton's method will converge if started at the right endpoint, $x_0 = n\pi$, provided $x_1 > (n - \frac{1}{2})\pi$. This is indeed the case: since the function $u/(2-u^2)$ on $[0,1]$ increases from 0 to 1, we have

$$x_1 = n\pi - \frac{\tanh n\pi}{2 - \tanh^2 n\pi} > n\pi - 1 > n\pi - \tfrac{1}{2}\pi.$$

32. See the text.

33. (a) Plotting the graphs of $y = \sin x$ and $y = \frac{1}{x}$, one sees immediately that they intersect in two points with abscissae α_1 and α_2 such that $0 < \alpha_1 < \frac{1}{2}\pi < \alpha_2 < \pi$.

    (b) Since $f(1) = \sin 1 - 1 < 0$ and $f(\frac{1}{2}\pi) > 0$, we have $1 < \alpha_1 < \frac{1}{2}\pi$. From

$$f'(x) = \sin x + x\cos x, \quad f''(x) = 2\cos x - x\sin x,$$

    one sees that $f''(0) > 0$ and f'' is monotonically decreasing on $[0, \frac{1}{2}\pi]$. Since $f''(\frac{1}{2}\pi) < 0$, there is a unique $\omega \in [0, \frac{1}{2}\pi]$ such that $f''(\omega) = 0$. In fact,

$$\omega \tan \omega = 2,$$

    which has the solution (obtained by Newton's method) $\omega = 1.07687\ldots$. Since $f(\omega) = -.05183\ldots < 0$, $f(\frac{1}{2}\pi) > 0$, and $f'' < 0$ on $[\omega, \frac{1}{2}\pi]$, Newton's method started with $x_0 = \omega$ converges monotonically increasing to α_1. If started with $x_0 = \pi/2$, the same is (numerically) observed to happen after the third Newton step.

    On the other hand, since $f(\frac{1}{2}\pi) > 0$, $f(\pi) < 0$, and $f'' < 0$ on $[\frac{1}{2}\pi, \pi]$, starting Newton's method with $x_0 = \pi$ yields monotone (decreasing) convergence to α_2.

34. See the text.

35. For x near α, we have by Taylor's theorem (assuming $f \in C^m$ near α), since $f(\alpha) = f'(\alpha) = \cdots = f^{(m-1)}(\alpha) = 0$, that

$$\begin{aligned} f(x) &= f(\alpha) + f'(\alpha)(x-\alpha) + \cdots + \frac{f^{(m-1)}(\alpha)}{(m-1)!}(x-\alpha)^{m-1} + \frac{f^{(m)}(\xi_n)}{m!}(x-\alpha)^m \\ &= \frac{f^{(m)}(\xi_n)}{m!}(x-\alpha)^m, \\ f'(x) &= f'(\alpha) + \cdots + \frac{f^{(m-1)}(\alpha)}{(m-2)!}(x-\alpha)^{m-2} + \frac{f^{(m)}(\xi_n')}{(m-1)!}(x-\alpha)^{m-1} \\ &= \frac{f^{(m)}(\xi_n')}{(m-1)!}(x-\alpha)^{m-1}, \end{aligned}$$

where ξ_n, ξ_n' are between x and α. Therefore, we get

$$\begin{aligned} x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} &= x_n - \frac{\dfrac{f^{(m)}(\xi_n)}{m!}(x_n-\alpha)^m}{\dfrac{f^{(m)}(\xi_n')}{(m-1)!}(x_n-\alpha)^{m-1}} \\ &= x_n - \frac{1}{m}\,\frac{f^{(m)}(\xi_n)}{f^{(m)}(\xi_n')}(x_n-\alpha). \end{aligned}$$

Subtracting α on both sides and dividing by $x_n - \alpha$ gives

$$\frac{x_{n+1}-\alpha}{x_n-\alpha} = 1 - \frac{1}{m}\,\frac{f^{(m)}(\xi_n)}{f^{(m)}(\xi_n')} \to 1 - \frac{1}{m} \quad \text{as } n \to \infty.$$

Thus, convergence is linear, and the asymptotic error constant is $c = 1 - \frac{1}{m}$.

36. We answer (a) by doing (b). The "m-times relaxed" Newton's method is

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

and thus has the form $x_{n+1} = \varphi(x_n)$ with

$$\varphi(x) = x - m\,\frac{f(x)}{f'(x)}.$$

Let α be a zero of f of exact multiplicity m. Then

$$f(x) = (x-\alpha)^m g(x), \qquad g(\alpha) \neq 0,$$

where, by Taylor expansion,

$$g(x) = \frac{1}{m!}f^{(m)}(\alpha) + \frac{x-\alpha}{(m+1)!}f^{(m+1)}(\alpha) + \cdots.$$

Thus,

$$\begin{aligned}\varphi(x) &= x - \frac{m(x-\alpha)^m g(x)}{m(x-\alpha)^{m-1}g(x) + (x-\alpha)^m g'(x)} \\ &= x - (x-\alpha)\frac{g(x)}{g(x) + \dfrac{1}{m}(x-\alpha)g'(x)}\,.\end{aligned}$$

Differentiating, we get

$$\varphi'(x) = 1 - \frac{g(x)}{g(x) + \dfrac{1}{m}(x-\alpha)g'(x)} - (x-\alpha)\left[\frac{g}{g + \dfrac{1}{m}(x-\alpha)g'}\right]',$$

and thus

$$\varphi'(\alpha) = 0.$$

The method converges quadratically, at least. We can write

$$\varphi'(x) = (x-\alpha)\left\{\frac{\dfrac{1}{m}\,g'(x)}{g(x) + \dfrac{1}{m}(x-\alpha)g'(x)} - \left[\frac{g}{g + \dfrac{1}{m}(x-\alpha)g'}\right]'\right\}.$$

Differentiating once more, and setting $x = \alpha$, gives

$$\begin{aligned}\varphi''(\alpha) &= \frac{1}{m}\,\frac{g'(\alpha)}{g(\alpha)} - \left.\frac{(g+\cdots\,)g' - g(g' + \dfrac{1}{m}\,g' + \cdots\,)}{(g + \dfrac{1}{m}(x-\alpha)g')^2}\right|_{\alpha} \\ &= \frac{1}{m}\,\frac{g'(\alpha)}{g(\alpha)} + \frac{1}{m}\,\frac{g'(\alpha)}{g(\alpha)} = \frac{2}{m}\,\frac{g'(\alpha)}{g(\alpha)}\,.\end{aligned}$$

We have

$$g(\alpha) = \frac{1}{m!}\,f^{(m)}(\alpha), \quad g'(\alpha) = \frac{1}{(m+1)!}\,f^{(m+1)}(\alpha),$$

so that

$$\varphi''(\alpha) = \frac{2}{m}\,\frac{f^{(m+1)}(\alpha)}{(m+1)!\dfrac{1}{m!}\,f^{(m)}(\alpha)} = \frac{2}{m(m+1)}\,\frac{f^{(m+1)}(\alpha)}{f^{(m)}(\alpha)}\,.$$

Convergence, therefore, is exactly of order 2 if $f^{(m+1)}(\alpha) \neq 0$. Under this condition, the asymptotic error constant, by (4.72), is

$$c = \frac{1}{2}\,\varphi''(\alpha) = \frac{1}{m(m+1)}\,\frac{f^{(m+1)}(\alpha)}{f^{(m)}(\alpha)}\,.$$

37. (a) Letting $f(x) = x\ln x - a$, we have $f(0) = -a < 0$, $f'(x) = \ln x + 1$, so that $f'(x) < 0$ for $0 < x < \frac{1}{\mathrm{e}}$, and $f'(x) > 0$ for $x > \frac{1}{\mathrm{e}}$. Since $f(x) \to +\infty$ as $x \to \infty$, there is exactly one positive root, $x = x(a)$.

(b) It is clear, first of all, that $x(a) \to \infty$ as $a \to \infty$. Differentiating the identity

$$x(a) \ln x(a) \equiv a$$

with respect to a gives

$$\frac{\mathrm{d}x}{\mathrm{d}a} \ln x(a) + x(a) \cdot \frac{1}{x(a)} \frac{\mathrm{d}x}{\mathrm{d}a} \equiv 1,$$

that is,

$$\frac{\mathrm{d}x}{\mathrm{d}a} = \frac{1}{1 + \ln x(a)}\,.$$

Since

$$\frac{x(a) \ln a}{a} = \frac{\ln a}{\ln x(a)}\,,$$

the rule of Bernoulli-L'Hospital yields

$$\lim_{a\to\infty} \frac{x(a) \ln a}{a} = \lim_{a\to\infty} \frac{\dfrac{1}{a}}{\dfrac{1}{x(a)} \cdot \dfrac{1}{1 + \ln x(a)}} = \lim_{a\to\infty} \frac{(1 + \ln x(a))x(a)}{a}$$

$$= \lim_{a\to\infty} \frac{x(a) + a}{a} = \lim_{a\to\infty} \left(\frac{1}{\ln x(a)} + 1 \right) = 1.$$

(c) Applying one step of Newton's method with initial approximation x_0 to the equation $f(x) = 0,\ f(x) = x \ln x - a$, we get

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - \frac{x_0 \ln x_0 - a}{\ln x_0 + 1}$$

$$= \frac{x_0(\ln x_0 + 1) - x_0 \ln x_0 + a}{\ln x_0 + 1}$$

$$= \frac{x_0 + a}{\ln x_0 + 1}\,.$$

Inserting $x_0 = \frac{a}{\ln a}$ gives

$$x_1 = \frac{\dfrac{a}{\ln a} + a}{\ln a - \ln\ln a + 1}$$

$$= \frac{a}{\ln a} \, \frac{1 + \ln a}{1 + \ln a - \ln\ln a}$$

$$= \frac{a}{\ln a} \, \frac{1}{1 - \dfrac{\ln\ln a}{1 + \ln a}}$$

as the improved approximation.

38. The roots are $\alpha_1 = \sqrt{2}$, $\alpha_2 = -\sqrt{2}$. Let the iteration function be $\varphi(x)$ in each case.

(a) Here,

$$\varphi(x) = \frac{2}{x}\,, \quad \varphi'(\alpha_{1,2}) = -\left.\frac{2}{x^2}\right|_{x=\pm\sqrt{2}} = -1.$$

Thus, one cannot expect linear convergence. In fact, for any $x_0 \neq 0$, we have

$$x_1 = \frac{2}{x_0}\,,$$
$$x_2 = \frac{2}{x_1} = \frac{2}{2/x_0} = x_0,$$

and the iteration cycles. Thus, it cannot converge unless $x_0 = \alpha$.

(b) Here we have

$$\varphi(x) = x^2 + x - 2, \quad \varphi'(x) = 2x + 1,$$
$$\varphi'(\alpha_1) = 2\sqrt{2} + 1 > 1, \quad \varphi'(\alpha_2) = 1 - 2\sqrt{2} < -1,$$

so that $|\varphi'(\alpha)| > 1$. The iteration does not converge, unless $x_0 = \alpha$.

(c) Here,

$$\varphi(x) = \frac{x+2}{x+1}\,, \quad \varphi'(x) = -\frac{1}{(x+1)^2}\,,$$
$$\varphi'(\alpha_1) = -\frac{1}{(\sqrt{2}+1)^2}\,, \quad \varphi'(\alpha_2) = -\frac{1}{(1-\sqrt{2})^2}\,.$$

Thus, $|\varphi'(\alpha_1)| < 1$, and the iteration converges locally to α_1, but $|\varphi'(\alpha_2)| > 1$, and it cannot converge to α_2 unless $x_0 = \alpha_2$.

39. This is a fixed point iteration with iteration function

$$\varphi(x) = \frac{x(x^2+3a)}{3x^2+a}.$$

Clearly, $\varphi(\alpha) = \alpha$. Differentiating repeatedly the identity

$$(3x^2 + a)\varphi(x) = x^3 + 3ax,$$

one gets first

$$6x\varphi(x) + (3x^2 + a)\varphi'(x) = 3x^2 + 3a,$$

hence $4a\varphi'(\alpha) = 6a - 6\alpha\varphi(\alpha) = 6a - 6\alpha^2 = 0$, then

$$6\varphi(x) + 12x\varphi'(x) + (3x^2 + a)\varphi''(x) = 6x,$$

hence $4a\varphi''(\alpha) = 6\alpha - 6\alpha = 0$, and finally

$$18\varphi'(x) + 18x\varphi''(x) + (3x^2 + a)\varphi'''(x) = 6,$$

hence $4a\varphi'''(\alpha) = 6$, that is, $\varphi'''(\alpha) \neq 0$. This shows that the iteration converges with order $p = 3$.

The asymptotic error constant, by (4.72), is

$$c = \frac{1}{3!}\,\varphi'''(\alpha) = \frac{1}{6} \cdot \frac{6}{4a} = \frac{1}{4a}\,.$$

40. See the text.

41. (a) The algorithm is the fixed point iteration $x_{n+1} = \varphi(x_n)$, where $\varphi(x) = \frac{1}{\omega}(x^2 - (3-\omega)x + 2)$. Local convergence to 1 requires that $|\varphi'(1)| < 1$. Since

$$\varphi'(x) = \frac{1}{\omega}(2x - (3-\omega)),$$

this gives

$$|\varphi'(1)| = \left|\,\frac{\omega - 1}{\omega}\,\right| < 1,$$

that is,

$$\tfrac{1}{2} < \omega < \infty.$$

(b) Similarly,

$$|\varphi'(2)| = \left|\,\frac{\omega + 1}{\omega}\,\right| < 1$$

is equivalent to

$$-\infty < \omega < -\tfrac{1}{2}.$$

(c) We have quadratic convergence to 1 precisely if $\varphi'(1) = 0$, that is, $\omega = 1$.

(d) The algorithm can be written in the form

$$x_{n+1} = x_n^2 - 2x_n + 2 = x_n - (3x_n - x_n^2 - 2) = x_n - \frac{F(x_n)}{F'(x_n)}$$

provided

$$\frac{F(x)}{F'(x)} = 3x - x^2 - 2 = -(x-1)(x-2),$$

or

$$\frac{F'(x)}{F(x)} = (\log F)' = -\frac{1}{(x-1)(x-2)} = \frac{1}{x-1} - \frac{1}{x-2}\,.$$

Integration followed by exponentiation gives

$$F(x) = \text{const}\;\frac{x-1}{x-2}\,.$$

From the graph of the function F (we may set the constant equal to 1), it follows that Newton's method converges to 1 precisely if $0 < x_0 < 2$, since $x_0 = 0$ yields $x_1 = 2$, and F is concave and monotonically decreasing to $-\infty$ on $[0, 2)$. Any $x_0 > 2$, and therefore also any $x_0 < 0$, produces an iteration diverging to ∞.

42. Apply Theorem 4.7.1 to Newton's method considered as the fixed point iteration

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

We have $\varphi(\alpha) = \alpha$, $\varphi'(\alpha) = 0$ (cf. the remarks after Theorem 4.7.1) and, since

$$\varphi'(x) = f(x)\frac{f''(x)}{[f'(x)]^2}\,,$$

using Leibniz's rule of differentiation,

$$\varphi^{(k+1)}(x) = \sum_{j=0}^{k} \binom{k}{j} f^{(j)}(x) \left(\frac{f''(x)}{[f'(x)]^2}\right)^{(k-j)}, \quad k = 1, 2, \ldots, p-1.$$

If $x = \alpha$, all terms vanish, since by assumption, $f^{(j)}(\alpha) = 0$, except possibly for the term with $j = 1$. But this one also vanishes if $k \leq p-2$ on account of the other derivative $(\cdots)^{(k-1)}$, which involves derivatives in the numerator of orders $p-1$ at most. If $k = p-1$, for the same reasons,

$$\begin{aligned}
\varphi^{(p)}(\alpha) &= (p-1)f'(\alpha)\left(\frac{f''(\alpha)}{[f'(\alpha)]^2}\right)^{(p-2)} \\
&= (p-1)f'(\alpha)\left\{\frac{f^{(p)}(\alpha)}{[f'(\alpha)]^2} + \binom{p-2}{1} f^{(p-1)}(\alpha)\left(\frac{1}{[f'(\alpha)]^2}\right)' + \cdots\right\} \\
&= (p-1)\frac{f^{(p)}(\alpha)}{f'(\alpha)} \neq 0.
\end{aligned}$$

Hence the order of convergence is p, and the asymptotic error constant is (cf. (4.72))

$$c = \frac{1}{p(p-2)!}\frac{f^{(p)}(\alpha)}{f'(\alpha)}\,.$$

43. See the text.

44. (a) Comparing coefficients of equal powers on the left and right of the identity, one gets

$$\begin{aligned}
a_{d-1} &= b_{d-1} - t, \\
a_{d-2} &= b_{d-2} - tb_{d-1} - s, \\
a_{d-3} &= b_{d-3} - tb_{d-2} - sb_{d-1}, \\
&\cdots \quad \cdots\cdots\cdots\cdots\cdots \\
a_2 &= b_2 - tb_3 - sb_4, \\
a_1 &= b_1 - tb_2 - sb_3, \\
a_0 &= b_0 - tb_1 - sb_2.
\end{aligned}$$

If we define $b_{d+1} = 0$, $b_d = 1$, and solve for the leading b on the right, one obtains

$$b_{d+1} = 0, \quad b_d = 1,$$
$$b_k = a_k + tb_{k+1} + sb_{k+2}, \quad k = d-1, d-2, \ldots, 0.$$

(b) Let $\alpha = \rho + \mathrm{i}\sigma$, with ρ the real part and σ the imaginary part of α. Then $\overline{\alpha} = \rho - \mathrm{i}\sigma$ is also a zero of f, so that f has the quadratic factor

$$(x-\alpha)(x-\overline{\alpha}) = x^2 - 2\rho x + \rho^2 + \sigma^2.$$

Letting therefore

$$t = 2\rho, \quad s = -(\rho^2 + \sigma^2)$$

in the algorithm of (a) will produce the coefficients b_i of the deflated polynomial

$$\frac{f(x)}{(x-\alpha)(x-\overline{\alpha})} = x^{d-2} + b_{d-1}x^{d-3} + \cdots + b_2.$$

(c) The polynomial $x^2 - tx - s$ is a quadratic factor of f if and only if

$$b_0(t,s) = 0,$$
$$b_1(t,s) = 0,$$

where b_0, b_1 are the terminal values obtained in the algorithm of (a). This is a system of two equations in two unknowns and can be solved by Newton's method:

$$\begin{bmatrix} \dfrac{\partial b_0}{\partial t} & \dfrac{\partial b_0}{\partial s} \\ \dfrac{\partial b_1}{\partial t} & \dfrac{\partial b_1}{\partial s} \end{bmatrix} (t_n, s_n) \cdot \begin{bmatrix} \delta_n \\ \varepsilon_n \end{bmatrix} = - \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} (t_n, s_n),$$

$$\begin{bmatrix} t_{n+1} \\ s_{n+1} \end{bmatrix} = \begin{bmatrix} t_n \\ s_n \end{bmatrix} + \begin{bmatrix} \delta_n \\ \varepsilon_n \end{bmatrix}, \quad n = 0, 1, \ldots ,$$

with t_0, s_0 suitable initial approximations to the coefficients of the quadratic factor.

To compute the Jacobian matrix, define

$$u_k = \frac{\partial b_{k-1}}{\partial t}, \quad v_k = \frac{\partial b_{k-2}}{\partial s}.$$

Differentiating the recurrence relations found in (a) partially with respect to t and s, one gets

$$u_{d+1} = 0, \quad u_d = 1,$$
$$u_k = b_k + tu_{k+1} + su_{k+2}, \quad k = d-1, d-2, \ldots, 1,$$

and

$$v_{d+1} = 0, \quad v_d = 1,$$
$$v_k = b_k + tv_{k+1} + sv_{k+2}, \quad k = d-1, d-2, \ldots, 2.$$

Since the second recursion (including the initial values) is exactly the same as the first except for notation and the terminal value 1 of k, we have

$$v_k = u_k \quad \text{for} \quad k = 2, 3, \ldots, d+1.$$

The Jacobian matrix, therefore, is

$$\begin{bmatrix} u_1 & v_2 \\ u_2 & v_3 \end{bmatrix}(t_n, s_n) = \begin{bmatrix} u_1 & u_2 \\ u_2 & u_3 \end{bmatrix}(t_n, s_n).$$

Note the symmetry of the Jacobian.

45. See the text.

46. Differentiating φ_r, we obtain

$$\varphi_r'(x) = 1 + \sum_{m=1}^{r} (-1)^m \left\{ \frac{1}{m!} \left(y^{[m]}(x) \right)' [f(x)]^m + \frac{1}{m!} y^{[m]}(x) \cdot m[f(x)]^{m-1} f'(x) \right\}.$$

In view of $\left(y^{[m]}(x)\right)' = f'(x)y^{[m+1]}(x)$, this gives

$$\varphi_r'(x) = 1 + \sum_{m=1}^{r} (-1)^m \left\{ \frac{1}{m!} f'(x) y^{[m+1]}(x)[f(x)]^m + \frac{1}{(m-1)!} f'(x) y^{[m]}(x)[f(x)]^{m-1}(x) \right\}.$$

With $a_m := \frac{1}{m!} f'(x) y^{[m+1]}(x)[f(x)]^m$, this is

$$\varphi_r'(x) = 1 + \sum_{m=1}^{r} (-1)^m (a_m + a_{m-1}) = 1 + (-1)^r a_r - a_0$$
$$= 1 + (-1)^r a_r - f'(x)y^{[1]}(x) = (-1)^r a_r,$$

hence,

$$\varphi_r'(x) = \frac{(-1)^r}{r!} f'(x) y^{[r+1]}(x)[f(x)]^r.$$

From this, differentiating successively $r-1$ times, one concludes

$$\varphi_r^{(s)}(\alpha) = 0 \quad \text{for} \quad s = 1, 2, \ldots r,$$

while

$$\varphi_r^{(r+1)}(\alpha) = (-1)^r [f'(\alpha)]^{r+1} y^{[r+1]}(\alpha) \neq 0.$$

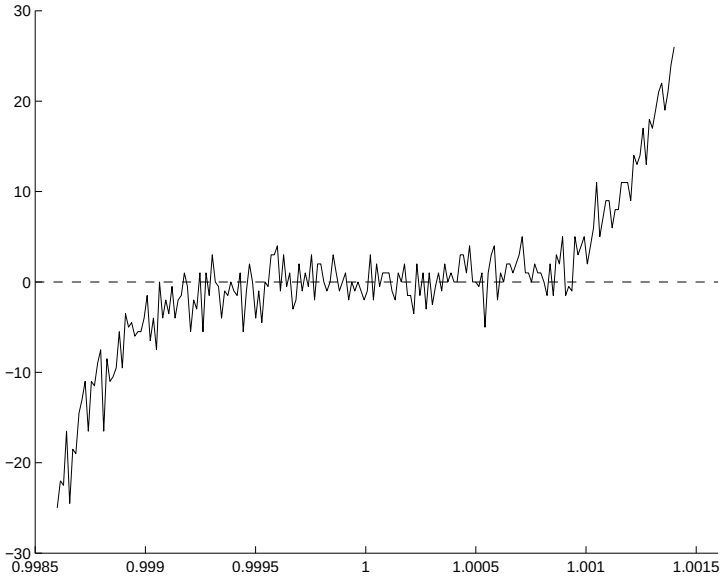
ANSWERS TO MACHINE ASSIGNMENTS

1. (a)

PROGRAM

```
%MAIV_1A
%
x=linspace(.9986,1.0014,200)';
p=ones(size(x));
p=x.*p-5; p=x.*p+10; p=x.*p-10;
p=x.*p+5; p=x.*p-1;
hold on
plot(x,p/eps)
axis([.9985 1.0016 -30 30])
plot([.9985 1.0016],[0 0],'--')
hold off
```

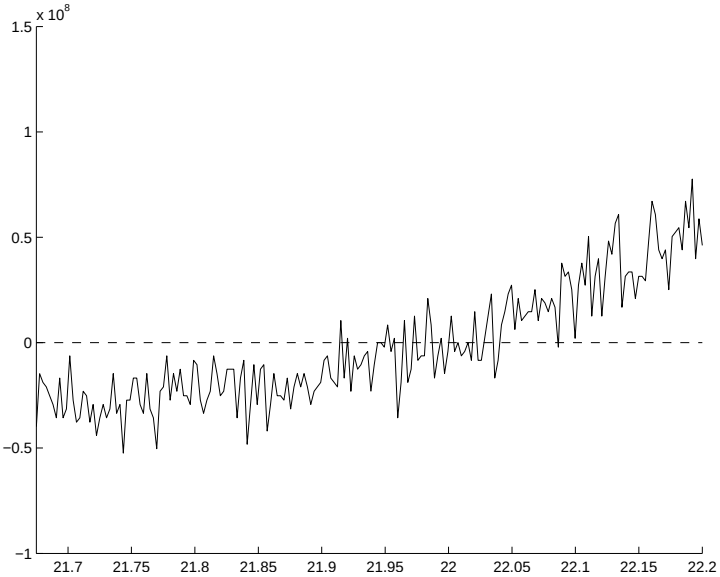
PLOT



Evaluation of $(x-1)^5$ in expanded form near $x = 1$ involves cancellation errors even in the sense of absolute errors; the errors are seen to be as large as 5 times the machine precision and they behave erratically. As a consequence, the computed (piecewise linear) function crosses the real axis many times near $x = 1$. The uncertainty interval is seen to be approximately $[.99925, 1.00075]$.

(b)

```
PROGRAM

%MAIV_1B
%
prec=eps('single');
x=single(linspace(21.675,22.2,200))';
p=ones(size(x),'single');
p=x.*p-100; p=x.*p+3995; p=x.*p-79700;
p=x.*p+794004; p=x.*p-3160080;
hold on
plot(x,p/prec);
axis([21.675 22.2 -1e8 1.5e8])
plot([21.6 22.2],[0 0],'--')
hold off

PLOT
```



Similar behavior as in (a), but with much larger errors and interval of uncertainty.

2. (a) The roots are symmetric with respect to the origin. It suffices to consider $x > 0$. If $x > \pi$, then

$$f(x) = \tfrac{1}{2}x - \sin x > \tfrac{\pi}{2} - 1 > 0,$$

and there are no real roots larger than π. On $[0, \pi]$, the function f is convex, zero at $x = 0$, negative at $x = \frac{\pi}{2}$, and positive at $x = \pi$. Hence, there is exactly one positive root between $\frac{\pi}{2}$ and π.

(b1)–(b4) Be sure, in the following Matlab routines, to place the correct function `f` resp. `fd` as subfunction in the routines `sbisec.m`, `sfalsepos.m`, etc.

```
 PROGRAMS

%MAIV_2B1
%
a=pi/2; b=pi; dig=35;
for tol=[.5e-7 .5e-15 .5e-33]
  [ntol,x]=sbisec(dig,a,b,tol)
end

%MAIV_2B2
%
a=pi/2; b=pi; nmax=100; dig=37;
for tol=[.5e-7 .5e-15 .5e-33]
  [n,x]=sfalsepos(dig,a,b,tol,nmax)
end

%MAIV_2B3
%
a=pi/2; b=pi; nmax=50; dig=35;
for tol=[.5e-7 .5e-15 .5e-33]
  [n,x]=ssecant(dig,a,b,tol,nmax)
end

%MAIV_2B4
%
a=pi; nmax=20; dig=35;
for tol=[.5e-7 .5e-15 .5e-33]
  [n,x]=snewton(dig,a,tol,nmax)
end
```

BISECTION

ntol	x
25	1.8954942
52	1.895494267033981
112	1.8954942670339809471440357380936 02

FALSE POSITION

n	x
16	1.8954942
33	1.895494267033981
73	1.8954942670339809471440357380936 01

SECANT

n	x
7	1.8954943
8	1.895494267033981
10	1.8954942670339809471440357380936 02

NEWTON

n	x
5	1.8954943
6	1.895494267033981
8	1.8954942670339809471440357380936 02

The methods perform as expected.

3. (a) A simple computation yields

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

(b) There are two sign changes in the coeficients of p_n. By Descartes' rule, p_n has either two positive zeros, or none. Since $p_n(0) > 0$, $p_n(1) < 0$, and $p_n(\infty) > 0$, it follows that there are exactly two positive zeros, one in $(0, 1)$, the other in $(1, \infty)$. The left-hand side of the original equation being invariant under the substitution $x \mapsto \frac{1}{x}$, each zero is the reciprocal of the other.

(c) We have

$$p_n'(x) = nx^{n-2}[2x^{n+1} - (n+1)x^2 - n + 1],$$

and

$$p_n''(x) = \begin{cases} 12x(x-1) & \text{if } n = 2, \\ nx^{n-3}\left[2(2n-1)x^{n+1} - n(n+1)x^2 - (n-1)(n-2)\right] & \\ \qquad\qquad \text{if } n > 2. & \end{cases}$$

Therefore,

$$p_n'(1) = 2n(1-n) < 0$$

and

$$p_n''(1) = \begin{cases} 0 \quad \text{if} \quad n = 2, \\ -2n(n-1)(n-2) < 0 \quad \text{if} \quad n > 2. \end{cases}$$

By Descartes' rule, p_n' and p_n'' both have at most one positive zero. Since $p_n'(1) < 0$ and clearly $p_n'(\alpha_n) > 0$, it follows that p_n' vanishes at some $x = \xi$, $1 < \xi < \alpha_n$, and remains positive thereafter. Since p_n has a minimum at $x = \xi$, we have $p_n''(\xi) > 0$. Since $p_n''(1) \leq 0$, it follows that $p_n''(x) > 0$ for $x > \xi$, in particular on $[\alpha_n, \alpha_{n-1}]$. Therefore, α_{n-1} is a valid initial approximation for α_n.

In the following Matlab routine be sure to place the correct function `f` resp. `fd` as subfunction in the routines `sbisec.m` resp. `snewton.m`.

PROGRAM

```
%MAIV_3C
%
global N
f0='%6.0f %12.8f %12.8f %6.0f %6.0f\n';
disp('     n   bisection      Newton     itbis  itnewt')
tol=.5e-8; dig=15; alpha=zeros(19,1); itmax=20;
for N=2:20
  if N==2, b=3; else b=alpha(N-2); end
  [itbis,xbis]=sbisec(dig,1,b,tol);
  [itnewt,xnewt]=snewton(dig,b,tol,itmax);
  alpha(N-1)=subs(xbis);
  fprintf(f0,N,subs(xbis),subs(xnewt),itbis,itnewt)
end
```

OUTPUT

```
>> MAIV_3C
     n   bisection      Newton     itbis  itnewt
     2   2.29663026   2.29663026     29      6
     3   1.93185165   1.93185165     28      6
     4   1.73908388   1.73908388     28      6
     5   1.61803399   1.61803399     28      6
     6   1.53417713   1.53417713     27      6
     7   1.47226911   1.47226911     27      6
     8   1.42447748   1.42447748     27      5
     9   1.38634148   1.38634148     27      5
    10   1.35512315   1.35512315     27      5
    11   1.32904341   1.32904341     27      5
    12   1.30689311   1.30689311     26      5
    13   1.28782004   1.28782005     26      5
```

```
14   1.27120531   1.27120530     26      5
15   1.25658788   1.25658788     26      5
16   1.24361694   1.24361694     26      5
17   1.23202049   1.23202049     26      5
18   1.22158422   1.22158422     26      5
19   1.21213684   1.21213684     26      5
20   1.20353971   1.20353971     26      5
```

4. (a)

```
PROGRAM

%MAIV_4A
%
x0=0; x1=1; n=0;
while abs(x1-x0)>eps
  n=n+1;
  x0=x1;
  x1=exp(-x0);
end
fprintf('   n=%2.0f  alpha=%17.15f\n',n,x1)

OUTPUT

>> MAIV_4A
    n=64  alpha=0.567143290409784
>>
```

(b) Let

$$\varphi(x) = \mathrm{e}^{-x}, \quad \varphi_\omega(x) = \frac{\omega \mathrm{e}^{-x} + x}{1+\omega}.$$

Then the condition for (ultimately) faster convergence is

$$|\varphi'_\omega(\alpha)| < |\varphi'(\alpha)|,$$

that is,

$$\left| \frac{1 - \omega \mathrm{e}^{-\alpha}}{1+\omega} \right| < |\mathrm{e}^{-\alpha}|.$$

Since $\mathrm{e}^{-\alpha} = \alpha$, this is equivalent to

$$\left| \frac{1}{\alpha} - \omega \right| < |1+\omega|.$$

By plotting both sides as functions of ω, one finds that the inequality is satisfied precisely when

$$\omega > \frac{1}{2}\left(\frac{1}{\alpha} - 1\right).$$

(c) The optimal value of ω is the one for which $\varphi'_\omega(\alpha) = 0$, namely, $\omega = \mathrm{e}^\alpha = \frac{1}{\alpha} = 1.763222834351897$ (from the result in (a)). Verification:

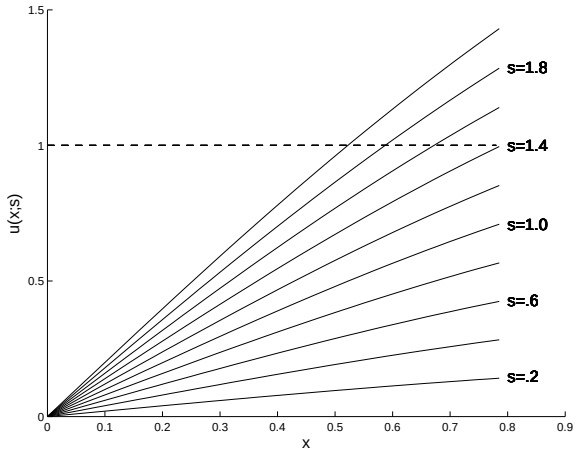
```
PROGRAM

%MAIV_4C
%
f0='%19.15f %4.0f\n';
disp('        omega         n')
for k=-5:5
  om=1.763222834351897+k/5;
  x0=0; x1=1; n=0;
  while abs(x1-x0)>eps
    n=n+1;
    x0=x1;
    x1=(om*exp(-x0)+x0)/(1+om);
  end
  fprintf(f0,om,n)
end

OUTPUT

  >> MAIV_4C
          omega         n
    0.763222834351897   32
    0.963222834351897   26
    1.163222834351897   20
    1.363222834351897   17
    1.563222834351897   13
    1.763222834351897    5
    1.963222834351897   12
    2.163222834351897   13
    2.363222834351897   16
    2.563222834351897   18
    2.763222834351897   19
  >>
```

5. (a)

```
 PROGRAMS

%MAIV_5A
%
xspan=[0 pi/4];
hold on
for s=.2:.2:2
```

```
  y0=[0 s]';
  [x,y]=ode45(@pend,xspan,y0);
  plot(x,y(:,1))
end
hold off

%PEND The function in the differential equation
%    for the pendulum
%
function yprime=pend(x,y)
yprime=zeros(2,1);
yprime(1)=y(2);
yprime(2)=-sin(y(1));
```

PLOT



(b) The graphs of (a) suggest initial values $s_0 = 1.2$, $s_1 = 1.6$.

PROGRAMS

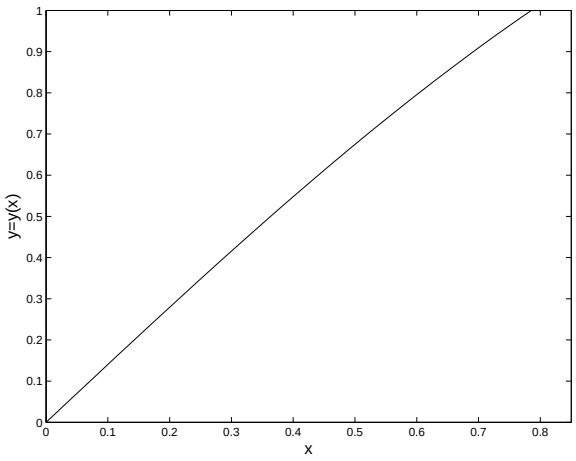
```
%MAIV_5B
%
s0=1.2; s1=1.6; tol=.5e-12;
[ntol,s]=bisec(s0,s1,tol);
fprintf('    ntol=%2.0f  s=%14.12f\n',ntol,s)

%BISEC  Bisection method
%
function [ntol,x]=bisec(a,b,tol)
ntol=ceil(log((b-a)/tol)/log(2));
```

```
for n=1:ntol
  x=(a+b)/2;
  fx=f(x);
  if fx<0
    a=x;
  else
    b=x;
  end
end

function y=f(x)
tspan=[0,pi/4]; y0=[0 x]';
options=odeset('AbsTol',.5e-12);
[t,z]=ode45(@pend,tspan,y0,options);
m=size(t,1);
y=z(m,1)-1;

 OUTPUT

>> MAIV_5B
    ntol=40  s=1.406054253886
>>
```

(c)

```
 PROGRAM

%MAIV_5C
%
xspan=[0 pi/4]; s=1.406054253885;
y0=[0 s]';
[x,y]=ode45(@pend,xspan,y0);
plot(x,y(:,1))
```

PLOT



6. See the text.

7. (a) Newton's method applied to

$$\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{0}, \quad \boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u} + h^2 \sin \boldsymbol{u} + \boldsymbol{e}_n$$

is

$$\boldsymbol{u}^{[i+1]} = \boldsymbol{u}^{[i]} + \boldsymbol{\Delta}^{[i]}, \quad \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{u}^{[i]})\boldsymbol{\Delta}^{[i]} = -\boldsymbol{f}(\boldsymbol{u}^{[i]}), \quad i = 0, 1, 2, \ldots,$$

where

$$\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{u}) = \boldsymbol{A} + h^2 \mathrm{diag}(\cos \boldsymbol{u}).$$

PROGRAM

```
%MAIV_7A
%
hold on
eps0=.5e-14;
for n=[10 100 1000]
  h=pi/(4*(n+1)); h2=h^2;
  A=zeros(n);
  A(1,1)=-2; A(1,2)=1; A(n,n-1)=1; A(n,n)=-2;
  for k=2:n-1
    A(k,k-1)=1; A(k,k)=-2; A(k,k+1)=1;
  end
  en=eye(n); en=en(:,n);
  u0=zeros(n,1); u1=zeros(n,1);
  uplot=zeros(n+2,1); uplot(n+2)=1;
  for k=1:n
```

```
    u1(k)=k*h;
  end
  it=0;
  while abs(norm(u1-u0,inf))>eps0
    it=it+1;
    u0=u1;
    r=-(A*u0+h2*sin(u0)+en);
    J=A+h2*diag(cos(u0));
    a=diag(J); b=ones(n-1,1); c=b;
    d=tridiag(n,a,b,c,r);
    u1=u0+d;
  end
  x=linspace(0,pi/4,n+2)'; uplot(2:n+1)=u1;
  fprintf('     it=%3.0f\n',it)
  plot(x,uplot)
end
hold off

 OUTPUT

  >> MAIV_7A
       it=  4
       it=  4
       it= 13
  >>
```

The graphs are essentially identical with the one in the solution to MA 5(c).

(b) The system of nonlinear equations becomes

$$\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{0}, \quad \boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u} - h\boldsymbol{g}(\boldsymbol{u}) + y0\boldsymbol{e}_1 + y1\boldsymbol{e}_n,$$

where $\boldsymbol{A}$ is the tridiagonal matrix defined in MA 6(c), $\boldsymbol{e}_1$ and $\boldsymbol{e}_n$ are the first, respectively last, column of the $n \times n$ unit matrix, and

$$[\boldsymbol{g}(\boldsymbol{u})]_k = \begin{cases} u_1(u_2 - y_0) \ \text{ if } k = 1, \\ u_k(u_{k+1} - u_{k-1}) \ \text{ if } 2 \le k \le n-1, \\ u_n(y_1 - u_{n-1}) \ \text{ if } k = n. \end{cases}$$

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

$$\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{u}) = \boldsymbol{A} - h\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{u}),$$

where

$$\boldsymbol{J_g}(\boldsymbol{u}) = \begin{bmatrix} u_2 - y_0 & u_1 & & & & \\ -u_2 & u_3 - u_1 & u_2 & & & \\ & -u_3 & u_4 - u_2 & u_3 & & \\ & \ddots & \ddots & \ddots & & \\ & & -u_{n-1} & u_n - u_{n-2} & u_{n-1} & \\ & & & -u_n & y_1 - u_{n-1} \end{bmatrix},$$

and so Newton's method is

$$\boldsymbol{u}^{[i+1]} = \boldsymbol{u}^{[i]} + \boldsymbol{\Delta}^{[i]}, \quad \boldsymbol{J_f}(\boldsymbol{u}^{[i]})\boldsymbol{\Delta}^{[i]} = -\boldsymbol{A}\boldsymbol{u}^{[i]} + h\boldsymbol{g}(\boldsymbol{u}^{[i]}) - y_0\boldsymbol{e}_1 - y_1\boldsymbol{e}_n, \quad i = 0, 1, 2, \ldots .$$

The program can be checked by running it with $y_0 = y_1 = 1$, in which case it should produce the obvious solution $y(x) \equiv 1$.

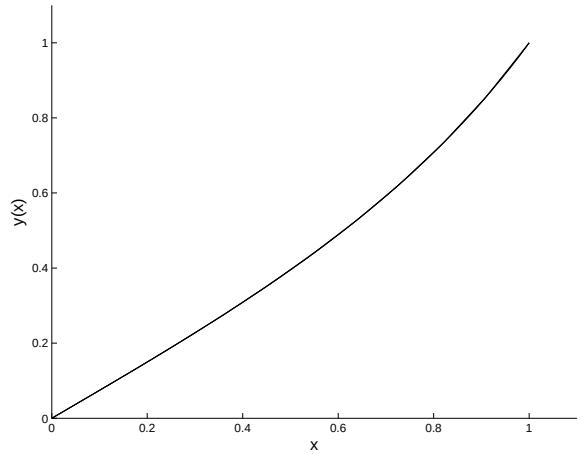
```
 PROGRAM

%MAIV_7B
%
hold on
eps0=.5e-6; y0=0; y1=1;
for n=[10 50 100]
  h=1/(n+1);
  A=zeros(n); g=zeros(n,1);
  A(1,1)=-2; A(1,2)=1; A(n,n-1)=1; A(n,n)=-2;
  for k=2:n-1
    A(k,k-1)=1; A(k,k)=-2; A(k,k+1)=1;
  end
  en=eye(n); gb=y0*en(:,1)+y1*en(:,n);
  u0=zeros(n,1); u1=zeros(n,1);
  uplot=zeros(n+2,1); uplot(1)=y0; uplot(n+2)=y1;
  for k=1:n
    u1(k)=y0+(k-1)*(y1-y0)/(n-1);
  end
  it=0;
  while abs(norm(u1-u0,inf))>eps0
    it=it+1;
    u0=u1;
    g(1)=u0(1)*(u0(2)-y0);
    g(2:n-1)=u0(2:n-1).*(u0(3:n)-u0(1:n-2));
    g(n)=u0(n)*(y1-u0(n-1));
```

```
    r=-A*u0+h*g-gb;
    dg=zeros(n);
    dg(1,1)=u0(2)-y0;
    for k=2:n-1
      dg(k,k-1)=-u0(k); dg(k,k+1)=u0(k);
      dg(k,k)=u0(k+1)-u0(k-1);
    end
    dg(n,n-1)=-u0(n); dg(n,n)=y1-u0(n-1);
    J=A-h*dg;
    a=diag(J); b=1+h*u0(2:n); c=1-h*u0(1:n-1);
    d=tridiag(n,a,b,c,r);
    u1=u0+d;
  end
  x=linspace(0,1,n+2)'; uplot(2:n+1)=u1;
  fprintf('     it=%3.0f\n',it)
  plot(x,uplot)
  axis([0 1.1 0 1.1])
  xlabel('x','Fontsize',14)
  ylabel('y(x)','FontSize',14)
end
hold off
```

OUTPUT

```
  >> MAIV_7B
       it=  4
       it=  3
       it=  3
  >>
```

PLOT



(c) The iteration function in question is

$$\boldsymbol{\varphi}(\boldsymbol{u}) = \boldsymbol{A}^{-1}(h\boldsymbol{g}(\boldsymbol{u}) - y_0\boldsymbol{e}_1 - y_1\boldsymbol{e}_n),$$

and thus

$$\boldsymbol{\varphi}(\boldsymbol{u}) - \boldsymbol{\varphi}(\boldsymbol{u}^*) = h\boldsymbol{A}^{-1}(\boldsymbol{g}(\boldsymbol{u}) - \boldsymbol{g}(\boldsymbol{u}^*)).$$

Here, $h = O(1/n)$ as $n \to \infty$, so that, by the *Hint*,

$$\|h\boldsymbol{A}^{-1}\|_\infty = O(\frac{1}{n} \cdot n^2) = O(n) \quad \text{as } n \to \infty.$$

Therefore, $\boldsymbol{\varphi}$ cannot be contractive. In fact, by an argument analogous to the one used in the proof of Theorem 4.9.1, (4.90), but with the inequalities reversed, one finds that the errors of the iterates tend to infinity.

8. See the text.

9. (a) Let π_n be the monic s-orthogonal polynomial of degree n. Define the monic polynomial $p_n(t) = (-1)^n\pi_n(-t)$. Using the change of variable $t \mapsto -t$ and symmetry, we have for an arbitrary $q \in \mathbb{P}_{n-1}$

$$\begin{aligned}\int_{-a}^{a} [p_n(t)]^{2s+1} q(t)w(t)\mathrm{d}t &= -\int_{a}^{-a} [p_n(-t)]^{2s+1} q(-t)w(-t)\mathrm{d}t \\ &= \int_{-a}^{a} [p_n(-t)]^{2s+1} q(-t)w(t)\mathrm{d}t \\ &= (-1)^{(2s+1)n} \int_{-a}^{a} [\pi_n(t)]^{2s+1} q(-t)w(t)\mathrm{d}t,\end{aligned}$$

and since $q(-t)$ is also an arbitrary polynomial of degree $< n$, we get by s-orthogonality of π_n,

$$\int_{-a}^{a} [p_n(t)]^{2s+1} q(t)w(t)\mathrm{d}t = 0, \quad q \in \mathbb{P}_{n-1}.$$

By uniqueness of π_n, there follows $p_n(t) \equiv \pi_n(t)$, i.e., $\pi_n(-t) \equiv (-1)^n\pi_n(t)$ as claimed in the *Hint*. For the same reason, $\pi_{k,n}(-t) = (-1)^k w_{k,n}(t)$ since $\mathrm{d}\lambda_n^s = \pi_n^{2s}\mathrm{d}\lambda$ is also symmetric.
It now follows (cf. Ch. 2, (2.44)) that $\alpha_\nu = 0$ for all ν, and therefore $f_{2\nu+1} = 0$ in (4.17), since the integrand is an odd function. Thus, we have only n unknowns β_ν, $\nu = 0, 1, \ldots, n-1$, and the system (4.17) simplifies to

$$\begin{aligned} f_0 &:= \beta_0 - \int_{\mathbb{R}} \pi_{n,n}^{2s}(t)\mathrm{d}\lambda(t) = 0, \\ f_{2\nu} &:= \int_{\mathbb{R}} [\beta_\nu \pi_{\nu-1,n}^2(t) - \pi_{\nu,n}^2(t)]\pi_{n,n}^{2s}(t)\mathrm{d}\lambda(t) = 0, \quad \nu = 1, \ldots, n-1. \end{aligned}$$

(b) When $n = 2$, we have $\pi_1(t) = t$, $\pi_2(t) = t\pi_1(t) - \beta_1 = t^2 - \beta_1$, and the system of equations can be written as

$$\beta_0 - \int_{\mathbb{R}} (t^2 - \beta_1)^{2s} \mathrm{d}\lambda(t) = 0,$$
$$\int_{\mathbb{R}} (\beta_1 - t^2)(t^2 - \beta_1)^{2s} \mathrm{d}\lambda(t) = 0.$$

The last equation, for $s = 1$, is a cubic equation in β_1, namely

$$\int_{\mathbb{R}} (\beta_1 - t^2)^3 \mathrm{d}\lambda(t) = \mu_0\beta_1^3 - 3\mu_2\beta_1^2 + 3\mu_4\beta_1 - \mu_6 = 0 \quad (s = 1),$$

where μ_k are the moments of $\mathrm{d}\lambda$. Once β_1 is found (by the known uniqueness of π_2, there can be only *one* real root), one computes β_0 from the first equation,

$$\beta_0 = \mu_4 - 2\mu_2\beta_1 + \mu_0\beta_1^2 \quad (s = 1).$$

Likewise, for $s = 2$, one obtains for β_1 a quintic equation,

$$\mu_0\beta_1^5 - 5\mu_2\beta_1^4 + 10\mu_4\beta_1^3 - 10\mu_6\beta_1^2 + 5\mu_8\beta_1 - \mu_{10} = 0 \quad (s = 2),$$

from which, once solved, β_0 is obtained by

$$\beta_0 = \mu_8 - 4\mu_6\beta_1 + 6\mu_4\beta_1^2 - 4\mu_2\beta_1^3 + \mu_0\beta_1^4 \quad (s = 2).$$

For the Legendre measure, we have

$$\mu_{2m} = \frac{2}{2m+1} \quad \text{(Legendre)}.$$

Thus, for $s = 1$,

$$\beta_1^3 - \beta_1^2 + \frac{3}{5}\beta_1 - \frac{1}{7} = 0, \ \ \beta_1 = .395906644039, \ \ \beta_0 = .185608616203 \ \text{(Legendre)}$$

and, for $s = 2$,

$$\beta_1^5 - \tfrac{5}{3}\beta_1^4 + 2\beta_1^3 - \tfrac{10}{7}\beta_1^2 + \tfrac{5}{9}\beta_1 - \tfrac{1}{11} = 0,$$
$$\beta_1 = .422536080289, \ \ \beta_0 = .0303941826257 \ \text{(Legendre)}.$$

In our program, for each $n \geq 2$, the initial approximations for $\beta_0, \beta_1, \ldots, \beta_{n-2}$ are the corresponding final values for $n - 1$, while β_{n-1} is taken to be the final value β_{n-2}.

(c)

```
 PROGRAM

%MAIV_9C
%
f0='%4.0f %19.12e %19.12e\n';
global n s ns xw
options=optimset('TolFun',.5e-12,'TolX',.5e-12);
disp('   n        beta                zeros')
s=1;
%s=2;
for n=2:10
  if n==2, x0=[.19 .4]'; end
  ns=(s+1)*n;
  ab=r_jacobi(ns); xw=gauss(ns,ab);
  x=fsolve(@turan,x0,options);
  x0=[x;x(n)];
  abturan=[zeros(n,1) x];
  xwturan=gauss(n,abturan);
  for k=1:n
    fprintf(f0,k,x(k),xwturan(k,1))
  end
  fprintf('\n')
end

%TURAN
%
function y=turan(x)
global n s ns xw
y=zeros(n,1); pi=zeros(n+1,1);
for k=1:ns
  t=xw(k,1); pi(1)=1; pi(2)=t;
  for j=2:n
    pi(j+1)=t*pi(j)-x(j)*pi(j-1);
  end
  y(1)=y(1)+xw(k,2)*(pi(n+1))^(2*s);
  for j=2:n
    y(j)=y(j)+xw(k,2)*(x(j)*(pi(j-1))^2 ...
      -(pi(j))^2)*(pi(n+1))^(2*s);
  end
end
y(1)=x(1)-y(1);
```

OUTPUT

s=1

```
>> MAIV_9C
   n        beta                zeros
Optimization terminated: first-order optimality is less than options.TolFun.
   1  1.856086162032e-01 -6.292111283499e-01
   2  3.959066440394e-01  6.292111283499e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  4.838648998030e-02 -8.144391855742e-01
   2  3.963906154348e-01  9.367506770275e-17
   3  2.669205715641e-01  8.144391855742e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  1.236597083504e-02 -8.896768136056e-01
   2  3.970139938677e-01 -3.588585270389e-01
   3  2.665433448640e-01  3.588585270389e-01
   4  2.567469363641e-01  8.896768136056e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  3.133547309797e-03 -9.271178696099e-01
   2  3.975143795566e-01 -5.608674191616e-01
   3  2.664214804359e-01 -8.219213553428e-17
   4  2.565093538962e-01  5.608674191616e-01
   5  2.536745921383e-01  9.271178696099e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  7.905553218592e-04 -9.483420576076e-01
   2  3.978988727494e-01 -6.824365742882e-01
   3  2.663827081094e-01 -2.479153922227e-01
   4  2.564049174909e-01  2.479153922227e-01
   5  2.535215703772e-01  6.824365742882e-01
   6  2.523263091276e-01  9.483420576076e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  1.989365689848e-04 -9.615030345523e-01
   2  3.981964532878e-01 -7.604121161599e-01
   3  2.663761200611e-01 -4.189145018946e-01
   4  2.563535417337e-01 -1.650801983293e-16
   5  2.534448028328e-01  4.189145018946e-01
   6  2.522224401216e-01  7.604121161599e-01
   7  2.516106737167e-01  9.615030345523e-01
```

```
Optimization terminated: first-order optimality is less than options.TolFun.
   1  4.997961259234e-05 -9.702155861293e-01
   2  3.984307066578e-01 -8.131296793316e-01
   3  2.663829274050e-01 -5.396700823082e-01
   4  2.563271092156e-01 -1.889104595029e-01
   5  2.534020136367e-01  1.889104595029e-01
   6  2.521657035084e-01  5.396700823082e-01
   7  2.515369573165e-01  8.131296793316e-01
   8  2.511837006862e-01  9.702155861293e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  1.254291114725e-05 -9.762772745723e-01
   2  3.986184404503e-01 -8.503253986884e-01
   3  2.663952652890e-01 -6.273217417917e-01
   4  2.563136195385e-01 -3.326619162821e-01
   5  2.533766746745e-01  2.940482448004e-17
   6  2.521317324448e-01  3.326619162821e-01
   7  2.514940832140e-01  6.273217417917e-01
   8  2.511294715822e-01  8.503253986884e-01
   9  2.509078315768e-01  9.762772745723e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  3.145366900599e-06 -9.806625959365e-01
   2  3.987714142764e-01 -8.775002209847e-01
   3  2.664095892884e-01 -6.926244251400e-01
   4  2.563072802521e-01 -4.432009919506e-01
   5  2.533611556214e-01 -1.524705876794e-01
   6  2.521101749001e-01  1.524705876794e-01
   7  2.514670877104e-01  4.432009919506e-01
   8  2.510963341675e-01  6.926244251400e-01
   9  2.508667578946e-01  8.775002209847e-01
  10  2.507189644598e-01  9.806625959365e-01
>>
                      s=2

>> MAIV_9C
   n        beta                zeros
Optimization terminated: first-order optimality is less than options.TolFun.
   1  3.039418262568e-02 -6.500277534758e-01
   2  4.225360802888e-01  6.500277534758e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  2.043984748496e-03 -8.292327380625e-01
   2  4.232211545600e-01 -7.112366251505e-17
   3  2.644057793147e-01  8.292327380625e-01
```

```
Optimization terminated: first-order optimality is less than options.TolFun.
   1  1.327030887978e-04 -8.998292126760e-01
   2  4.241025816349e-01 -3.659243547587e-01
   3  2.638488490075e-01  3.659243547587e-01
   4  2.556418147484e-01  8.998292126760e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  8.488833743153e-06 -9.343689749286e-01
   2  4.248126649785e-01 -5.690324387923e-01
   3  2.636394138044e-01  1.156924084810e-17
   4  2.553211416427e-01  5.690324387923e-01
   5  2.530700772813e-01  9.343689749286e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  5.389068533395e-07 -9.537394817237e-01
   2  4.253614030717e-01 -6.900313943868e-01
   3  2.635503063580e-01 -2.512763505456e-01
   4  2.551708906467e-01  2.512763505456e-01
   5  2.528713367043e-01  6.900313943868e-01
   6  2.519481918009e-01  9.537394817237e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  3.406161328030e-08 -9.656626615626e-01
   2  4.257888730849e-01 -7.670751801332e-01
   3  2.635119821833e-01 -4.235791133204e-01
   4  2.550905821217e-01  1.044003205992e-16
   5  2.527674893217e-01  4.235791133204e-01
   6  2.518162306917e-01  7.670751801332e-01
   7  2.513528157506e-01  9.656626615626e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  2.146876193310e-09 -9.735135417885e-01
   2  4.261275068850e-01 -8.188788001555e-01
   3  2.634984271752e-01 -5.446647773413e-01
   4  2.550443346246e-01 -1.908528748441e-01
   5  2.527065918393e-01  1.908528748441e-01
   6  2.517417284875e-01  5.446647773413e-01
   7  2.512599092962e-01  8.188788001555e-01
   8  2.509971465943e-01  9.735135417885e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  1.350651227393e-10 -9.789532522124e-01
   2  4.264009031563e-01 -8.552762568552e-01
   3  2.634954777013e-01 -6.322102949267e-01
```

```
   4  2.550167018716e-01 -3.356221896577e-01
   5  2.526685560198e-01 -3.030861827561e-16
   6  2.516958795270e-01  3.356221896577e-01
   7  2.512054283310e-01  6.322102949267e-01
   8  2.509295070676e-01  8.552762568552e-01
   9  2.507666030849e-01  9.789532522124e-01

Optimization terminated: first-order optimality is less than options.TolFun.
   1  8.489223366415e-12 -9.828838901362e-01
   2  4.264009031563e-01 -8.818010639719e-01
   3  2.634954777013e-01 -6.972573977883e-01
   4  2.550167018716e-01 -4.466618391288e-01
   5  2.526685560198e-01 -1.537436978499e-01
   6  2.516958795270e-01  1.537436978499e-01
   7  2.512054283310e-01  4.466618391288e-01
   8  2.509295070676e-01  6.972573977883e-01
   9  2.507666030849e-01  8.818010639719e-01
  10  2.507666030849e-01  9.828838901362e-01
>>
```

EXERCISES AND MACHINE ASSIGNMENTS TO CHAPTER 5

EXERCISES

1. Consider the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \kappa(y + y^3), \quad 0 \le x \le 1; \quad y(0) = s,$$

where $\kappa > 0$ (in fact, $\kappa \gg 1$) and $s > 0$. Under what conditions on s does the solution $y(x) = y(x; s)$ exist on the whole interval $[0, 1]$? {*Hint*: find y explicitly.}

2. Prove (5.46).

3. Prove

$$(\boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f})_{\boldsymbol{y}}\boldsymbol{f} = \boldsymbol{f}^T\boldsymbol{f}_{\boldsymbol{yy}}\boldsymbol{f} + \boldsymbol{f}_{\boldsymbol{y}}^2\boldsymbol{f}.$$

4. Let

$$\boldsymbol{f}(x, \boldsymbol{y}) = \begin{bmatrix} f^1(x, \boldsymbol{y}) \\ f^2(x, \boldsymbol{y}) \\ \vdots \\ f^d(x, \boldsymbol{y}) \end{bmatrix}$$

be a C^1 map from $[a, b] \times \mathbb{R}^d$ to $\mathbb{R}^d$. Assume that

$$\left| \frac{\partial f^i(x, \boldsymbol{y})}{\partial y^j} \right| \le M_{ij} \quad \text{on} \quad [a, b] \times \mathbb{R}^d, \quad i, j = 1, 2, \ldots, d,$$

where M_{ij} are constants independent of x and $\boldsymbol{y}$, and let $\boldsymbol{M} = [M_{ij}] \in \mathbb{R}_+^{d \times d}$. Determine a Lipschitz constant L of $\boldsymbol{f}$

   (a) in the ℓ_1 vector norm;

   (b) in the ℓ_2 vector norm;

   (c) in the ℓ_∞ vector norm.

   Express L, if possible, in terms of a matrix norm of $\boldsymbol{M}$.

5. (a) Write the system of differential equations

$$u''' = x^2uu'' - uv',$$
$$v'' = xvv' + 4u'$$

   as a first-order system of differential equations, $\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$.

   (b) Determine the Jacobian matrix $\boldsymbol{f}_{\boldsymbol{y}}(x, \boldsymbol{y})$ for the system in (a).

   (c) Determine a Lipschitz constant L for $\boldsymbol{f}$ on $[0, 1] \times \mathcal{D}$, where $\mathcal{D} = \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y}\|_1 \le 1\}$, using, respectively, the ℓ_1, ℓ_2, and ℓ_∞ norm (cf. Ex. 4).

6. For the (scalar) differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = y^{\lambda}, \quad \lambda > 0,$$

   (a) determine the principal error function of the general explicit two-stage Runge–Kutta method (5.56), (5.57);

   (b) compare the local accuracy of the modified Euler method with that of Heun's method;

   (c) determine a λ-interval such that for each λ in this interval, there is a two-stage explicit Runge–Kutta method of order $p = 3$ having parameters $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$, and $0 < \mu < 1$.

7. For the implicit Euler method

$$\boldsymbol{y}_{\text{next}} = \boldsymbol{y} + h\boldsymbol{f}(x + h, \boldsymbol{y}_{\text{next}}),$$

   (a) state a condition under which $\boldsymbol{y}_{\text{next}}$ is uniquely defined;

   (b) determine the order and principal error function.

8. Show that any explicit two-stage Runge–Kutta method of order $p = 2$ integrates the special scalar differential equation $\mathrm{d}y/\mathrm{d}x = f(x)$, $f \in \mathbb{P}_1$, exactly.

9. The (scalar) second-order differential equation

$$\frac{\mathrm{d}^2 z}{\mathrm{d}x^2} = g(x, z),$$

   in which g does not depend on $\mathrm{d}z/\mathrm{d}x$, can be written as a first-order system

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \begin{bmatrix} y^2 \\ g(x, y^1) \end{bmatrix}$$

   by letting, as usual, $y^1 = z$, $y^2 = \mathrm{d}z/\mathrm{d}x$. For this system, consider a one-step method $\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + h\boldsymbol{\Phi}(x_n, \boldsymbol{u}_n; h)$ with

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \begin{bmatrix} y^2 + \frac{1}{2}hk(x, \boldsymbol{y}; h) \\ k(x, \boldsymbol{y}; h) \end{bmatrix}, \quad k = g(x + \mu h, y^1 + \mu h y^2), \quad \boldsymbol{y} = \begin{bmatrix} y^1 \\ y^2 \end{bmatrix}.$$

   (Note that this method requires only one evaluation of g per step.)

   (a) Can the method be made to have order $p = 2$, and if so, for what value(s) of μ?

   (b) Determine the principal error function of any method obtained in (a).

10. Show that the first condition in (5.67) is equivalent to the condition that

$$\boldsymbol{k}_s(x, \boldsymbol{y}; h) = \boldsymbol{u}'(x + \mu_s h) + O(h^2), \quad s \geq 2,$$

   where $\boldsymbol{u}(t)$ is the reference solution through the point $(x, \boldsymbol{y})$.

11. Suppose that

$$\int_x^{x+h} z(t)\mathrm{d}t = h\sum_{k=1}^{\nu} w_k z(x+\vartheta_k h) + ch^{\mu+1}z^{(\mu)}(\xi)$$

is a quadrature formula with $w_k \in \mathbb{R}$, $\vartheta_k \in [0,1]$, $c \neq 0$, and $\xi \in (x, x+h)$, for z sufficiently smooth. Given increment functions $\overline{\boldsymbol{\Phi}}_k(x, \boldsymbol{y}; h)$ defining methods of order $\overline{p}_k$, $k = 1, 2, \ldots, \nu$, show that the one-step method defined by

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \sum_{k=1}^{\nu} w_k \boldsymbol{f}(x+\vartheta_k h, \boldsymbol{y} + \vartheta_k h \overline{\boldsymbol{\Phi}}_k(x, \boldsymbol{y}; \vartheta_k h))$$

has order p at least equal to $\min(\mu, \overline{p}+1)$, where $\overline{p} = \min \overline{p}_k$.

12. Let $\boldsymbol{g}(x, \boldsymbol{y}) = (\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f})(x, \boldsymbol{y})$. Show that the one-step method defined by the increment function

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \boldsymbol{f}(x, \boldsymbol{y}) + \tfrac{1}{2}h\boldsymbol{g}(x + \tfrac{1}{3}h, \boldsymbol{y} + \tfrac{1}{3}h\boldsymbol{f}(x, \boldsymbol{y}))$$

has order $p = 3$. Express the principal error function in terms of $\boldsymbol{g}$ and its derivatives.

13. Let $\boldsymbol{f}(x, \boldsymbol{y})$ satisfy a Lipschitz condition in $\boldsymbol{y}$ on $[a, b] \times \mathbb{R}^d$, with Lipschitz constant L.

    (a) Show that the increment function $\boldsymbol{\Phi}$ of the second-order Runge–Kutta method

$$\begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}(x, \boldsymbol{y}), \\ \boldsymbol{k}_2 &= \boldsymbol{f}(x+h, \boldsymbol{y} + h\boldsymbol{k}_1), \\ \boldsymbol{\Phi}(x, \boldsymbol{y}; h) &= \tfrac{1}{2}(\boldsymbol{k}_1 + \boldsymbol{k}_2) \end{aligned}$$

    also satisfies a Lipschitz condition whenever $x+h \in [a, b]$, and determine a respective Lipschitz constant M.

    (b) What would the result be for the classical Runge–Kutta method?

    (c) What would it be for the general implicit Runge–Kutta method?

14. Describe the application of Newton's method to implement the implicit Runge–Kutta method.

15. Consider the following scheme of constructing an estimator $\boldsymbol{r}(x, \boldsymbol{y}; h)$ for the

principal error function $\boldsymbol{\tau}(x, \boldsymbol{y})$ of Heun's method:

$$\begin{aligned}
\boldsymbol{k}_1 &= \boldsymbol{f}(x, \boldsymbol{y}),\\
\boldsymbol{k}_2 &= \boldsymbol{f}(x + h, \boldsymbol{y} + h\boldsymbol{k}_1),\\
\boldsymbol{y}_h &= \boldsymbol{y} + \tfrac{1}{2}\, h(\boldsymbol{k}_1 + \boldsymbol{k}_2),\\
\boldsymbol{k}_3 &= \boldsymbol{f}(x + h, \boldsymbol{y}_h),\\
\boldsymbol{k}_4 &= \boldsymbol{f}(x + h + \mu h, \boldsymbol{y}_h + \mu h \boldsymbol{k}_3),\\
\boldsymbol{r}(x, \boldsymbol{y}; h) &= h^{-2}(\beta_1 \boldsymbol{k}_1 + \beta_2 \boldsymbol{k}_2 + \beta_3 \boldsymbol{k}_3 + \beta_4 \boldsymbol{k}_4).
\end{aligned}$$

(Note that this scheme requires one additional function evaluation, $\boldsymbol{k}_4$, beyond what would be required anyhow to carry out Heun's method.) Obtain the conditions on the parameters μ, β_1, β_2, β_3, β_4 in order that

$$\boldsymbol{r}(x, \boldsymbol{y}; h) = \boldsymbol{\tau}(x, \boldsymbol{y}) + O(h).$$

Show, in particular, that there is a unique set of βs for any μ with $\mu(\mu+1) \neq 0$. What is a good choice of the parameters, and why?

16. Apply the asymptotic error formula (5.104) to the (scalar) initial value problem $\mathrm{d}y/\mathrm{d}x = \lambda y$, $y(0) = 1$, on $[0, 1]$, when solved by the classical fourth-order Runge–Kutta method. In particular, determine

$$\lim_{h \to 0} h^{-4} \frac{u_N - y(1)}{y(1)},$$

where u_N is the Runge–Kutta approximation to $y(1)$ obtained with step $h = 1/N$.

17. Consider $y' = \lambda y$ on $[0, \infty)$ for complex λ with $\operatorname{Re}\lambda < 0$. Let $\{u_n\}$ be the approximations to $\{y(x_n)\}$ obtained by the classical fourth-order Runge–Kutta method with the step h held fixed. (That is, $x_n = nh$, $h > 0$, and $n = 0, 1, 2, \ldots$.)

    (a) Show that $y(x) \to 0$ as $x \to \infty$, for any initial value y_0.

    (b) Under what condition on h can we assert that $u_n \to 0$ as $n \to \infty$? In particular, what is the condition if λ is real (negative)?

    (c) What is the analogous result for Euler's method?

    (d) Generalize to systems $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A}$ is a constant matrix all of whose eigenvalues have negative real parts.

18. Show that any one-step method of order p, which, when applied to the model problem $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}$, yields

$$\boldsymbol{y}_{\text{next}} = \varphi(h\boldsymbol{A})\boldsymbol{y}, \qquad \varphi \text{ a polynomial of degree } q \geq p,$$

must have

$$\varphi(z) = 1 + z + \tfrac{1}{2!}z^2 + \cdots + \tfrac{1}{p!}z^p + z^{p+1}\chi(z),$$

where χ is identically zero if $q = p$, and a polynomial of degree $q - p - 1$ otherwise. In particular, show that $\chi \equiv 0$ for a p-stage explicit Runge–Kutta method of order p, $1 \le p \le 4$, and for the Taylor expansion method of order $p \ge 1$.

19. Consider the linear homogeneous system

$$(*) \qquad \boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}, \quad \boldsymbol{y} \in \mathbb{R}^d,$$

with constant coefficient matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$.

(a) For Euler's method applied to (*), determine $\varphi(z)$ (cf. (5.140)) and the principal error function.

(b) Do the same for the classical fourth-order Runge–Kutta method.

20. Consider the model equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = a(x)[y - b(x)], \quad 0 \le x < \infty,$$

where $a(x)$, $b(x)$ are continuous and bounded on $\mathbb{R}_+$, and $a(x)$ negative with $|a(x)|$ large, say,

$$a \le |a(x)| \le A \ \text{ on } \mathbb{R}_+, \quad a \gg 1.$$

For the explicit and implicit Euler methods, derive a condition (if any) on the step length h that ensures boundedness of the respective approximations u_n as $x_n = nh \to \infty$ for $h > 0$ fixed. (Assume, in the case of the explicit Euler method, that a is so large that $ah > 1$.)

21. Consider the implicit one-step method

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \boldsymbol{k}(x, \boldsymbol{y}; h),$$

where $\boldsymbol{k} : [a, b] \times \mathbb{R}^d \times (0, h_0] \to \mathbb{R}^d$ is implicitly defined, in terms of total derivatives of $\boldsymbol{f}$, by

$$\boldsymbol{k} = \sum_{s=1}^{r} h^{s-1}[\alpha_s \boldsymbol{f}^{[s-1]}(x, \boldsymbol{y}) - \beta_s \boldsymbol{f}^{[s-1]}(x + h, \boldsymbol{y} + h\boldsymbol{k})],$$

with suitable constants α_s and β_s (Ehle's method; cf. Sect. 5.9.3(4)).

(a) Show how the method works on the model problem $\mathrm{d}y/\mathrm{d}x = \lambda y$. What is the maximum possible order in this case? Is the resulting method (of maximal order) A-stable?

(b) We may associate with the one-step method the quadrature rule

$$\int_x^{x+h} g(t)\mathrm{d}t = \sum_{s=1}^{r} h^s[\alpha_s g^{(s-1)}(x) - \beta_s g^{(s-1)}(x+h)] + E(g).$$

Given any p with $r \le p \le 2r$, show that α_s, β_s can be chosen so as to have $E(g) = O(h^{p+1})$ when $g(t) = e^{t-x}$.

(c) With α_s, β_s chosen as in (b), prove that $E(g) = O(h^{p+1})$ for any $g \in C^p$ (not just for $g(t) = e^{t-x}$). {*Hint*: expand $E(g)$ in powers of h through h^p inclusive; then specialize to $g(t) = e^{t-x}$ and draw appropriate conclusions.}

(d) With α_s, β_s chosen as in (b), show that the implicit one-step method has order p if $\boldsymbol{f} \in C^p$. {*Hint*: use the definition of truncation error and Lipschitz conditions on the total derivatives $\boldsymbol{f}^{[s-1]}$.}

(e) Work out the optimal one-step method with $r = 2$ and order $p = 4$.

(f) How can you make the method L-stable (cf. (5.170)) and have maximum possible order? Illustrate with $r = 2$.

MACHINE ASSIGNMENTS

1. (a) Write Matlab routines implementing the basic step $(x, \boldsymbol{y}) \mapsto (x + h, \boldsymbol{y}_{\text{next}})$ in the case of Euler's method and the classical fourth-order Runge–Kutta method, entering the function $\boldsymbol{f}$ of the differential equationi $\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$ as an input function.

(b) Consider the initial value problem

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}, \quad 0 \le x \le 1, \quad \boldsymbol{y}(0) = \mathbf{1},$$

where

$$\boldsymbol{A} = \frac{1}{2}\begin{bmatrix} \lambda_2+\lambda_3 & \lambda_3-\lambda_1 & \lambda_2-\lambda_1 \\ \lambda_3-\lambda_2 & \lambda_1+\lambda_3 & \lambda_1-\lambda_2 \\ \lambda_2-\lambda_3 & \lambda_1-\lambda_3 & \lambda_1+\lambda_2 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

The exact solution is

$$\boldsymbol{y}(x) = \begin{bmatrix} y^1 \\ y^2 \\ y^3 \end{bmatrix}, \quad \begin{aligned} y^1 &= -e^{\lambda_1 x} + e^{\lambda_2 x} + e^{\lambda_3 x}, \\ y^2 &= e^{\lambda_1 x} - e^{\lambda_2 x} + e^{\lambda_3 x}, \\ y^3 &= e^{\lambda_1 x} + e^{\lambda_2 x} - e^{\lambda_3 x}. \end{aligned}$$

Integrate the initial value problem with constant step length $h = 1/N$ by

(i) Euler's method (order $p = 1$);

(ii) the classical Runge–Kutta method (order $p = 4$),

using the programs written in (a). In each case, along with the approximation vectors $\boldsymbol{u}_n \in \mathbb{R}^3$, $n = 1, 2, \ldots, N$, generate vectors $\boldsymbol{v}_n \in \mathbb{R}^3$, $n = 1, 2, \ldots, N$, that approximate the solution of the variational equation according to Theorem 5.8.1. (For the estimate $\boldsymbol{r}(x, \boldsymbol{y}; h)$ of the principal error function take the true value $\boldsymbol{r}(x, \boldsymbol{y}; h) = \boldsymbol{\tau}(x, \boldsymbol{y})$ according to Ex. 19.) In this way obtain estimates $\tilde{\boldsymbol{e}}_n = h^p \boldsymbol{v}_n$ ($p =$ order of the method) of the global errors $\boldsymbol{e}_n = \boldsymbol{u}_n - \boldsymbol{y}(x_n)$. Use $N = 5, 10, 20, 40, 80$, and print x_n, $\|\boldsymbol{e}_n\|_\infty$, and $\|\tilde{\boldsymbol{e}}_n\|_\infty$ for $x_n = .2 : .2 : 1$.

Suggested λ-values are

(i) $\lambda_1 = -1,\ \lambda_2 = 0,\ \lambda_3 = 1$;

(ii) $\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -10$;

(iii) $\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -40$;

(iv) $\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -160$.

Summarize what you learn from these examples and from others that you may wish to run.

2. Consider the initial value problem

$$y'' = \cos(xy), \quad y(0) = 1, \quad y'(0) = 0, \quad 0 \le x \le 1.$$

(a) Does the solution $y(x)$ exist on the whole interval $0 \le x \le 1$? Explain.

(b) Use a computer algebra system, for example Maple, to determine the Maclaurin expansion of the solution $y(x)$ up to, and including, the term with x^{50}. Evaluate the expansion to 15 decimal digits for $x = .25 : .25 : 1.0$.

(c) Describe in detail the generic step of the classical fourth-order Runge–Kutta method applied to this problem.

(d) Use the 4th-order Runge–Kutta routine `RK4.m` of MA1(a), in conjunction with a function `fMAV_2.m` appropriate for this assignment, to produce approximations $\boldsymbol{u}_n \approx \boldsymbol{y}(x_n)$ at $x_n = n/N$, $n = 0, 1, 2, \ldots, N$, for $N = [4, 16, 64, 256]$. Print the results $y(x_n), y'(x_n)$ to 12 decimal places for $x_n = .25 : .25 : 1.0$, including the errors $e_n = |u_n^1 - y(x_n)|$. (Use the Taylor expansion of (b) to compute $y(x_n)$.) Plot the solution y, y' obtained with $N = 256$.

3. On the interval $[2\sqrt{q}, 2\sqrt{q} + 1]$, $q \ge 0$ an integer, consider the initial value problem (Fehlberg, 1968)

$$\frac{\mathrm{d}^2 c}{\mathrm{d}x^2} = -\pi^2 x^2 c - \pi \frac{s}{\sqrt{c^2 + s^2}}\ ,$$
$$\frac{\mathrm{d}^2 s}{\mathrm{d}x^2} = -\pi^2 x^2 s + \pi \frac{c}{\sqrt{c^2 + s^2}}\ ,$$

with initial conditions at $x = 2\sqrt{q}$ given by

$$c = 1, \quad \frac{\mathrm{d}c}{\mathrm{d}x} = 0, \quad s = 0, \quad \frac{\mathrm{d}s}{\mathrm{d}x} = 2\pi\sqrt{q}.$$

(a) Show that the exact solution is

$$c(x) = \cos(\tfrac{\pi}{2}x^2), \quad s(x) = \sin(\tfrac{\pi}{2}x^2).$$

(b) Write the problem as an initial value problem for a system of first-order differential equations.

(c) Consider the Runge–Kutta–Fehlberg $(3, 4)$ pair $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}^*$ given by

$$\boldsymbol{k}_1 = \boldsymbol{f}(x, \boldsymbol{y}),$$

$$\boldsymbol{k}_2 = \boldsymbol{f}(x + \tfrac{2}{7}h, \boldsymbol{y} + \tfrac{2}{7}h\boldsymbol{k}_1),$$

$$\boldsymbol{k}_3 = \boldsymbol{f}(x + \tfrac{7}{15}h, \boldsymbol{y} + \tfrac{77}{900}h\boldsymbol{k}_1 + \tfrac{343}{900}h\boldsymbol{k}_2),$$

$$\boldsymbol{k}_4 = \boldsymbol{f}(x + \tfrac{35}{38}h, \boldsymbol{y} + \tfrac{805}{1444}h\boldsymbol{k}_1 - \tfrac{77175}{54872}h\boldsymbol{k}_2 + \tfrac{97125}{54872}h\boldsymbol{k}_3),$$

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \tfrac{79}{490}\boldsymbol{k}_1 + \tfrac{2175}{3626}\boldsymbol{k}_3 + \tfrac{2166}{9065}\boldsymbol{k}_4,$$

respectively

$$\boldsymbol{k}_1,\ \boldsymbol{k}_2,\ \boldsymbol{k}_3,\ \boldsymbol{k}_4 \ \text{ as previously},$$

$$\boldsymbol{k}_5 = \boldsymbol{f}(x + h, \boldsymbol{y} + h\boldsymbol{\Phi}(x, \boldsymbol{y}; h)),$$

$$\boldsymbol{\Phi}^*(x, \boldsymbol{y}; h) = \tfrac{229}{1470}\boldsymbol{k}_1 + \tfrac{1125}{1813}\boldsymbol{k}_3 + \tfrac{13718}{81585}\boldsymbol{k}_4 + \tfrac{1}{18}\boldsymbol{k}_5.$$

Solve the initial value problem in (b) for $q = 0(1)3$ by the method $\boldsymbol{\Phi}$, using constant step length $h = .2$. Repeat the integration with half the step length, and keep repeating (and halving the step) until $\max_n \|\boldsymbol{u}_n - \boldsymbol{y}(x_n)\|_\infty \le .5 \times 10^{-6}$, where $\boldsymbol{u}_n$, $\boldsymbol{y}(x_n)$ are the approximate resp. exact solution vectors at $x_n = 2\sqrt{q} + nh$. For each run print

$$q, \quad h, \quad \max_n \|\boldsymbol{u}_n - \boldsymbol{y}(x_n)\|_\infty, \quad \max_n |c_n^2 + s_n^2 - 1|,$$

where c_n, s_n are the approximate values obtained for $c(x_n)$ resp. $s(x_n)$, and the maxima are taken over $n = 1, 2, \ldots, N$ with N such that $Nh = 1$.

(d) For the same values of q as in (c) and $h = .2, .1, .05, .025, .0125$, print the global and (estimated) local errors,

$$q, \quad h, \quad \|\boldsymbol{u}_{n+1} - \boldsymbol{y}(x_{n+1})\|_\infty, \quad h\|\boldsymbol{\Phi}(x_n, \boldsymbol{u}_n; h) - \boldsymbol{\Phi}^*(x_n, \boldsymbol{u}_n; h)\|_\infty,$$

for $x_n = 0(.2).8$.

(e) Implement Theorem 5.8.1 on global error estimation, using the Runge–Kutta–Fehlberg $(3,4)$ method $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}^*$ of (c) and the estimator $\boldsymbol{r}(x,\boldsymbol{y};h) = h^{-3}[\boldsymbol{\Phi}(x,\boldsymbol{y};h) - \boldsymbol{\Phi}^*(x,\boldsymbol{y};h)]$ of the principal error function of $\boldsymbol{\Phi}$. For the same values of q as in (d), and for $h = .05, .025, .0125$, print the exact and estimated global errors,

$$q, \quad h, \quad \|\boldsymbol{u}_{n+1} - \boldsymbol{y}(x_{n+1})\|_\infty, \quad h^3\|\boldsymbol{v}_{n+1}\|_\infty \quad \text{for} \quad x_n = 0 : .2 : .8.$$

4. (a) Let $f(z) = 1 + \frac{1}{1!}z + \frac{1}{2!}z^2 + \cdots + \frac{1}{p!}z^p$. For $p = 1(1)4$ write a Matlab program, using the `contour` command, to plot the lines along which $|f(z)| = r$, $r = .1(.1)1$ (level lines of f) and the lines along which $\arg f(z) = \theta$, $\theta = 0(\frac{1}{8}\pi)2\pi - \frac{1}{8}\pi$ (phase lines of f).

   (b) For any analytic function f, derive differential equations for the level and phase lines of f. {*Hint*: write $f(z) = r\exp(\mathrm{i}\theta)$ and use θ as the independent variable for the level lines, and r as the independent variable for the phase lines. In each case, introduce arc length as the final independent variable.}

   (c) Use the Matlab function `ode45` to compute from the differential equation of (b) the level lines $|f(z)| = 1$ of the function f given in (a), for $p = 1(1)21$; these determine the regions of absolute stability of the Taylor expansion method (cf. Ex. 18). {*Hint*: use initial conditions at the origin. Produce only those parts of the curves that lie in the upper half-plane (why?). To do so in Matlab, let `ode45` run sufficiently long, interpolate between the first pair of points lying on opposite sides of the real axis to get a point on the axis, and then delete the rest of the data before plotting.}

5. Newton's equations for the motion of a particle on a planar orbit (with eccentricity ε, $0 < \varepsilon < 1$) are

$$x'' = -\frac{x}{r^3}, \quad x(0) = 1 - \varepsilon, \quad x'(0) = 0,$$

$$y'' = -\frac{y}{r^3}, \quad y(0) = 0, \quad y'(0) = \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\,, \qquad t \geq 0,$$

   where

$$r^2 = x^2 + y^2.$$

   (a) Verify that the solution can be written in the form $x(t) = \cos u - \varepsilon$, $y(t) = \sqrt{1-\varepsilon^2}\sin u$, where u is the solution of Kepler's equation $u - \varepsilon\sin u - t = 0$.

   (b) Reformulate the problem as an initial value problem for a system of first-order differential equations.

(c) Write a Matlab program for solving the initial value problem in (b) on the interval $[0, 20]$ by the classical Runge–Kutta method, for $\varepsilon = .3, .5, \text{and} .7$. Use step lengths $h = 1/N$, $N = [40, 80, 120]$ and, along with the approximate solution $u(t; h)$, $v(t; h)$, compute and plot the approximate principal error functions $r(t; h) = h^{-4}[u(t; h) - x(t)]$, $s(t; h) = h^{-4}[v(t; h) - y(t)]$ when $N = 120$ (i.e., $h = .008333\ldots$). Compute the exact solution from the formula given in (a), using Newton's method to solve Kepler's equation.

ANSWERS TO THE EXERCISES AND MACHINE ASSIGNMENTS OF CHAPTER 5

ANSWERS TO EXERCISES

1. Considering x as a function of y, we can write the initial value problem equivalently in the form

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{\kappa}\left(\frac{1}{y} - \frac{y}{y^2+1}\right), \quad x(s) = 0.$$

Integration from s to y yields

$$x = \frac{1}{\kappa}\left[\ln y - \frac{1}{2}\,\ln(y^2+1)\right]_s^y = \frac{1}{\kappa}\ln\left(\frac{y}{\sqrt{y^2+1}}\,\frac{\sqrt{s^2+1}}{s}\right),$$

hence

$$\mathrm{e}^{\kappa x} = \frac{y}{\sqrt{y^2+1}}\,\frac{\sqrt{s^2+1}}{s}\,.$$

Squaring both sides and multiplying the result by y^2+1 gives, after rearrangement,

$$y^2\left(1 + \frac{1}{s^2} - \mathrm{e}^{2\kappa x}\right) = \mathrm{e}^{2\kappa x},$$

that is,

$$y(x;s) = \frac{\mathrm{e}^{\kappa x}}{\sqrt{1 + \frac{1}{s^2} - \mathrm{e}^{2\kappa x}}}\,, \quad x \geq 0.$$

The solution is monotonically increasing and remains bounded on $[0,1]$ if and only if

$$1 + \frac{1}{s^2} - \mathrm{e}^{2\kappa} > 0.$$

Thus, the condition on s is

$$0 < s < \frac{1}{\sqrt{\mathrm{e}^{2\kappa}-1}}\,.$$

This severely limits s if κ is large.

2. Equation (5.46) is true for $k = 0$. Assume it is true for some $k \geq 0$. Then differentiating both sides with respect to t gives

$$\begin{aligned}\boldsymbol{u}^{(k+2)}(t) &= \boldsymbol{f}_x^{[k]}(t,\boldsymbol{u}(t)) + \boldsymbol{f}_{\boldsymbol{y}}^{[k]}(t,\boldsymbol{u}(t))\boldsymbol{u}'(t)\\ &= \boldsymbol{f}_x^{[k]}(t,\boldsymbol{u}(t)) + \boldsymbol{f}_{\boldsymbol{y}}^{[k]}(t,\boldsymbol{u}(t))\boldsymbol{f}(t,\boldsymbol{u}(t))\\ &= \boldsymbol{f}^{[k+1]}(t,\boldsymbol{u}(t)),\end{aligned}$$

where in the second equality the differential equation (i.e., (5.46) for $k = 0$) has been used, and in the last equality the definition (2.8) of $\boldsymbol{f}^{[k+1]}$. Thus, (5.46) holds for $k+1$, and hence, by induction, for every $k \geq 0$.

3. We use superscripts to index components, and subscripts to indicate partial derivatives with respect to the components of $\boldsymbol{y}$:

$$f^i_j = \frac{\partial f^i}{\partial y^j}, \quad f^i_{jk} = \frac{\partial^2 f^i}{\partial y^j \partial y^k}, \quad \text{etc.}$$

It is also convenient to use the summation convention for repeated indices, e.g.,

$$f^i_j f^j = \sum_j f^i_j f^j, \quad \text{etc.}$$

Then the ith component on the left of the asserted identity is

$$(f^i_k f^k)_j f^j = (f^i_{kj} f^k + f^i_k f^k_j) f^j = f^i_{kj} f^k f^j + f^i_k f^k_j f^j,$$

whereas the ith component on the right is

$$f^k f^i_{kj} f^j + f^i_k f^k_j f^j,$$

where the first term is just the definition of the ith component of $\boldsymbol{f}^T \boldsymbol{f}_{\boldsymbol{yy}} \boldsymbol{f}$, and $f^i_k f^k_j$ the (i,j)-element of $(\boldsymbol{f}_{\boldsymbol{y}})^2$. Both expressions are evidently the same.

4. By the mean value theorem for functions of several variables, we have

$$f^i(x, \boldsymbol{y}) - f^i(x, \boldsymbol{y}^*) = \sum_{j=1}^d \frac{\partial f^i}{\partial y^j}(x, \overline{\boldsymbol{y}}_i)[y^j - (y^*)^j],$$

where $\overline{\boldsymbol{y}}_i$ is a point on the line connecting $\boldsymbol{y}$ and $\boldsymbol{y}^*$ which depends on i.

(a) Using the triangle inequality, we have, by assumption,

$$\begin{aligned}
\|\boldsymbol{f}(x, \boldsymbol{y}) - \boldsymbol{f}(x, \boldsymbol{y}^*)\|_1 &= \sum_{i=1}^d |f^i(x, \boldsymbol{y}) - f^i(x, \boldsymbol{y}^*)| \\
&\le \sum_{i=1}^d \sum_{j=1}^d \left| \frac{\partial f^i(x, \overline{\boldsymbol{y}}_i)}{\partial y^j} \right| |y^j - (y^*)^j| \\
&\le \sum_{i=1}^d \sum_{j=1}^d M_{ij} |y^j - (y^*)^j| = \sum_{j=1}^d |y^j - (y^*)^j| \sum_{i=1}^d M_{ij} \\
&\le \left(\max_j \sum_{i=1}^d M_{ij} \right) \sum_{j=1}^d |y^j - (y^*)^j|,
\end{aligned}$$

that is,

$$\|\boldsymbol{f}(x, \boldsymbol{y}) - \boldsymbol{f}(x, \boldsymbol{y}^*)\|_1 \le L \|\boldsymbol{y} - \boldsymbol{y}^*\|_1, \quad L = \|\boldsymbol{M}\|_1,$$

where $\| \cdot \|_1$ on the far right is the 1-matrix norm.

(b) Using again the triangle inequality, and combining it with the Schwarz inequality for summation (in the last inequality below), we have

$$\|\boldsymbol{f}(x,\boldsymbol{y})-\boldsymbol{f}(x,\boldsymbol{y}^*)\|_2=\left(\sum_{i=1}^d|f^i(x,\boldsymbol{y})-f^i(x,\boldsymbol{y}^*)|^2\right)^{1/2}$$
$$=\left(\sum_{i=1}^d\left|\sum_{j=1}^d\frac{\partial f^i}{\partial y^j}(x,\overline{\boldsymbol{y}}_i)[y^j-(y^*)^j]\right|^2\right)^{1/2}\le\left(\sum_{i=1}^d\left[\sum_{j=1}^dM_{ij}|y^j-(y^*)^j|\right]^2\right)^{1/2}$$
$$\le\left(\sum_{i=1}^d\sum_{j=1}^dM_{ij}^2\sum_{j=1}^d|y^j-(y^*)^j|^2\right)^{1/2},$$

so that

$$\|\boldsymbol{f}(x,\boldsymbol{y})-\boldsymbol{f}(x,\boldsymbol{y}^*)\|_2\le L\|\boldsymbol{y}-\boldsymbol{y}^*\|_2,\quad L=\left(\sum_{i=1}^d\sum_{j=1}^dM_{ij}^2\right)^{1/2}=\|\boldsymbol{M}\|_F,$$

where $\|\cdot\|_F$ is the Frobenius matrix norm.

(c) Similarly as before,

$$\|\boldsymbol{f}(x,\boldsymbol{y})-\boldsymbol{f}(x,\boldsymbol{y}^*)\|_\infty=\max_i|f^i(x,\boldsymbol{y})-f^i(x,\boldsymbol{y}^*)|$$
$$=\max_i\left|\sum_{j=1}^d\frac{\partial f^i}{\partial y^j}(x,\overline{\boldsymbol{y}}_i)[y^j-(y^*)^j]\right|\le\max_i\sum_{j=1}^dM_{ij}|y^j-(y^*)^j|$$
$$\le\left(\max_i\sum_{j=1}^dM_{ij}\right)\max_j|y^j-(y^*)^j|,$$

hence

$$\|\boldsymbol{f}(x,\boldsymbol{y})-\boldsymbol{f}(x,\boldsymbol{y}^*)\|_\infty\le L\|\boldsymbol{y}-\boldsymbol{y}^*\|_\infty,\quad L=\|\boldsymbol{M}\|_\infty,$$

where $\|\cdot\|_\infty$ on the far right is the ∞-matrix norm.

5. (a) We let $\boldsymbol{y}^T=[y_1,y_2,\ldots,y_5]\in\mathbb{R}^5$, where

$$y_1=u,\quad y_2=u',\quad y_3=u'',\quad y_4=v,\quad y_5=v'.$$

Then

$$\begin{aligned}
\frac{\mathrm{d}y_1}{\mathrm{d}x} &= y_2,\\
\frac{\mathrm{d}y_2}{\mathrm{d}x} &= y_3,\\
\frac{\mathrm{d}y_3}{\mathrm{d}x} &= x^2y_1y_3 - y_1y_5,\\
\frac{\mathrm{d}y_4}{\mathrm{d}x} &= y_5,\\
\frac{\mathrm{d}y_5}{\mathrm{d}x} &= xy_4y_5 + 4y_2.
\end{aligned}$$

(b) We have

$$\boldsymbol{f}_{\boldsymbol{y}}(x,\boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ x^2y_3 - y_5 & 0 & x^2y_1 & 0 & -y_1\\ 0 & 0 & 0 & 0 & 1\\ 0 & 4 & 0 & xy_5 & xy_4 \end{bmatrix}.$$

(c) By the result in (b), we have

$$\left|\frac{\partial f_i}{\partial y_j}\right| \le m_{ij}(\boldsymbol{y}), \quad \boldsymbol{m}(\boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ |y_3| + |y_5| & 0 & |y_1| & 0 & |y_1|\\ 0 & 0 & 0 & 0 & 1\\ 0 & 4 & 0 & |y_5| & |y_4| \end{bmatrix}.$$

On $[0,1]\times\mathcal{D}$, therefore,

$$\boldsymbol{m}(\boldsymbol{y}) \le \begin{bmatrix} 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 1 & 0 & 1 & 0 & 1\\ 0 & 0 & 0 & 0 & 1\\ 0 & 4 & 0 & 1 & 1 \end{bmatrix} =: \boldsymbol{M},$$

since $\|\boldsymbol{y}\|_1 \le 1$. According to the answers to Ex. 4, valid Lipschitz constants L in the ℓ_1, ℓ_2, and ℓ_∞ norm are respectively $L = \|\boldsymbol{M}\|_1$, $L = \|\boldsymbol{M}\|_F$, and $L = \|\boldsymbol{M}\|_\infty$, that is, in the present case, $L = 5$, $L = \sqrt{24} = 4.89897\ldots$, $L = 6$.

6. (a) For a general scalar differential equation, we have from (5.64) that

$$\tau(x,y) = \frac{1}{2}\left[\left(\frac{1}{4\alpha_2} - \frac{1}{3}\right)(f_{xx} + 2f_{xy}f + f_{yy}f^2) - \frac{1}{3}\, f_y(f_x + f_yf)\right].$$

In the case at hand, $f(x,y) = y^\lambda$, so that

$$\begin{aligned}
&f_x = f_{xx} = f_{xy} = 0,\\
&f = y^\lambda,\ f_y = \lambda y^{\lambda-1},\ f_{yy} = \lambda(\lambda-1)y^{\lambda-2}.
\end{aligned}$$

There follows

$$\tau(x,y) = \frac{\lambda}{2}\left[\left(\frac{1}{4\alpha_2} - \frac{1}{3}\right)(\lambda - 1) - \frac{1}{3}\lambda\right] y^{3\lambda-2}.$$

(b) The modified Euler method has $\alpha_2 = 1$, hence

$$\tau^{ME}(x,y) = -\frac{\lambda}{24}(5\lambda - 1)y^{3\lambda-2}.$$

Heun's method corresponds to $\alpha_2 = \frac{1}{2}$, so that

$$\tau^{H}(x,y) = -\frac{\lambda}{12}(\lambda + 1)y^{3\lambda-2}.$$

We have $|\tau^{ME}| \le |\tau^{H}|$ if and only if $|5\lambda - 1| \le 2(\lambda + 1)$. Together with $\lambda > 0$, this will be the case precisely if $0 < \lambda \le 1$. If $\lambda > 1$, then $|\tau^{ME}| > |\tau^{H}|$.

(c) For the method to have order $p = 3$, we must have $\tau(x,y) = 0$, which holds precisely if

$$\left(\frac{1}{4\alpha_2} - \frac{1}{3}\right)(\lambda - 1) - \frac{1}{3}\lambda = 0.$$

Clearly, $\lambda \neq 1$. Solving for α_2, we get

$$\alpha_2 = \frac{3(\lambda - 1)}{4(2\lambda - 1)}.$$

If $\lambda > 1$, one has $0 < \alpha_2 < \frac{3}{8}$, so that by (5.63), $\mu = \frac{1}{2\alpha_2} > \frac{4}{3}$, which is excluded. If $\frac{1}{2} < \lambda < 1$, then $\alpha_2 < 0$, hence $\mu < 0$, which is also excluded. In the remaining interval $0 < \lambda < \frac{1}{2}$, we have $\frac{3}{4} < \alpha_2 < \infty$, and $\frac{3}{4} < \alpha_2 < 1$ precisely if $0 < \lambda < \frac{1}{5}$. Since by (5.63), $\alpha_1 = 1 - \alpha_2$, we then have $0 < \alpha_1 < \frac{1}{4}$ and $\frac{1}{2} < \mu = \frac{1}{2\alpha_2} < \frac{2}{3}$, both being acceptable. The λ-interval in question, therefore, is $0 < \lambda < \frac{1}{5}$.

7. (a) The function

$$\boldsymbol{\varphi}(\boldsymbol{u}) := \boldsymbol{y} + h\boldsymbol{f}(x + h, \boldsymbol{u}), \quad \boldsymbol{u} \in \mathbb{R}^d,$$

is contractive on $\mathbb{R}^d$ if $hL < 1$, where L is a uniform Lipschitz constant for $\boldsymbol{f}$. Therefore, if $hL < 1$ there is a unique fixed point, $\boldsymbol{u}^* = \boldsymbol{y}_{\text{next}}$, of $\boldsymbol{\varphi}$ satisfying $\boldsymbol{u}^* = \boldsymbol{\varphi}(\boldsymbol{u}^*)$ (cf. Theorem 4.9.1).

(b) We have

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \boldsymbol{f}(x + h, \boldsymbol{y}_{\text{next}}),$$

where $\boldsymbol{y}_{\text{next}}$ is the solution of (cf. (a))

$$\boldsymbol{y}_{\text{next}} = \boldsymbol{y} + h\boldsymbol{f}(x + h, \boldsymbol{y}_{\text{next}}).$$

Let $\boldsymbol{y}_{\text{next}} = \boldsymbol{y} + \boldsymbol{c}_1 h + \boldsymbol{c}_2 h^2 + \cdots$. Then, using Taylor expansion (cf. (5.58)),

$$\begin{aligned}\boldsymbol{y} + \boldsymbol{c}_1 h + \boldsymbol{c}_2 h^2 + \cdots &= \boldsymbol{y} + h\boldsymbol{f}(x+h, \boldsymbol{y} + \boldsymbol{c}_1 h + \boldsymbol{c}_2 h^2 + \cdots) \\ &= \boldsymbol{y} + h[\boldsymbol{f} + h\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}(\boldsymbol{c}_1 h + O(h^2))] \\ &= \boldsymbol{y} + h\boldsymbol{f} + h^2(\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{c}_1) + O(h^3),\end{aligned}$$

with $\boldsymbol{f}$ and its partial derivatives evaluated at $(x, \boldsymbol{y})$. Comparing coefficients of like powers on the left and right, we get

$$\boldsymbol{c}_1 = \boldsymbol{f}, \quad \boldsymbol{c}_2 = \boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{c}_1 = \boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f}, \quad \ldots,$$

so that

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \frac{1}{h}(\boldsymbol{y}_{\text{next}} - \boldsymbol{y}) = \boldsymbol{c}_1 + \boldsymbol{c}_2 h + \cdots = \boldsymbol{f} + h(\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f}) + O(h^2).$$

On the other hand (cf. (5.60)),

$$\frac{1}{h}[\boldsymbol{u}(x+h) - \boldsymbol{u}(x)] = \boldsymbol{f} + \tfrac{1}{2}h(\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f}) + O(h^2),$$

so that

$$\boldsymbol{T}(x, \boldsymbol{y}; h) = \boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \frac{1}{h}[\boldsymbol{u}(x+h) - \boldsymbol{u}(x)] = \tfrac{1}{2}h(\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f}) + O(h^2).$$

The order is thus $p = 1$, and the principal error function

$$\boldsymbol{\tau}(x, \boldsymbol{y}) = \tfrac{1}{2}(\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f})(x, \boldsymbol{y}),$$

the same, except for the sign, as the principal error function of the explicit Euler method (cf. (5.44)).

8. The truncation error (5.32) in this special case, the reference solution being $u(t) = y + \int_x^t f(\tau)\mathrm{d}\tau$, is

$$\begin{aligned}T(x, y; h) &= \Phi(x, y; h) - \frac{1}{h}[u(x+h) - u(x)] \\ &= \alpha_1 f(x) + \alpha_2 f(x + \mu h) - \frac{1}{h}\int_x^{x+h} f(\tau)\mathrm{d}\tau.\end{aligned}$$

Thus, $-hT(x, y; h)$ is the remainder term $E(f)$ of the two-point quadrature formula

$$\int_x^{x+h} f(\tau)\mathrm{d}\tau = h[\alpha_1 f(x) + \alpha_2 f(x + \mu h)] + E(f).$$

For $f(\tau) \equiv 1$ and $f(\tau) \equiv \tau$, we have that $E(f)$ is respectively a linear and a quadratic function of h. But since $E(f)$ must be $O(h^3)$ (the method having order $p = 2$), it follows by the linearity of the functional E that $E(f) = 0$ for any $f \in \mathbb{P}_1$. Therefore also $T(x, y; h) = 0$ whenever $f \in \mathbb{P}_1$, and the method has zero error.

9. (a) Given $x \in \mathbb{R}$ and $\boldsymbol{y} = [y^1, y^2]^T \in \mathbb{R}^2$, the truncation error is

$$\boldsymbol{T}(x, \boldsymbol{y}; h) = \boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \frac{1}{h}\,[\boldsymbol{u}(x+h) - \boldsymbol{u}(x)],$$

where $\boldsymbol{u}(t)$ is the reference solution, i.e., the solution of the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u^1 \\ u^2 \end{bmatrix} = \begin{bmatrix} u^2 \\ g(t, u^1) \end{bmatrix}, \quad \begin{bmatrix} u^1 \\ u^2 \end{bmatrix}(x) = \begin{bmatrix} y^1 \\ y^2 \end{bmatrix}.$$

It follows from this that

$$(u^1)' = u^2, \quad (u^1)'' = (u^2)' = g(t, u^1),$$
$$(u^2)' = g(t, u^1), \quad (u^2)'' = g_x(t, u^1) + g_z(t, u^1)(u^1)' = g_x(t, u^1) + g_z(t, u^1)u^2,$$

and thus, at $t = x$,

$$(u^1)'(x) = y^2, \quad (u^1)''(x) = g; \qquad (u^2)'(x) = g, \quad (u^2)''(x) = g_x + g_z y^2,$$

where g, g_x, g_z are evaluated at (x, y^1). Now, the two components of $\boldsymbol{T}(x, \boldsymbol{y}; h)$ are

$$T^1(x, \boldsymbol{y}; h) = y^2 + \frac{1}{2}hg(x + \mu h, y^1 + \mu h y^2) - \frac{1}{h}\,[u^1(x+h) - y^1],$$
$$T^2(x, \boldsymbol{y}; h) = g(x + \mu h, y^1 + \mu h y^2) - \frac{1}{h}\,[u^2(x+h) - y^2],$$

which, expanded in powers of h, up to the first power, give

$$\begin{aligned} T^1(x, \boldsymbol{y}; h) &= y^2 + \frac{1}{2}hg - \frac{1}{h}\,[h(u^1)'(x) + \frac{1}{2}h^2(u^1)''(x)] + O(h^2) \\ &= y^2 + \frac{1}{2}hg - y^2 - \frac{1}{2}hg + O(h^2) = O(h^2), \\ T^2(x, \boldsymbol{y}; h) &= g + \mu h(g_x + g_z y^2) - \frac{1}{h}\,[h(u^2)'(x) + \frac{1}{2}h^2(u^2)''(x)] + O(h^2) \\ &= g + \mu h(g_x + g_z y^2) - g - \frac{1}{2}h(g_x + g_z y^2) + O(h^2) \\ &= (\mu - \frac{1}{2})h(g_x + g_z y^2) + O(h^2), \end{aligned}$$

with g and its partial derivatives evaluated at (x, y^1). Thus, we have order $p = 2$, that is, $\boldsymbol{T}(x, \boldsymbol{y}; h) = O(h^2)$ (cf. Definition 5.5.3), precisely if $\mu = \frac{1}{2}$.

(b) To obtain the principal error function, we need to expand $\boldsymbol{T}(x, \boldsymbol{y}; h)$ one step further. This requires third derivatives of $\boldsymbol{u}$,

$$\begin{aligned} (u^1)''' &= g_x + g_z(u^1)' = g_x + g_z u^2, \\ (u^2)''' &= g_{xx} + g_{xz}(u^1)' + [g_{xz} + g_{zz}(u^1)']u^2 + g_z(u^2)' \\ &= g_{xx} + 2g_{xz}u^2 + g_{zz}(u^2)^2 + g_z g, \end{aligned}$$

which, at $t = x$, give

$$(u^1)'''(x) = g_x + g_z y^2,$$
$$(u^2)'''(x) = g_{xx} + 2g_{xz}y^2 + g_{zz}(y^2)^2 + g_z g.$$

Therefore, the additional term needed in the expansion of $\boldsymbol{T}(x, \boldsymbol{y}; h)$ is

$$\tfrac{1}{2}h^2(\mu g_x + \mu g_z y^2) - \tfrac{1}{6}h^2(g_x + g_z y^2) = \tfrac{1}{6}h^2(3\mu - 1)(g_x + g_z y^2)$$

for T^1, and

$$\begin{aligned}\tfrac{1}{2}(\mu h)^2[g_{xx} + 2g_{xz}y^2 + g_{zz}(y^2)^2] - \tfrac{1}{6}h^2[g_{xx} + 2g_{xz}y^2 + g_{zz}(y^2)^2 + g_z g] \\ = \tfrac{1}{6}h^2\{(3\mu^2 - 1)[g_{xx} + 2g_{xz}y^2 + g_{zz}(y^2)^2] - g_z g\}\end{aligned}$$

for T^2. When $\mu = \frac{1}{2}$, the principal error function, therefore, is

$$\boldsymbol{\tau}(x, \boldsymbol{y}) = \begin{bmatrix} \frac{1}{12}(g_x + g_z y^2) \\ -\frac{1}{24}[g_{xx} + 2g_{xz}y^2 + g_{zz}(y^2)^2] - \frac{1}{6}g_z g \end{bmatrix},$$

or, in terms of the reference solution u^1,

$$\boldsymbol{\tau}(x, \boldsymbol{y}) = \begin{bmatrix} \frac{1}{12}(u^1)'''(x) \\ -\frac{1}{24}(u^1)^{(4)}(x) - \frac{1}{8}g_z(x, u^1(x))(u^1)''(x) \end{bmatrix},$$

since $(u^2)''' = (u^1)^{(4)}$.

10. Taylor expansion gives

$$\begin{aligned}\boldsymbol{k}_s(x, \boldsymbol{y}; h) &= \boldsymbol{f}(x + \mu_s h, \boldsymbol{y} + h\sum_{j=1}^{s-1} \lambda_{sj}\boldsymbol{k}_j) \\ &= \boldsymbol{f} + h\mu_s \boldsymbol{f}_x + h\boldsymbol{f}_{\boldsymbol{y}} \sum_{j=1}^{s-1} \lambda_{sj}\boldsymbol{k}_j(x, \boldsymbol{y}; 0) + O(h^2),\end{aligned}$$

where $\boldsymbol{f}$, $\boldsymbol{f}_x$, $\boldsymbol{f}_{\boldsymbol{y}}$ are evaluated at $(x, \boldsymbol{y})$. But $\boldsymbol{k}_j(x, \boldsymbol{y}; 0) = \boldsymbol{f}(x, \boldsymbol{y})$, and so

$$\boldsymbol{k}_s(x, \boldsymbol{y}; h) = \boldsymbol{f} + h\left[\mu_s \boldsymbol{f}_x + \left(\sum_{j=1}^{s-1} \lambda_{sj}\right)\boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f}\right] + O(h^2).$$

On the other hand,

$$\begin{aligned}\boldsymbol{u}'(x + \mu_s h) &= \boldsymbol{f}(x + \mu_s h, \boldsymbol{u}(x + \mu_s h)) \\ &= \boldsymbol{f}(x + \mu_s h, \boldsymbol{u}(x) + \mu_s h\boldsymbol{u}'(x) + O(h^2)),\end{aligned}$$

and, since $\boldsymbol{u}(x) = \boldsymbol{y}$, $\boldsymbol{u}'(x) = \boldsymbol{f}(x, \boldsymbol{y})$, we get

$$\boldsymbol{u}'(x + \mu_s h) = \boldsymbol{f} + h\mu_s(\boldsymbol{f}_x + \boldsymbol{f}_{\boldsymbol{y}}\boldsymbol{f}) + O(h^2).$$

Comparison with the expansion of $\boldsymbol{k}_s$ shows that $\boldsymbol{k}_s(x, \boldsymbol{y}; h) = \boldsymbol{u}'(x+\mu_s h) + O(h^2)$ is equivalent to $\mu_s = \sum_{j=1}^{s-1} \lambda_{sj}$.

11. We have for the truncation error of the method $\boldsymbol{\Phi}$,

$$\boldsymbol{T}(x, \boldsymbol{y}; h) = \boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \frac{1}{h}\,[\boldsymbol{u}(x + h) - \boldsymbol{u}(x)],$$

where $\boldsymbol{u}$ is the reference solution assumed to be in $C^{\mu+1}[x, x + h]$. Then, applying the given quadrature rule to $\boldsymbol{u}'$, we get

$$\boldsymbol{T}(x, \boldsymbol{y}; h) = \sum_{k=1}^{\nu} w_k \boldsymbol{f}(x + \vartheta_k h, \boldsymbol{y} + \vartheta_k h \overline{\boldsymbol{\Phi}}_k(x, \boldsymbol{y}; \vartheta_k h)) - \frac{1}{h}\int_x^{x+h} \boldsymbol{u}'(t)\mathrm{d}t$$
$$= \sum_{k=1}^{\nu} w_k[\boldsymbol{f}(x + \vartheta_k h, \boldsymbol{y} + \vartheta_k h \overline{\boldsymbol{\Phi}}_k(x, \boldsymbol{y}; \vartheta_k h)) - \boldsymbol{f}(x + \vartheta_k h, \boldsymbol{u}(x + \vartheta_k h)] - ch^{\mu}\boldsymbol{u}^{(\mu+1)}(\xi),$$

and therefore, on applying the Lipschitz condition,

$$\|\boldsymbol{T}(x, \boldsymbol{y}; h)\| \le L \sum_{k=1}^{\nu} |w_k| \cdot \|\boldsymbol{y} + \vartheta_k h \overline{\boldsymbol{\Phi}}_k(x, \boldsymbol{y}; \vartheta_k h) - \boldsymbol{u}(x + \vartheta_k h)\| + O(h^{\mu}).$$

But, by assumption, $\overline{\boldsymbol{\Phi}}_k$ has order $\overline{p}_k$, so that

$$\overline{\boldsymbol{\Phi}}_k(x, \boldsymbol{y}; \vartheta_k h) - \frac{1}{\vartheta_k h}\,[\boldsymbol{u}(x + \vartheta_k h) - \boldsymbol{u}(x)] = O(h^{\overline{p}_k}).$$

Therefore, since $\boldsymbol{u}(x) = \boldsymbol{y}$,

$$\|\boldsymbol{y} + \vartheta_k h \overline{\boldsymbol{\Phi}}_k(x, \boldsymbol{y}; \vartheta_k h) - \boldsymbol{u}(x + \vartheta_k h)\| = O(h^{\overline{p}_k+1}),$$

and the assertion follows, since $O(h^{\overline{p}_k+1}) = O(h^{\overline{p}+1})$.

12. Since $\boldsymbol{g}(x, \boldsymbol{y}) = \boldsymbol{f}^{[1]}(x, \boldsymbol{y})$, the truncation error of the method is given by

$$\boldsymbol{T}(x, \boldsymbol{y}; h) = \boldsymbol{f}(x, \boldsymbol{y}) + \tfrac{1}{2}h\boldsymbol{f}^{[1]}(x + \tfrac{1}{3}h, \boldsymbol{y} + \tfrac{1}{3}h\boldsymbol{f}(x, \boldsymbol{y})) - \frac{1}{h}\,[\boldsymbol{u}(x + h) - \boldsymbol{u}(x)].$$

Expanding it in powers of h yields (with $\boldsymbol{f}$, $\boldsymbol{f}^{[1]}$, $\boldsymbol{f}_x^{[1]}$, etc., evaluated at

$(x, \boldsymbol{y})$)

$$\begin{aligned}
\boldsymbol{T} &= \boldsymbol{f} + \tfrac{1}{2}h[\boldsymbol{f}^{[1]} + \tfrac{1}{3}h(\boldsymbol{f}_x^{[1]} + \boldsymbol{f}_y^{[1]}\boldsymbol{f}) \\
&\qquad + \tfrac{1}{2}(\tfrac{1}{3}h)^2(\boldsymbol{f}_{xx}^{[1]} + 2\boldsymbol{f}_{x\boldsymbol{y}}^{[1]}\boldsymbol{f} + \boldsymbol{f}^T\boldsymbol{f}_{\boldsymbol{yy}}^{[1]}\boldsymbol{f}) + O(h^3)] \\
&\qquad - \frac{1}{h}\left[h\boldsymbol{u}'(x) + \tfrac{1}{2}h^2\boldsymbol{u}''(x) + \tfrac{1}{6}h^3\boldsymbol{u}'''(x) + \tfrac{1}{24}h^4\boldsymbol{u}^{(4)}(x) + O(h^5)\right] \\
&= \boldsymbol{f} + \tfrac{1}{2}h\boldsymbol{f}^{[1]} + \tfrac{1}{6}h^2\boldsymbol{f}^{[2]} + \tfrac{1}{36}h^3(\boldsymbol{f}_{xx}^{[1]} + 2\boldsymbol{f}_{x\boldsymbol{y}}^{[1]}\boldsymbol{f} + \boldsymbol{f}^T\boldsymbol{f}_{\boldsymbol{yy}}^{[1]}\boldsymbol{f}) \\
&\qquad - [\boldsymbol{f} + \tfrac{1}{2}h\boldsymbol{f}^{[1]} + \tfrac{1}{6}h^2\boldsymbol{f}^{[2]} + \tfrac{1}{24}h^3\boldsymbol{f}^{[3]}] + O(h^4) \\
&= \tfrac{1}{72}h^3[2(\boldsymbol{f}_{xx}^{[1]} + 2\boldsymbol{f}_{x\boldsymbol{y}}^{[1]}\boldsymbol{f} + \boldsymbol{f}^T\boldsymbol{f}_{\boldsymbol{yy}}^{[1]}\boldsymbol{f}) - 3\boldsymbol{f}^{[3]}] + O(h^4).
\end{aligned}$$

This shows that the method has order $p = 3$. The principal error function is the coefficient of h^3 in the last line of the formula above. It can be expressed in terms of $\boldsymbol{f}^{[1]} = \boldsymbol{g}$ by noting that

$$\begin{aligned}
\boldsymbol{f}^{[2]} &= \boldsymbol{f}_x^{[1]} + \boldsymbol{f}_{\boldsymbol{y}}^{[1]}\boldsymbol{f}, \\
\boldsymbol{f}^{[3]} &= \boldsymbol{f}_x^{[2]} + \boldsymbol{f}_{\boldsymbol{y}}^{[2]}\boldsymbol{f} \\
&= \boldsymbol{f}_{xx}^{[1]} + 2\boldsymbol{f}_{x\boldsymbol{y}}^{[1]}\boldsymbol{f} + \boldsymbol{f}_{\boldsymbol{y}}^{[1]}\boldsymbol{f}_x + (\boldsymbol{f}_{\boldsymbol{y}}^{[1]}\boldsymbol{f})_{\boldsymbol{y}}\boldsymbol{f}.
\end{aligned}$$

Thus,

$$\boldsymbol{\tau}(x, \boldsymbol{y}) = \tfrac{1}{72}\left[2\boldsymbol{f}^T\boldsymbol{g}_{\boldsymbol{yy}}\boldsymbol{f} - (\boldsymbol{g}_{xx} + 2\boldsymbol{g}_{x\boldsymbol{y}}\boldsymbol{f}) - 3(\boldsymbol{g}_{\boldsymbol{y}}\boldsymbol{f}_x + (\boldsymbol{g}_{\boldsymbol{y}}\boldsymbol{f})_{\boldsymbol{y}}\boldsymbol{f})\right].$$

13. Let $\boldsymbol{k}_i = \boldsymbol{k}_i(x, \boldsymbol{y}; h)$, $\boldsymbol{k}_i^* = \boldsymbol{k}_i(x, \boldsymbol{y}^*; h)$.

(a) For the second-order Runge–Kutta method, we have

$$\begin{aligned}
\boldsymbol{k}_1 - \boldsymbol{k}_1^* &= \boldsymbol{f}(x, \boldsymbol{y}) - \boldsymbol{f}(x, \boldsymbol{y}^*), \\
\boldsymbol{k}_2 - \boldsymbol{k}_2^* &= \boldsymbol{f}(x + h, \boldsymbol{y} + h\boldsymbol{k}_1) - \boldsymbol{f}(x + h, \boldsymbol{y}^* + h\boldsymbol{k}_1^*),
\end{aligned}$$

and

$$\begin{aligned}
\|\boldsymbol{k}_1 - \boldsymbol{k}_1^*\| &\le L\|\boldsymbol{y} - \boldsymbol{y}^*\|, \\
\|\boldsymbol{k}_2 - \boldsymbol{k}_2^*\| &\le L\|\boldsymbol{y} - \boldsymbol{y}^* + h(\boldsymbol{k}_1 - \boldsymbol{k}_1^*)\| \\
&\le L(1 + hL)\|\boldsymbol{y} - \boldsymbol{y}^*\|.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\|\boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \boldsymbol{\Phi}(x, \boldsymbol{y}^*; h)\| &= \tfrac{1}{2}\|(\boldsymbol{k}_1 - \boldsymbol{k}_1^*) + (\boldsymbol{k}_2 - \boldsymbol{k}_2^*)\| \\
&\le \tfrac{1}{2}(L + L(1 + hL))\|\boldsymbol{y} - \boldsymbol{y}^*\| \\
&= L(1 + \tfrac{1}{2}hL)\|\boldsymbol{y} - \boldsymbol{y}^*\|,
\end{aligned}$$

so that

$$M = L(1 + \tfrac{1}{2}hL).$$

(b) For the classical Runge–Kutta method, we have

$$\begin{aligned}
\boldsymbol{k}_1 - \boldsymbol{k}_1^* &= \boldsymbol{f}(x, \boldsymbol{y}) - \boldsymbol{f}(x, \boldsymbol{y}^*),\\
\boldsymbol{k}_2 - \boldsymbol{k}_2^* &= \boldsymbol{f}(x + \tfrac{1}{2}h, \boldsymbol{y} + \tfrac{1}{2}h\boldsymbol{k}_1) - \boldsymbol{f}(x + \tfrac{1}{2}h, \boldsymbol{y}^* + \tfrac{1}{2}h\boldsymbol{k}_1^*),\\
\boldsymbol{k}_3 - \boldsymbol{k}_3^* &= \boldsymbol{f}(x + \tfrac{1}{2}h, \boldsymbol{y} + \tfrac{1}{2}h\boldsymbol{k}_2) - \boldsymbol{f}(x + \tfrac{1}{2}h, \boldsymbol{y}^* + \tfrac{1}{2}h\boldsymbol{k}_2^*),\\
\boldsymbol{k}_4 - \boldsymbol{k}_4^* &= \boldsymbol{f}(x + h, \boldsymbol{y} + h\boldsymbol{k}_3) - \boldsymbol{f}(x + h, \boldsymbol{y}^* + h\boldsymbol{k}_3^*),
\end{aligned}$$

from which

$$\begin{aligned}
\|\boldsymbol{k}_1 - \boldsymbol{k}_1^*\| &\le L\|\boldsymbol{y} - \boldsymbol{y}^*\|,\\
\|\boldsymbol{k}_2 - \boldsymbol{k}_2^*\| &\le L\|\boldsymbol{y} - \boldsymbol{y}^* + \tfrac{1}{2}h(\boldsymbol{k}_1 - \boldsymbol{k}_1^*)\|\\
&\le L(1 + \tfrac{1}{2}hL)\|\boldsymbol{y} - \boldsymbol{y}^*\|,\\
\|\boldsymbol{k}_3 - \boldsymbol{k}_3^*\| &\le L\|\boldsymbol{y} - \boldsymbol{y}^* + \tfrac{1}{2}h(\boldsymbol{k}_2 - \boldsymbol{k}_2^*)\|\\
&\le L(1 + \tfrac{1}{2}hL + \tfrac{1}{4}h^2L^2)\|\boldsymbol{y} - \boldsymbol{y}^*\|,\\
\|\boldsymbol{k}_4 - \boldsymbol{k}_4^*\| &\le L\|\boldsymbol{y} - \boldsymbol{y}^* + h(\boldsymbol{k}_3 - \boldsymbol{k}_3^*)\|\\
&\le L(1 + hL + \tfrac{1}{2}h^2L^2 + \tfrac{1}{4}h^3L^3)\|\boldsymbol{y} - \boldsymbol{y}^*\|.
\end{aligned}$$

Since

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \boldsymbol{\Phi}(x, \boldsymbol{y}^*; h) = \tfrac{1}{6}\left[(\boldsymbol{k}_1 - \boldsymbol{k}_1^*) + 2(\boldsymbol{k}_2 - \boldsymbol{k}_2^*) + 2(\boldsymbol{k}_3 - \boldsymbol{k}_3^*) + (\boldsymbol{k}_4 - \boldsymbol{k}_4^*)\right],$$

we get

$$\begin{aligned}
&\|\boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \boldsymbol{\Phi}(x, \boldsymbol{y}^*; h)\|\\
&\quad\le \tfrac{1}{6}L[1 + (2 + hL) + (2 + hL + \tfrac{1}{2}h^2L^2) + (1 + hL + \tfrac{1}{2}h^2L^2 + \tfrac{1}{4}h^3L^3)]\|\boldsymbol{y} - \boldsymbol{y}^*\|\\
&\quad= L(1 + \tfrac{1}{2}hL + \tfrac{1}{6}h^2L^2 + \tfrac{1}{24}h^3L^3)\|\boldsymbol{y} - \boldsymbol{y}^*\|,
\end{aligned}$$

and thus

$$M = L(1 + \tfrac{1}{2}hL + \tfrac{1}{6}h^2L^2 + \tfrac{1}{24}h^3L^3).$$

(c) For the implicit r-stage Runge–Kutta method, we have

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \boldsymbol{\Phi}(x, \boldsymbol{y}^*; h) = \sum_{s=1}^{r} \alpha_s(\boldsymbol{k}_s - \boldsymbol{k}_s^*),$$

where

$$\boldsymbol{k}_s = \boldsymbol{f}(x + \mu_s h, \boldsymbol{y} + h\sum_{j=1}^{r} \lambda_{sj}\boldsymbol{k}_j), \quad \boldsymbol{k}_s^* = \boldsymbol{f}(x + \mu_s h, \boldsymbol{y}^* + h\sum_{j=1}^{r} \lambda_{sj}\boldsymbol{k}_j^*).$$

Therefore,

$$\begin{aligned}\|\boldsymbol{k}_s - \boldsymbol{k}_s^*\| &\le L\|\boldsymbol{y} - \boldsymbol{y}^* + h\sum_{j=1}^{r}\lambda_{sj}(\boldsymbol{k}_j - \boldsymbol{k}_j^*)\| \\ &\le L\|\boldsymbol{y} - \boldsymbol{y}^*\| + \lambda L h\sum_{j=1}^{r}\|\boldsymbol{k}_j - \boldsymbol{k}_j^*\|,\end{aligned}$$

where $\lambda = \max_{s,j}|\lambda_{sj}|$. Summing over s gives

$$\sum_{s=1}^{r}\|\boldsymbol{k}_s - \boldsymbol{k}_s^*\| \le rL\|\boldsymbol{y} - \boldsymbol{y}^*\| + r\lambda L h\sum_{j=1}^{r}\|\boldsymbol{k}_j - \boldsymbol{k}_j^*\|.$$

Assuming h small enough to have $r\lambda L h < \frac{1}{2}$, hence $(1 - r\lambda L h)^{-1} < 2$, we get

$$\sum_{s=1}^{r}\|\boldsymbol{k}_s - \boldsymbol{k}_s^*\| \le 2rL\|\boldsymbol{y} - \boldsymbol{y}^*\|.$$

Therefore, letting $\alpha = \max_s|\alpha_s|$, we obtain

$$\|\boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \boldsymbol{\Phi}(x, \boldsymbol{y}^*; h)\| \le \alpha\sum_{s=1}^{r}\|\boldsymbol{k}_s - \boldsymbol{k}_s^*\| \le 2r\alpha L\|\boldsymbol{y} - \boldsymbol{y}^*\|,$$

so that

$$M = 2r\alpha L \quad (\text{if } r\lambda L h < \tfrac{1}{2}).$$

14. The system to be solved is

$$\boldsymbol{F}_s(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_r) := \boldsymbol{k}_s - \boldsymbol{f}(x + \mu_s h, \boldsymbol{y} + h\sum_{j=1}^{r}\lambda_{sj}\boldsymbol{k}_j) = \boldsymbol{0}, \quad s = 1, \ldots, r.$$

Let

$$\boldsymbol{k} = \begin{bmatrix}\boldsymbol{k}_1 \\ \boldsymbol{k}_2 \\ \vdots \\ \boldsymbol{k}_r\end{bmatrix}, \quad \boldsymbol{F}(\boldsymbol{k}) = \begin{bmatrix}\boldsymbol{F}_1(\boldsymbol{k}) \\ \boldsymbol{F}_2(\boldsymbol{k}) \\ \vdots \\ \boldsymbol{F}_r(\boldsymbol{k})\end{bmatrix}.$$

The Jacobian matrix of the system is

$$\frac{\partial \boldsymbol{F}(\boldsymbol{k})}{\partial \boldsymbol{k}} = \left[\frac{\partial \boldsymbol{F}_s}{\partial \boldsymbol{k}_t}\right] = \begin{bmatrix} \boldsymbol{I} - h\lambda_{11}\boldsymbol{f}_{\boldsymbol{y}} & -h\lambda_{12}\boldsymbol{f}_{\boldsymbol{y}} & \cdots & -h\lambda_{1r}\boldsymbol{f}_{\boldsymbol{y}} \\ -h\lambda_{21}\boldsymbol{f}_{\boldsymbol{y}} & \boldsymbol{I} - h\lambda_{22}\boldsymbol{f}_{\boldsymbol{y}} & \cdots & -h\lambda_{2r}\boldsymbol{f}_{\boldsymbol{y}} \\ \cdots & \cdots & \cdots & \cdots \\ -h\lambda_{r1}\boldsymbol{f}_{\boldsymbol{y}} & -h\lambda_{r2}\boldsymbol{f}_{\boldsymbol{y}} & \cdots & \boldsymbol{I} - h\lambda_{rr}\boldsymbol{f}_{\boldsymbol{y}} \end{bmatrix},$$

where the Jacobian $\boldsymbol{f_y}$ in the (s,t)-block of this matrix is to be evaluated at $(x+\mu_s h, \boldsymbol{y}+h\sum_{j=1}^{r}\lambda_{sj}\boldsymbol{k}_j)$. Newton's method consists in the iteration

$$\boldsymbol{k}^{[i+1]} = \boldsymbol{k}^{[i]} + \boldsymbol{\Delta}^{[i]}, \quad i = 0, 1, 2, \dots ,$$

where $\boldsymbol{\Delta}^{[i]}$ is the solution of the linear system

$$\frac{\partial \boldsymbol{F}(\boldsymbol{k}^{[i]})}{\partial \boldsymbol{k}}\,\boldsymbol{\Delta}^{[i]} = -\boldsymbol{F}(\boldsymbol{k}^{[i]}).$$

As initial approximation one can take the solution for $h=0$, that is,

$$\boldsymbol{k}^{[0]} = \begin{bmatrix} \boldsymbol{f}(x,\boldsymbol{y}) \\ \boldsymbol{f}(x,\boldsymbol{y}) \\ \vdots \\ \boldsymbol{f}(x,\boldsymbol{y}) \end{bmatrix} \in \mathbb{R}^{rd}.$$

15. See the text.

16. We have

$$y_{\text{next}} = \varphi(h\lambda)y,$$

where (cf. Ex. 18)

$$\varphi(z) = 1 + z + \tfrac{1}{2}z^2 + \tfrac{1}{6}z^3 + \tfrac{1}{24}z^4.$$

Therefore,

$$T(x,y;h) = \frac{1}{h}\left[y_{\text{next}} - u(x+h)\right] = \frac{1}{h}\left[\varphi(h\lambda)y - u(x+h)\right],$$

where $u(t)$ is the reference solution, i.e., the solution of

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \lambda u, \quad u(x) = y.$$

Clearly, $u(t) = ye^{\lambda(t-x)}$, so that

$$\begin{aligned} T(x,y;h) &= \frac{1}{h}\left[1 + \lambda h + \frac{1}{2}(\lambda h)^2 + \frac{1}{6}(\lambda h)^3 + \frac{1}{24}(\lambda h)^4 - e^{\lambda h}\right]y \\ &= -\frac{\lambda^5}{120}h^4 y + O(h^5). \end{aligned}$$

The principal error function thus is

$$\tau(x,y) = -\frac{\lambda^5}{120}\,y,$$

and the variational differential equation

$$\frac{\mathrm{d}e}{\mathrm{d}x} = \lambda e - \frac{\lambda^5}{120}\,\mathrm{e}^{\lambda x}, \quad 0 \le x \le 1; \quad e(0) = 0.$$

This is readily solved, giving

$$e(x) = -\frac{\lambda^5}{120} x \mathrm{e}^{\lambda x}.$$

Therefore, by (5.104),

$$u_N - y(1) = e(1)h^4 + O(h^5) = -\frac{\lambda^5 h^4}{120}\mathrm{e}^{\lambda} + O(h^5),$$

and, in particular, since $y(1) = \mathrm{e}^{\lambda}$,

$$\lim_{h \to 0} h^{-4} \, \frac{u_N - y(1)}{y(1)} = -\frac{\lambda^5}{120}\,.$$

17. (a) Every solution has the form $y(x) = y_0 \mathrm{e}^{\lambda x}$. Here, $|\mathrm{e}^{\lambda x}| = \mathrm{e}^{(\mathrm{Re}\,\lambda) x}$, which, if $\mathrm{Re}\,\lambda < 0$, tends to zero as $x \to \infty$.

(b) At a generic point (x, y), the Runge–Kutta method computes

$$\begin{aligned}
k_1 &= \lambda y, \\
k_2 &= \lambda(y + \tfrac{1}{2} h k_1) = \lambda(1 + \tfrac{1}{2} h\lambda) y, \\
k_3 &= \lambda(y + \tfrac{1}{2} h k_2) = \lambda(1 + \tfrac{1}{2} h\lambda + \tfrac{1}{4}(h\lambda)^2) y, \\
k_4 &= \lambda(y + h k_3) = \lambda(1 + h\lambda + \tfrac{1}{2}(h\lambda)^2 + \tfrac{1}{4}(h\lambda)^3) y,
\end{aligned}$$

in terms of which

$$\begin{aligned}
y_{\mathrm{next}} &= y + \frac{h}{6}\,(k_1 + 2k_2 + 2k_3 + k_4) \\
&= [1 + h\lambda + \tfrac{1}{2}(h\lambda)^2 + \tfrac{1}{6}(h\lambda)^3 + \tfrac{1}{24}(h\lambda)^4] y.
\end{aligned}$$

Thus, in the notation of (5.140),

$$y_{\mathrm{next}} = \varphi(h\lambda) y, \qquad \varphi(z) = 1 + z + \tfrac{1}{2!} z^2 + \tfrac{1}{3!} z^3 + \tfrac{1}{4!} z^4.$$

In particular,

$$u_{n+1} = \varphi(h\lambda) u_n, \qquad n = 0, 1, 2, \dots\,,$$

hence,

$$u_n = [\varphi(h\lambda)]^n u_0.$$

We therefore have $u_n \to 0$ if and only if $|\varphi(h\lambda)| < 1$. Thus, $h\lambda$ has to lie in the region $\mathcal{D}$ of the complex plane in which $|\varphi(z)| < 1$, $\mathrm{Re}\, z < 0$ (cf. Fig. 5.4 where $p = 4$). If λ, hence $z = h\lambda$, is real, then $|\varphi(z)| < 1$ holds precisely if

$$-1 < 1 + z + \tfrac{1}{2} z^2 + \tfrac{1}{6} z^3 + \tfrac{1}{24} z^4 < 1.$$

The inequality on the left is

$$2 + z + \tfrac{1}{2}z^2 + \tfrac{1}{6}z^3 + \tfrac{1}{24}z^4 > 0,$$

which is true for any real z, since the polynomial on the left has only complex zeros (which are easily obtained, with Matlab for example). The inequality on the right is

$$z(1 + \tfrac{1}{2}z + \tfrac{1}{6}z^2 + \tfrac{1}{24}z^3) < 0,$$

which, for $z < 0$, amounts to

$$1 + \tfrac{1}{2}z + \tfrac{1}{6}z^2 + \tfrac{1}{24}z^3 > 0 \quad (z < 0).$$

Here, the left-hand side increases monotonically from $-\infty$ to 1 when z increases from $-\infty$ to 0. Hence, there is a unique root $\alpha < 0$, and we have

$$|\varphi(h\lambda)| < 1 \text{ if and only if } \alpha < h\lambda < 0.$$

One computes (using Matlab, for example) that $\alpha = -2.78529356\ldots$. Thus, the region $\mathcal{D}$ intersects the real axis at the point α.

(c) For Euler's method, one has

$$y_{\text{next}} = (1 + h\lambda)y,$$

giving

$$u_n = [\varphi(h\lambda)]^n u_0, \qquad \varphi(z) = 1 + z.$$

Clearly, $|\varphi(z)| < 1$ and $\operatorname{Re} z < 0$ precisely if z lies in the (open) disk of radius 1 centered at -1. For real λ, the condition, therefore, is

$$-2 < h\lambda < 0.$$

(d) For the system $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}$, one obtains

$$\boldsymbol{y}_{\text{next}} = \varphi(h\boldsymbol{A})\boldsymbol{y}, \qquad \boldsymbol{u}_n = [\varphi(h\boldsymbol{A})]^n \boldsymbol{u}_0,$$

and $\|\boldsymbol{u}_n\| \to 0$ if and only if the spectral radius of $\varphi(h\boldsymbol{A})$ is strictly less than 1. This means that

$$|\varphi(h\lambda_i)| < 1 \quad \text{for all eigenvalues } \lambda_i \text{ of } \boldsymbol{A}.$$

Thus, each $h\lambda_i$ has to lie in the region $\mathcal{D}$ defined above. If all eigenvalues λ_i are real, then, from (b) and (c), the condition on h is that for each i we must have

$$-2.78529356\ldots < h\lambda_i < 0 \text{ (Runge} - \text{Kutta)}, \quad -2 < h\lambda_i < 0 \text{ (Euler)}.$$

18. Since $\mathrm{e}^z = \varphi(z) + O(z^{p+1}),\ z \to 0$, for any method of order p, the assertion follows at once, if φ is a polynomial, as assumed. It is clear that every explicit p-stage Runge–Kutta method, when applied to the model problem, produces a polynomial φ of degree p, and the order can be made equal to p exactly when $1 \le p \le 4$ (cf. (5.69), (5.70)). For the Taylor expansion method of order p, the identity $\chi \equiv 0$ is obvious.

19. (a) For Euler's method,

$$\boldsymbol{y}_{\text{next}} = \boldsymbol{y} + h\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \boldsymbol{y} + h\boldsymbol{A}\boldsymbol{y} = (\boldsymbol{I} + h\boldsymbol{A})\boldsymbol{y} = \varphi(h\boldsymbol{A})\boldsymbol{y},$$

where $\varphi(z) = 1 + z$. The truncation error is

$$\boldsymbol{T}(x, \boldsymbol{y}; h) = \frac{1}{h}\left[\boldsymbol{y}_{\text{next}} - \boldsymbol{u}(x+h)\right]$$

with the reference solution $\boldsymbol{u}(t)$ being given by

$$\boldsymbol{u}(t) = \mathrm{e}^{(t-x)\boldsymbol{A}}\,\boldsymbol{y} = \left(\boldsymbol{I} + (t-x)\boldsymbol{A} + \tfrac{1}{2!}(t-x)^2\boldsymbol{A}^2 + \cdots\right)\boldsymbol{y}.$$

Therefore,

$$\begin{aligned}\boldsymbol{T}(x, \boldsymbol{y}; h) &= \frac{1}{h}\left[\boldsymbol{I} + h\boldsymbol{A} - (\boldsymbol{I} + h\boldsymbol{A} + \tfrac{1}{2!}h^2\boldsymbol{A}^2 + \cdots)\right]\boldsymbol{y} \\ &= -\left(\tfrac{1}{2}\,h\boldsymbol{A}^2 + \tfrac{1}{6}h^2\boldsymbol{A}^3 + \cdots\right)\boldsymbol{y} \\ &= -\tfrac{1}{2}\,h\boldsymbol{A}^2\boldsymbol{y} + O(h^2),\end{aligned}$$

so that the principal error function is $\boldsymbol{\tau}(x, \boldsymbol{y}) = -\frac{1}{2}\boldsymbol{A}^2\boldsymbol{y}$.

(b) For the classical 4th-order Runge–Kutta method, one gets similarly (cf. Ex. 17(b))

$$\varphi(z) = 1 + z + \tfrac{1}{2!}z^2 + \tfrac{1}{3!}z^3 + \tfrac{1}{4!}z^4$$

and

$$\boldsymbol{\tau}(x, \boldsymbol{y}) = -\tfrac{1}{5!}\,\boldsymbol{A}^5\boldsymbol{y}.$$

20. The explicit Euler method applied to the given model problem amounts to

$$u_{n+1} = u_n + ha_n(u_n - b_n), \quad u_0 = y_0,$$

where $a_n = a(x_n)$, $b_n = b(x_n)$. Thus,

$$u_{n+1} = (1 + ha_n)u_n - ha_nb_n, \quad u_0 = y_0 \quad \text{(explicit Euler)}.$$

Similarly, we have for the implicit Euler method

$$u^*_{n+1} = u^*_n + ha_{n+1}(u^*_{n+1} - b_{n+1}), \quad u^*_0 = y_0,$$

that is, since $1 - ha_{n+1} > 0$ by the negativity of a_{n+1},

$$u^*_{n+1} = \frac{1}{1 - ha_{n+1}}(u^*_n - ha_{n+1}b_{n+1}), \quad u^*_0 = y_0 \quad \text{(implicit Euler)}.$$

Assume $|b(x)| \le B$ on $\mathbb{R}_+$. Consider first the explicit Euler method.

Since we assumed $ah > 1$, we have also $Ah > 1$, and from $-A \le a_n \le -a$ there follows $1 - Ah \le 1 + ha_n \le 1 - ah < 0$, hence $|1 + ha_n| \le Ah - 1$. Thus,

$$|u_{n+1}| \le (Ah - 1)|u_n| + hAB, \ n = 0, 1, 2, \dots, \quad u_0 = y_0.$$

As in the proof of Lemma 5.7.1, we have $|u_n| \le E_n$, where

$$E_{n+1} = (Ah - 1)E_n + hAB, \quad E_0 = |y_0|.$$

A simple calculation gives

$$E_n = \left(|y_0| + \frac{hAB}{hA - 2}\right)(hA - 1)^n - \frac{hAB}{hA - 2}.$$

This remains bounded as $n \to \infty$ if $Ah - 1 \le 1$, that is, $Ah \le 2$, and therefore $h \le 2/A$.

For the implicit method, in contrast,

$$|u^*_{n+1}| \le \frac{1}{1 + ha}(|u^*_n| + hAB), \quad |u^*_0| = |y_0|,$$

giving $|u^*_n| \le E^*_n$ with

$$E^*_n = \frac{1}{(1 + ha)^n}\left(|y_0| - \frac{A}{a}B\right) + \frac{A}{a}B.$$

This remains bounded as $n \to \infty$ for arbitrary $h > 0$.

21. See the text.

ANSWERS TO MACHINE ASSIGNMENTS

1. (a)

```
 PROGRAMS

%EULER  Euler step
%
function ynext=Euler(f,x,y.h)
ynext=y+h*f(x,y);
```

```
%RK4  Fourth-order Runge Kutta step
%
function ynext=RK4(f,x,y,h)
k1=f(x,y);
k2=f(x+.5*h,y+.5*h*k1);
k3=f(x+.5*h,y+.5*h*k2);
k4=f(x+h,y+h*k3);
ynext=y+h*(k1+2*k2+2*k3+k4)/6;
```

(b)

```
 PROGRAMS (Euler's method)

%VARMAV_1B  Variational integration step for MAV_1B
%
function ynext=varMAV_1B(f,x,y,z,h)
ynext=y+h*f(x,y,z);

%fMAV_1B  Differential equation for MAV_1B
%
function yprime=fMAV_1B(x,y)
global A
yprime=A*y;

%FVARMAV_1B  Variational differential equation for MAV_1B
%
function yprime=fvarMAV_1B(x,y,z)
global A
yprime=A*(y-.5*A*z);   % Euler
%yprime=A*(y-(1/120)*A^4*z);   % Runge-Kutta

%MAV_1B
%
f0='%8.2f %12.4e %11.4e  h=%6.4f\n';
f1='%8.2f %12.4e %11.4e\n';
global A
disp('     x         e        e tilde')
lam1=-1; lam2=0; lam3=1;
%lam1=0; lam2=-1; lam3=-10;
%lam1=0; lam2=-1; lam3=-40;
%lam1=0; lam2=-1; lam3=-160;
A=.5*[lam2+lam3 lam3-lam1 lam2-lam1; ...
      lam3-lam2 lam1+lam3 lam1-lam2; ...
      lam2-lam3 lam1-lam3 lam1+lam2];
```

```
for N=[5 10 20 40 80]
  h=1/N;
  u1=[1;1;1]; v=[0;0;0];
  for n=1:N
    x=(n-1)*h; x1=x+h;
    u=u1;
    u1=Euler(@fMAV_1B,x,u,h);  % Euler
%    u1=RK4(@fMAV_1B,x,u,h);  % Runge-Kutta
    v=varMAV_1B(@fvarMAV_1B,x,v,u,h);
    y=[-exp(lam1*x1)+exp(lam2*x1)+exp(lam3*x1); ...
        exp(lam1*x1)-exp(lam2*x1)+exp(lam3*x1); ...
        exp(lam1*x1)+exp(lam2*x1)-exp(lam3*x1)];
    err=norm(u1-y,inf);
    erra=norm(h*v,inf);  % Euler
%    erra=norm(h^4*v,inf);  % Runge-Kutta
    if floor(5*n/N)==5*n/N
      if n==N/5
        fprintf(f0,x1,err,erra,h)
      else
        fprintf(f1,x1,err,erra)
      end
    end
  end
  fprintf('\n')
end
```

```
 OUTPUT
```

(i) Euler's method

$\lambda_1 = -1,\ \lambda_2 = 0,\ \lambda_3 = 1$

```
>> MAV_1B
     x         e         e tilde
  0.20   4.0134e-02  4.0000e-02  h=0.2000
  0.40   8.2145e-02  8.0000e-02
  0.60   1.3093e-01  1.2480e-01
  0.80   1.9167e-01  1.7920e-01
  1.00   2.7016e-01  2.4832e-01

  0.20   2.0134e-02  2.0000e-02  h=0.1000
  0.40   4.1945e-02  4.1200e-02
  0.60   6.7928e-02  6.6030e-02
```

```
    0.80   1.0081e-01  9.7081e-02
    1.00   1.4374e-01  1.3727e-01

    0.20   1.0121e-02  1.0075e-02  h=0.0500
    0.40   2.1269e-02  2.1054e-02
    0.60   3.4714e-02  3.4187e-02
    0.80   5.1869e-02  5.0844e-02
    1.00   7.4378e-02  7.2608e-02

    0.20   5.0788e-03  5.0657e-03  h=0.0250
    0.40   1.0719e-02  1.0662e-02
    0.60   1.7563e-02  1.7424e-02
    0.80   2.6330e-02  2.6062e-02
    1.00   3.7865e-02  3.7402e-02

    0.20   2.5446e-03  2.5411e-03  h=0.0125
    0.40   5.3820e-03  5.3671e-03
    0.60   8.8354e-03  8.7998e-03
    0.80   1.3268e-02  1.3200e-02
    1.00   1.9108e-02  1.8990e-02
>>
```

There is relatively good agreement between $\boldsymbol{e}_n$ and $\tilde{\boldsymbol{e}}_n$, to about 2–3 correct digits when $h = .0125$. The errors are approximately halved from one h to the next, confirming the order $p = 1$ of Euler's method.

$\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -10$

```
>> MAV_1B
     x          e         e tilde
    0.20   1.1541e+00  2.0200e+00  h=0.2000
    0.40   1.0120e+00  4.0320e+00
    0.60   1.0393e+00  6.0384e+00
    0.80   1.0394e+00  8.0410e+00
    1.00   1.0402e+00  1.0041e+01

    0.20   1.4407e-01  9.0000e-03  h=0.1000
    0.40   3.2536e-02  1.4580e-02
    0.60   1.9849e-02  1.7715e-02
    0.80   1.9197e-02  1.9132e-02
    1.00   1.9246e-02  1.9371e-02

    0.20   7.7060e-02  6.6787e-02  h=0.0500
```

```
    0.40   2.1309e-02  1.4796e-02
    0.60   1.0686e-02  9.2644e-03
    0.80   9.5225e-03  9.3269e-03
    1.00   9.4380e-03  9.4386e-03

    0.20   3.7301e-02  3.5465e-02  h=0.0250
    0.40   1.1693e-02  1.0102e-02
    0.60   5.6454e-03  5.1929e-03
    0.80   4.7815e-03  4.6958e-03
    1.00   4.6823e-03  4.6736e-03

    0.20   1.8300e-02  1.7902e-02  h=0.0125
    0.40   6.0636e-03  5.6756e-03
    0.60   2.9044e-03  2.7816e-03
    0.80   2.4010e-03  2.3747e-03
    1.00   2.3338e-03  2.3301e-03
>>
```

For $\|\boldsymbol{u}_n\|$ to remain bounded, we must have $|1 + h\lambda| \leq 1$ for all λ in the spectrum of $\boldsymbol{A}$, thus $h|\lambda| \leq 2$. In the present case, this requires $h \leq .2$. The results for $h = .2$ indeed confirm boundedness of $\|\boldsymbol{u}_n\|$, but there is little agreement between $\boldsymbol{e}_n$ and $\tilde{\boldsymbol{e}}_n$. This gradually improves as h is made smaller, the agreement being 1–3 decimal digits by the time $h = .0125$.

$\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -40$

```
>> MAV_1B
     X         e         e tilde
    0.20   7.0191e+00  3.2020e+01  h=0.2000
    0.40   4.9030e+01  4.4803e+02
    0.60   3.4304e+02  4.7040e+03
    0.80   2.4010e+03  4.3904e+04
    1.00   1.6807e+04  3.8416e+05

    0.20   9.0084e+00  4.8009e+01  h=0.1000
    0.40   8.1014e+01  8.6401e+02
    0.60   7.2902e+02  1.1664e+04
    0.80   6.5610e+03  1.3997e+05
    1.00   5.9049e+04  1.5746e+06

    0.20   1.0039e+00  8.0043e+00  h=0.0500
    0.40   1.0069e+00  1.6007e+01
    0.60   1.0085e+00  2.4009e+01
```

```
    0.80   1.0092e+00  3.2009e+01
    1.00   1.0094e+00  4.0009e+01

    0.20   2.4144e-03  2.0940e-03  h=0.0250
    0.40   3.4000e-03  3.4201e-03
    0.60   4.1701e-03  4.1896e-03
    0.80   4.5465e-03  4.5619e-03
    1.00   4.6470e-03  4.6568e-03

    0.20   1.3516e-03  1.0961e-03  h=0.0125
    0.40   1.6879e-03  1.6927e-03
    0.60   2.0714e-03  2.0762e-03
    0.80   2.2598e-03  2.2636e-03
    1.00   2.3113e-03  2.3137e-03
>> ‘
```

Here, we must have $h \leq \frac{2}{40} = .05$ for boundedness of $\|\boldsymbol{u}_n\|$, which is again confirmed by the numerical results. After this critical value of h, the errors $\boldsymbol{e}_n$ and $\tilde{\boldsymbol{e}}_n$ again agree within 1–3 decimal digits.

$\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -160$

```
>> MAV_1B
     X         e        e tilde
    0.20   3.1019e+01  5.1202e+02  h=0.2000
    0.40   9.6103e+02  3.1744e+04
    0.60   2.9791e+04  1.4761e+06
    0.80   9.2352e+05  6.1012e+07
    1.00   2.8629e+07  2.3642e+09

    0.20   2.2501e+02  3.8400e+03  h=0.1000
    0.40   5.0625e+04  1.7280e+06
    0.60   1.1391e+07  5.8320e+08
    0.80   2.5629e+09  1.7496e+11
    1.00   5.7665e+11  4.9208e+13

    0.20   2.4010e+03  4.3904e+04  h=0.0500
    0.40   5.7648e+06  2.1083e+08
    0.60   1.3841e+10  7.5929e+11
    0.80   3.3233e+13  2.4308e+15
    1.00   7.9792e+16  7.2953e+18

    0.20   6.5610e+03  1.3997e+05  h=0.0250
```

```
    0.40   4.3047e+07  1.8367e+09
    0.60   2.8243e+11  1.8075e+13
    0.80   1.8530e+15  1.5812e+17
    1.00   1.2158e+19  1.2968e+21

    0.20   1.0010e+00  3.2001e+01  h=0.0125
    0.40   1.0017e+00  6.4002e+01
    0.60   1.0021e+00  9.6002e+01
    0.80   1.0023e+00  1.2800e+02
    1.00   1.0023e+00  1.6000e+02
>>
```

Now we have boundedness of $\|\boldsymbol{u}_n\|$ only for $h \leq \frac{2}{160} = .0125$, and the numerical results are in agreement with this. Since, however, we are on the boundary of the interval of absolute stability, the error estimates $\|\tilde{\boldsymbol{e}}\|$ are off by as much as two decimal orders.

(ii) Runge–Kutta method

$\lambda_1 = -1,\ \lambda_2 = 0,\ \lambda_3 = 1$

```
>> MAV_1B
     X         e        e tilde
    0.20   5.3384e-06  5.3333e-06  h=0.2000
    0.40   1.0963e-05  1.0774e-05
    0.60   1.7533e-05  1.6967e-05
    0.80   2.5767e-05  2.4587e-05
    1.00   3.6489e-05  3.4375e-05

    0.20   3.3564e-07  3.3417e-07  h=0.1000
    0.40   7.0044e-07  6.9177e-07
    0.60   1.1366e-06  1.1140e-06
    0.80   1.6908e-06  1.6456e-06
    1.00   2.4176e-06  2.3378e-06

    0.20   2.1096e-08  2.1029e-08  h=0.0500
    0.40   4.4372e-08  4.4054e-08
    0.60   7.2499e-08  7.1707e-08
    0.80   1.0847e-07  1.0691e-07
    1.00   1.5578e-07  1.5303e-07

    0.20   1.3230e-09  1.3206e-09  h=0.0250
```

```
    0.40   2.7936e-09  2.7829e-09
    0.60   4.5799e-09  4.5537e-09
    0.80   6.8710e-09  6.8196e-09
    1.00   9.8889e-09  9.7988e-09

    0.20   8.2847e-11  8.2766e-11  h=0.0125
    0.40   1.7527e-10  1.7492e-10
    0.60   2.8782e-10  2.8697e-10
    0.80   4.3238e-10  4.3073e-10
    1.00   6.2294e-10  6.2005e-10
>>
```

There is good agreement between $\boldsymbol{e}_n$ and $\tilde{\boldsymbol{e}}_n$, and the error reduction from one h to the next is generally close to $\frac{1}{16} = .0625$, confirming order $p = 4$ of the method.

$\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -10$

```
>> MAV_1B
     X         e         e tilde
    0.20   1.9800e-01  2.6667e-01  h=0.2000
    0.40   9.2800e-02  1.7778e-01
    0.60   3.4563e-02  2.0741e-01
    0.80   1.2016e-02  1.9754e-01
    1.00   4.0756e-03  2.0083e-01

    0.20   5.2899e-03  3.1252e-03  h=0.1000
    0.40   1.4600e-03  4.3970e-04
    0.60   3.0246e-04  6.2097e-05
    0.80   5.5929e-05  9.0153e-06
    1.00   9.9270e-06  1.5529e-06

    0.20   2.1450e-04  1.7818e-04  h=0.0500
    0.40   5.8116e-05  3.5301e-05
    0.60   1.1822e-05  5.4970e-06
    0.80   2.1512e-06  8.0568e-07
    1.00   3.8088e-07  1.2918e-07

    0.20   1.0859e-05  9.9534e-06  h=0.0250
    0.40   2.9401e-06  2.3444e-06
    0.60   5.9780e-07  4.1803e-07
    0.80   1.0887e-07  6.7611e-08
    1.00   1.9439e-08  1.1209e-08
```

```
    0.20   6.1127e-07  5.8581e-07  h=0.0125
    0.40   1.6550e-07  1.4849e-07
    0.60   3.3653e-08  2.8322e-08
    0.80   6.1343e-09  4.8618e-09
    1.00   1.1009e-09  8.3737e-10
>>
```

Boundedness of $\|\boldsymbol{u}_n\|$ requires $|\varphi(\lambda h)| \leq 1$ for all λ in the spectrum of A, where $\varphi(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$. This is equivalent to

$$\alpha \leq \lambda h \leq 0, \quad \alpha = -2.78529356\ldots,$$

where α is the unique negative root of (cf. Ex. 17(b))

$$1 + \tfrac{1}{2}x + \tfrac{1}{6}x^2 + \tfrac{1}{24}x^3 = 0.$$

In the case at hand ($\lambda = -10$), this amounts to $h \leq .278529356\ldots$. All h-values satisfy this condition, the first just barely so. Consequently, the errors for $h = .2$ are relatively large, and $\tilde{\boldsymbol{e}}_n$ does not estimate $\boldsymbol{e}_n$ very well. As h decreases, the errors improve, but agreement between $\boldsymbol{e}_n$ and $\tilde{\boldsymbol{e}}_n$ remains relatively poor, except for $h = .0125$, where there is agreement to about one significant digit. The reason for this is that Euler's method used to integrate the variational equation is subject to the constraint $|\lambda|h \leq 2$ to ensure boundedness, thus $h \leq .2$ in the present case.

$\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -40$

```
>> MAV_1B
     X          e         e tilde
    0.20   1.1033e+02  2.7307e+02  h=0.2000
    0.40   1.2173e+04  2.8217e+04
    0.60   1.3431e+06  3.1266e+06
    0.80   1.4819e+08  3.4488e+08
    1.00   1.6351e+10  3.8052e+10

    0.20   2.5000e+01  1.7067e+01  h=0.1000
    0.40   6.2500e+02  5.8027e+02
    0.60   1.5625e+04  1.5889e+04
    0.80   3.9063e+05  4.0967e+05
    1.00   9.7656e+06  1.0354e+07

    0.20   1.2010e-02  1.9753e-01  h=0.0500
```

```
    0.40   1.5232e-04  1.9997e-01
    0.60   1.8995e-06  2.0000e-01
    0.80   4.2750e-08  2.0000e-01
    1.00   2.0263e-08  2.0000e-01

    0.20   5.5604e-05  8.6909e-06  h=0.0250
    0.40   4.1289e-08  4.2913e-09
    0.60   1.1165e-09  1.0964e-09
    0.80   1.1948e-09  1.1939e-09
    1.00   1.2227e-09  1.2203e-09

    0.20   2.1317e-06  7.8622e-07  h=0.0125
    0.40   1.4899e-09  3.3258e-10
    0.60   6.8419e-11  6.7805e-11
    0.80   7.3899e-11  7.3869e-11
    1.00   7.5629e-11  7.5551e-11
>>
```

Here the restrictions are $h \leq \frac{2.78529356\ldots}{40} = .069632339\ldots$ for the Runge–Kutta method, and $h \leq .05$ for Euler's method. Accordingly, for the first two h-values, we have unboundedness of both $\|\boldsymbol{u}_n\|$ and $\|\boldsymbol{v}_n\|$. The third value, $h = .05$, falls within the stability region of the Runge–Kutta method, and hence gives reasonable errors, but lies on the boundary of the stability interval for Euler's method, which explains the poor estimates $\tilde{\boldsymbol{e}}_n$. Both the errors, and to a lesser extent the error estimates, improve as h is further decreased.

$\lambda_1 = 0,\ \lambda_2 = -1,\ \lambda_3 = -160$

```
>> MAV_1B
     x         e        e tilde
    0.20   3.8710e+04  2.7962e+05  h=0.2000
    0.40   1.4985e+09  1.0816e+10
    0.60   5.8007e+13  4.1867e+14
    0.80   2.2455e+18  1.6207e+19
    1.00   8.6923e+22  6.2738e+23

    0.20   4.6699e+06  1.8752e+07  h=0.1000
    0.40   2.1808e+13  8.7575e+13
    0.60   1.0184e+20  4.0897e+20
    0.80   4.7560e+26  1.9098e+27
    1.00   2.2210e+33  8.9188e+33
```

```
    0.20   1.4819e+08  3.4488e+08  h=0.0500
    0.40   2.1961e+16  5.1109e+16
    0.60   3.2545e+24  7.5740e+24
    0.80   4.8229e+32  1.1224e+33
    1.00   7.1472e+40  1.6633e+41

    0.20   3.9063e+05  4.0967e+05  h=0.0250
    0.40   1.5259e+11  1.6271e+11
    0.60   5.9605e+16  6.3578e+16
    0.80   2.3283e+22  2.4835e+22
    1.00   9.0949e+27  9.7013e+27

    0.20   2.3264e-08  2.0000e-01  h=0.0125
    0.40   5.5123e-11  2.0000e-01
    0.60   6.7695e-11  2.0000e-01
    0.80   7.3899e-11  2.0000e-01
    1.00   7.5629e-11  2.0000e-01
>>
```

The restrictions $h \leq \frac{2.78529336\ldots}{160} = .0174080835\ldots$ for the Runge–Kutta method, and $h \leq .0125$ for Euler's method, are now violated for all h except the last, $h = .0125$, where we are within the stability region for Runge–Kutta and on the boundary of the stability interval for Euler. Accordingly, the errors are huge for the first four values of h, and quite small for the last h-value, but the error estimate fails.
As expected, increasing stiffness of the differential system (and its variational equation) has a deteriorating effect on the accuracy of the solution, and even more so on the reliability of the asymptotic error estimation. For nonstiff systems, as the first set of λ-values suggests, the integration and estimation procedures are quite satisfactory.

2. See the text.

3. (a) We have

$$
\begin{aligned}
c'(x) &= -\sin(\tfrac{\pi}{2}x^2)\cdot \pi x,\\
c''(x) &= -\cos(\tfrac{\pi}{2}x^2)(\pi x)^2 - \sin(\tfrac{\pi}{2}x^2)\cdot \pi = -\pi^2 x^2 c(x) - \pi s(x);\\
s'(x) &= \cos(\tfrac{\pi}{2}x^2)\cdot \pi x,\\
s''(x) &= -\sin(\tfrac{\pi}{2}x^2)(\pi x)^2 + \cos(\tfrac{\pi}{2}x^2)\cdot \pi = -\pi^2 x^2 s(x) + \pi c(x);
\end{aligned}
$$

this is consistent with the given system, since $c^2(x) + s^2(x) = 1$. Also,

at $x = 2\sqrt{q}$, we have, q being an integer,

$$c = \cos(2\pi q) = 1, \quad c' = -\sin(2\pi q) \cdot 2\pi\sqrt{q} = 0;$$
$$s = \sin(2\pi q) = 0, \quad s' = \cos(2\pi q) \cdot 2\pi\sqrt{q} = 2\pi\sqrt{q},$$

which are the given initial conditions.

(b) Letting $u = c$, $v = s$, $w = c'$, $z = s'$, we have

$$\begin{aligned}
\frac{\mathrm{d}u}{\mathrm{d}x} &= w, \\
\frac{\mathrm{d}v}{\mathrm{d}x} &= z, \\
\frac{\mathrm{d}w}{\mathrm{d}x} &= -\pi^2 x^2 u - \pi \frac{v}{\sqrt{u^2 + v^2}}, \\
\frac{\mathrm{d}z}{\mathrm{d}x} &= -\pi^2 x^2 v + \pi \frac{u}{\sqrt{u^2 + v^2}},
\end{aligned}$$

with

$$u = 1, \quad v = 0, \quad w = 0, \quad z = 2\pi\sqrt{q} \quad \text{at} \quad x = 2\sqrt{q}.$$

(c)

PROGRAMS

```
%RKF34  Runge-Kutta-Fehlberg(3,4) step
%
function [ynext,phi,phistar]=RKF34(f,x,y,h)
k1=f(x,y);
k2=f(x+2*h/7,y+(2*h/7)*k1);
k3=f(x+7*h/15,y+(77*h/900)*k1+(343*h/900)*k2);
k4=f(x+35*h/38,y+(805*h/1444)*k1-(77175*h/54872)*k2 ...
   +(97125*h/54872)*k3);
phi=(79/490)*k1+(2175/3626)*k3+(2166/9065)*k4;
ynext=y+h*phi;
k5=f(x+h,y+h*phi);
phistar=(229/1470)*k1+(1125/1813)*k3+(13718/81585)*k4+(1/18)*k5;

%fMAV_3C  Differential equation for MAV_3C
%
function yprime=fMAV_3C(x,y)
sq=sqrt(y(1)^2+y(2)^2);
yprime=[y(3);y(4);-pi^2*x^2*y(1)-pi*y(2)/sq; ...
  -pi^2*x^2*y(2)+pi*y(1)/sq];
```

```
%MAV_3C
%
f0='%6.1f %12.8f %12.4e %12.4e\n';
f1='%19.8f %12.4e %12.4e\n';
disp('    q        h         ||u-y||    ||c^2+s^2-1||')
for q=0:3
  N=5/2; err=1;
  while err>.5e-6
    N=2*N; h=1/N;
    e=zeros(N,1); ecs=zeros(N,1);
    u1=[1;0;0;2*pi*sqrt(q)];
    for n=1:N
      x=2*sqrt(q)+(n-1)*h; x1=x+h; xp=pi*(x1^2)/2;
      u=u1; u1=RKF34(@fMAV_3C,x,u,h);
      y=[cos(xp);sin(xp);-pi*x1*sin(xp);pi*x1*cos(xp)];
      e(n)=norm(u1-y,inf); ecs(n)=abs(u1(1)^2+u1(2)^2-1);
    end
    err=norm(e,inf); errcs=norm(ecs,inf);
    if n==5
      fprintf(f0,q,h,err,errcs)
    else
      fprintf(f1,h,err,errcs)
    end
  end
  fprintf('\n')
end
```

OUTPUT

```
>> MAV_3C
    q        h          ||u-y||    ||c^2+s^2-1||
   0.0   0.20000000   2.6370e-03   1.8026e-03
         0.10000000   4.2779e-04   5.4852e-05
         0.05000000   5.1506e-05   1.3514e-05
         0.02500000   6.0316e-06   2.6312e-06
         0.01250000   7.1941e-07   3.8703e-07
         0.00625000   8.7481e-08   5.1976e-08

   1.0   0.20000000   3.4327e+00   5.8768e-01
         0.10000000   3.0156e-01   1.8450e-02
         0.05000000   2.4967e-02   2.1391e-03
         0.02500000   2.2685e-03   4.4341e-04
         0.01250000   2.8909e-04   6.2378e-05
         0.00625000   3.7527e-05   8.0969e-06
         0.00312500   4.7577e-06   1.0265e-06
```

```
          0.00156250   5.9826e-07   1.2908e-07
          0.00078125   7.4985e-08   1.6178e-08

   2.0    0.20000000   8.7222e+00   9.7878e-01
          0.10000000   1.3228e+00   1.2147e-01
          0.05000000   1.1292e-01   3.9646e-03
          0.02500000   1.1082e-02   1.2842e-03
          0.01250000   1.1714e-03   1.8923e-04
          0.00625000   1.3759e-04   2.4753e-05
          0.00312500   1.6787e-05   3.1405e-06
          0.00156250   2.0773e-06   3.9477e-07
          0.00078125   2.5850e-07   4.9461e-08

   3.0    0.20000000   1.5128e+01   9.8868e-01
          0.10000000   2.9882e+00   3.2062e-01
          0.05000000   3.0663e-01   2.9319e-03
          0.02500000   2.5886e-02   2.4075e-03
          0.01250000   2.6686e-03   3.7353e-04
          0.00625000   3.3191e-04   4.9311e-05
          0.00312500   4.2800e-05   6.2659e-06
          0.00156250   5.4185e-06   7.8767e-07
          0.00078125   6.8118e-07   9.8674e-08
          0.00039063   8.5376e-08   1.2346e-08
>>
```

Since with increasing q the solution oscillates more rapidly, one expects slower convergence of the numerical method for larger q. This is confirmed by the results. It is also seen that satisfaction of an identity such as $c^2 + s^2 = 1$ to within a certain accuracy is not an entirely reliable indicator of global accuracy, since the error in this identity, as shown in the last column, is typically about one order of accuracy smaller than the actual global error.

(d) The local error at $x = x_n$ is given by $h\boldsymbol{T}(x_n, \boldsymbol{u}_n; h) = h^{p+1}\boldsymbol{\tau}(x_n, \boldsymbol{u}_n) + O(h^{p+2})$. By (5.126) we therefore have

$$h\boldsymbol{T}(x_n, \boldsymbol{u}_n; h) = h^{p+1}\boldsymbol{r}(x_n, \boldsymbol{u}_n; h) + O(h^{p+2}) \approx h[\boldsymbol{\Phi}(x_n, \boldsymbol{u}_n; h) - \boldsymbol{\Phi}^*(x_n, \boldsymbol{u}_n; h)].$$

The program below compares the local error with the global error.

```
 PROGRAM

%MAV_3D
%
disp('    q         h       global err   est local err')
for q=0:3
  for N=[20 40 80]
```

```
    h=1/N;
    u1=[1;0;0;2*pi*sqrt(q)];
    for n=1:N
      x=2*sqrt(q)+(n-1)*h; x1=x+h; xp=pi*(x1^2)/2;
      u=u1; [u1,phi,phistar]=RKF34(@fMAV_3C,x,u,h);
      y=[cos(xp);sin(xp);-pi*x1*sin(xp);pi*x1*cos(xp)];
      e=norm(u1-y,inf); ea=h*norm(phi-phistar,inf);
      if (n-1)*5/N-fix((n-1)*5/N)==0
        if N==20 & n==1
          fprintf('%6.1f %12.8f %12.4e %13.4e\n',q,h,e,ea)
        elseif n==1
          fprintf('%19.8f %12.4e %13.4e\n',h,e,ea)
        else
          fprintf('%32.4e %13.4e\n',e,ea)
        end
      end
    end
    fprintf('\n')
  end
end
```

OUTPUT

```
>> MAV_3D
        q          h        global err   est local err
      0.0   0.05000000   2.7056e-07    2.7055e-07
                         1.3670e-06    2.6837e-07
                         2.3996e-06    7.0381e-07
                         8.1762e-06    2.2615e-06
                         2.6411e-05    6.1851e-06

            0.02500000   1.6909e-08    1.6909e-08
                         1.5207e-07    1.6606e-08
                         2.8158e-07    4.3086e-08
                         9.7778e-07    1.3976e-07
                         3.0323e-06    3.8017e-07

            0.01250000   1.0568e-09    1.0568e-09
                         1.7880e-08    1.0336e-09
                         3.4234e-08    2.6566e-09
                         1.1869e-07    8.6638e-09
                         3.6231e-07    2.3531e-08

      1.0   0.05000000   2.9080e-04    2.7855e-04
                         1.9729e-03    3.8584e-04
```

```
                    4.8583e-03    5.3669e-04
                    9.5448e-03    8.0120e-04
                    1.5239e-02    1.2852e-03

          0.02500000  1.7483e-05    1.7197e-05
                    1.5184e-04    2.4508e-05
                    3.9848e-04    3.4875e-05
                    7.8698e-04    5.2350e-05
                    1.3445e-03    8.2583e-05

          0.01250000  1.0739e-06    1.0665e-06
                    1.8945e-05    1.5412e-06
                    4.7512e-05    2.2167e-06
                    9.0938e-05    3.3355e-06
                    1.7373e-04    5.2209e-06

2.0       0.05000000  1.6624e-03    1.5516e-03
                    1.1546e-02    2.3284e-03
                    2.9909e-02    3.0959e-03
                    4.9010e-02    3.1160e-03
                    7.5437e-02    5.7466e-03

          0.02500000  9.8335e-05    9.6212e-05
                    1.1203e-03    1.3957e-04
                    2.6881e-03    1.7529e-04
                    3.7911e-03    2.0896e-04
                    7.5855e-03    3.4313e-04

          0.01250000  6.0254e-06    5.9805e-06
                    1.2387e-04    8.5266e-06
                    2.8058e-04    1.0384e-05
                    4.1554e-04    1.3457e-05
                    8.6102e-04    2.0912e-05

3.0       0.05000000  4.7721e-03    4.2512e-03
                    3.8298e-02    5.8639e-03
                    6.7700e-02    6.9280e-03
                    1.4644e-01    9.5174e-03
                    1.9738e-01    1.2983e-02

          0.02500000  2.7180e-04    2.6423e-04
                    3.1854e-03    3.2816e-04
                    5.7394e-03    4.4091e-04
                    1.2240e-02    5.1076e-04
                    1.9024e-02    7.6882e-04
```

```
          0.01250000   1.6588e-05    1.6443e-05
                       3.1631e-04    1.9305e-05
                       7.0363e-04    2.7714e-05
                       1.2002e-03    2.9337e-05
                       2.1060e-03    4.6582e-05
>>
```

The local error is seen to be generally about one order of accuracy smaller than the global error, except for relatively large values of h. Error estimation based on local errors thus underestimates the true (global) error. Note, however, that at the beginning of each integration, the local and global errors are essentially the same, as they should be.

(e) The variational differential equation (5.103), apart from an estimate of the principal error function, requires the Jacobian matrix $\boldsymbol{f}_{\boldsymbol{y}}(x, \boldsymbol{y})$ for the system of differential equations in (b). An elementary computation yields

$$\boldsymbol{f}_{\boldsymbol{y}}(x, \boldsymbol{y}) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\pi^2 x^2 + \pi uv/sq & -\pi u^2/sq & 0 & 0 \\ \pi v^2/sq & -\pi^2 x^2 - \pi uv/sq & 0 & 0 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} u \\ v \\ w \\ z \end{bmatrix},$$

where $sq = (u^2 + v^2)^{3/2}$.

PROGRAMS

```
%VARMAV_3E  Variational integration step for MAV_3E
%
function ynext=varMAV_3E(f,x,y,z,h,pr)
ynext=y+h*f(x,y,z,pr);

%FVARMAV_3E  Variational differential equation for MAV_3E
%
function yprime=fvarMAV_3E(x,y,z,pr)
sq=(z(1)^2+z(2)^2)^(3/2);
J=[0,0,1,0;0,0,0,1;-pi^2*x^2+pi*z(1)*z(2)/sq,-pi*z(1)^2/sq, ...
0,0;pi*z(2)^2/sq,-pi^2*x^2-pi*z(1)*z(2)/sq,0,0];
yprime=J*y+pr;
```

```
%MAV_3E
%
%disp('    q         h       global err   est local err')
for q=0:3
  for N=[20 40 80]
    h=1/N;
    u1=[1;0;0;2*pi*sqrt(q)]; v=[0;0;0;0];
    for n=1:N
      x=2*sqrt(q)+(n-1)*h; x1=x+h; xp=pi*(x1^2)/2;
      u=u1;
      [u1,phi,phistar]=RKF34(@fMAV_3C,x,u,h);
      y=[cos(xp);sin(xp);-pi*x1*sin(xp);pi*x1*cos(xp)];
      pr=h^(-3)*(phi-phistar);
      v=varMAV_3E(@fvarMAV_3E,x,v,u,h,pr);
      e=norm(u1-y,inf); ea=h^3*norm(v,inf);
      if (n-1)*5/N-fix((n-1)*5/N)==0
        if N==20 & n==1
          fprintf('%6.1f %12.8f %12.4e %13.4e\n',q,h,e,ea)
        elseif n==1
          fprintf('%19.8f %12.4e %13.4e\n',h,e,ea)
        else
          fprintf('%32.4e %13.4e\n',e,ea)
        end
      end
    end
    fprintf('\n')
  end
end
```

OUTPUT

```
>> MAV_3E
        q         h       global err   est global err
      0.0   0.05000000   2.7056e-07    2.7055e-07
                         1.3670e-06    1.3494e-06
                         2.3996e-06    2.7405e-06
                         8.1762e-06    8.8658e-06
                         2.6411e-05    2.6329e-05

            0.02500000   1.6909e-08    1.6909e-08
                         1.5207e-07    1.5129e-07
                         2.8158e-07    3.0349e-07
                         9.7778e-07    1.0141e-06
                         3.0323e-06    3.0277e-06
```

```
          0.01250000   1.0568e-09   1.0568e-09
                       1.7880e-08   1.7839e-08
                       3.4234e-08   3.5520e-08
                       1.1869e-07   1.2077e-07
                       3.6231e-07   3.6202e-07

1.0       0.05000000   2.9080e-04   2.7855e-04
                       1.9729e-03   1.5549e-03
                       4.8583e-03   3.8782e-03
                       9.5448e-03   8.7687e-03
                       1.5239e-02   1.7020e-02

          0.02500000   1.7483e-05   1.7197e-05
                       1.5184e-04   1.7393e-04
                       3.9848e-04   4.5215e-04
                       7.8698e-04   9.0453e-04
                       1.3445e-03   1.8356e-03

          0.01250000   1.0739e-06   1.0665e-06
                       1.8945e-05   2.0344e-05
                       4.7512e-05   5.3626e-05
                       9.0938e-05   1.0559e-04
                       1.7373e-04   2.0515e-04

2.0       0.05000000   1.6624e-03   1.5516e-03
                       1.1546e-02   1.1471e-02
                       2.9909e-02   3.2109e-02
                       4.9010e-02   6.1284e-02
                       7.5437e-02   1.3602e-01

          0.02500000   9.8335e-05   9.6212e-05
                       1.1203e-03   1.1426e-03
                       2.6881e-03   2.7679e-03
                       3.7911e-03   4.7770e-03
                       7.5855e-03   1.1146e-02

          0.01250000   6.0254e-06   5.9805e-06
                       1.2387e-04   1.2573e-04
                       2.8058e-04   2.8283e-04
                       4.1554e-04   5.2819e-04
                       8.6102e-04   1.0501e-03

3.0       0.05000000   4.7721e-03   4.2512e-03
                       3.8298e-02   3.3854e-02
                       6.7700e-02   7.3756e-02
```

```
                          1.4644e-01    2.5112e-01
                          1.9738e-01    4.1651e-01

               0.02500000   2.7180e-04    2.6423e-04
                          3.1854e-03    2.9567e-03
                          5.7394e-03    7.8688e-03
                          1.2240e-02    1.6007e-02
                          1.9024e-02    3.4171e-02

               0.01250000   1.6588e-05    1.6443e-05
                          3.1631e-04    3.0221e-04
                          7.0363e-04    8.3966e-04
                          1.2002e-03    1.3316e-03
                          2.1060e-03    2.8433e-03
>>
```

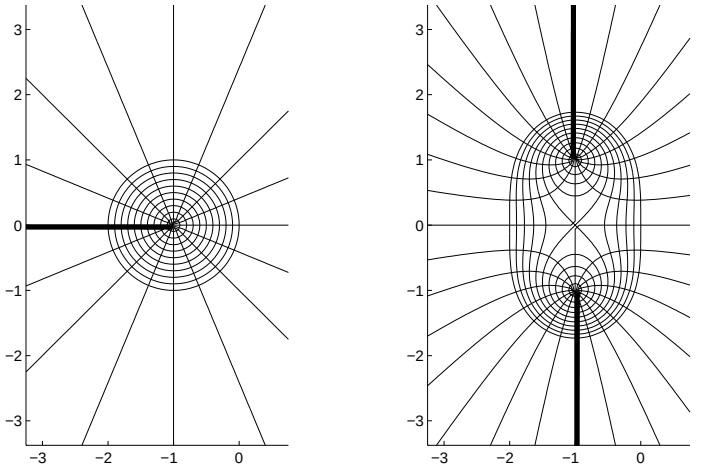
The estimated global error is consistently fairly close to the true global error and in this example is a good indicator of the accuracy obtained.

4. For literature on this topic, see W. Gautschi and J. Waldvogel, Contour plots of analytic functions, in *Solving problems in scientific computing using Maple and Matlab* (W. Gander and J. Hřebíček, eds.), 3d ed., Springer, Berlin, 1997, 359–372.

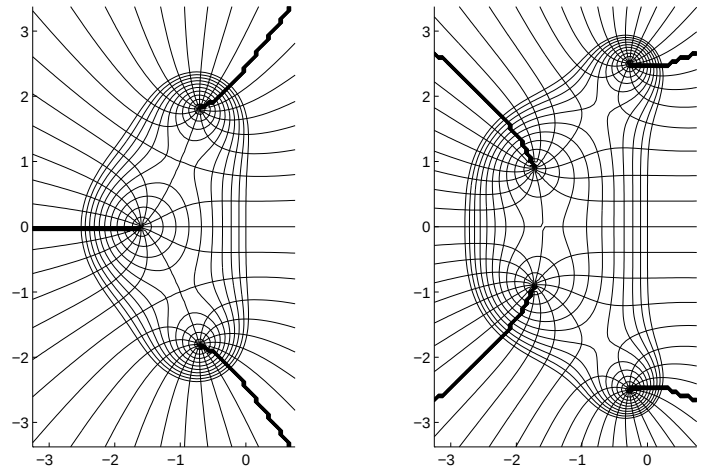
   (a)

```
 PROGRAM

%MAV_4A
%
h=1/16; bounds=[-3.25 .75 -3.375 3.375]; nmax=4;
vabs=[0:10]/10; vang=[-7:8]*pi/8;
x=bounds(1):h:bounds(2);
y=bounds(3):h:bounds(4);
a=ones(size(y'))*x+i*y'*ones(size(x));
t=ones(size(a)); s=t;
for n=1:nmax
  t=t.*a/n; s=s+t;
  figure(n);
  hold on;
  axis(bounds);
  axis('image');
  contour(x,y,abs(s),vabs);
  contour(x,y,angle(s),vang);
  hold off;
end
```

PLOTS



$n = 1$ $n = 2$



$n = 3$ $n = 4$

(b) Write

$$w = f(z), \quad w = r\mathrm{e}^{\mathrm{i}\theta}, \quad z = x + \mathrm{i}y.$$

For the level lines $r = \text{const}$, take θ as independent variable. Differentiating

$$f(z(\theta)) \equiv r\mathrm{e}^{\mathrm{i}\theta}, \quad r = \text{const},$$

with respect to θ then gives

$$f'(z)\frac{\mathrm{d}z}{\mathrm{d}\theta} = \mathrm{i}r\mathrm{e}^{\mathrm{i}\theta} = \mathrm{i}f(z),$$

that is,

$$\frac{\mathrm{d}z}{\mathrm{d}\theta} = \mathrm{i}\frac{f(z)}{f'(z)}\,.$$

With s the arc length, one has

$$\frac{\mathrm{d}s}{\mathrm{d}\theta} = \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}\theta}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}\theta}\right)^2} = \left|\frac{\mathrm{d}z}{\mathrm{d}\theta}\right| = \left|\frac{f(z)}{f'(z)}\right|,$$

so that

$$\frac{\mathrm{d}z}{\mathrm{d}s} = \frac{\mathrm{d}z}{\mathrm{d}\theta}\,\frac{\mathrm{d}\theta}{\mathrm{d}s} = \mathrm{i}\frac{f(z)}{f'(z)}\,\frac{1}{\left|\frac{f(z)}{f'(z)}\right|}\,.$$

Written as a system of differential equations, this is

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}s} &= -\mathrm{Im}\left[\frac{f(z)}{f'(z)}\Big/\left|\frac{f(z)}{f'(z)}\right|\right],\\ \frac{\mathrm{d}y}{\mathrm{d}s} &= \mathrm{Re}\left[\frac{f(z)}{f'(z)}\Big/\left|\frac{f(z)}{f'(z)}\right|\right],\end{aligned} \qquad z = x + \mathrm{i}y.$$

For the phase lines $\theta = \text{const}$, take r as independent variable. Differentiating

$$f(z(r)) = r\mathrm{e}^{\mathrm{i}\theta}, \quad \theta = \text{const},$$

with respect to r then gives

$$\frac{\mathrm{d}z}{\mathrm{d}r} = \frac{1}{r}\,\frac{f(z)}{f'(z)}\,.$$

For the arc length s we now have

$$\frac{\mathrm{d}s}{\mathrm{d}r} = \left|\frac{\mathrm{d}z}{\mathrm{d}r}\right| = \frac{1}{r}\,\left|\frac{f(z)}{f'(z)}\right|,$$

so that

$$\frac{\mathrm{d}z}{\mathrm{d}s} = \frac{\mathrm{d}z}{\mathrm{d}r}\,\frac{\mathrm{d}r}{\mathrm{d}s} = \frac{1}{r}\,\frac{f(z)}{f'(z)}\,\frac{r}{\left|\frac{f(z)}{f'(z)}\right|} = \frac{f(z)}{f'(z)}\,\frac{1}{\left|\frac{f(z)}{f'(z)}\right|}\,,$$

that is,

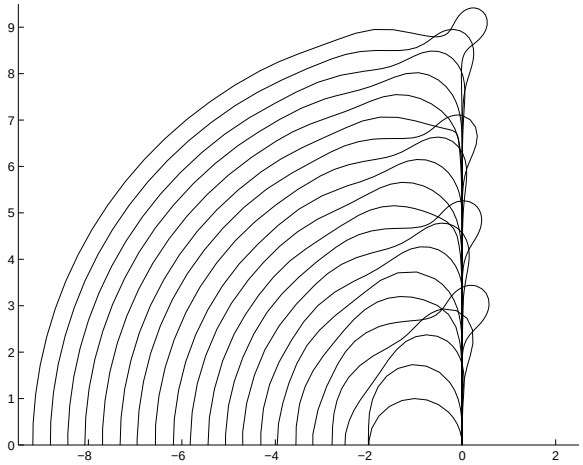
$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}s} &= \mathrm{Re}\left[\frac{f(z)}{f'(z)}\Big/\left|\frac{f(z)}{f'(z)}\right|\right],\\ \frac{\mathrm{d}y}{\mathrm{d}s} &= \mathrm{Im}\left[\frac{f(z)}{f'(z)}\Big/\left|\frac{f(z)}{f'(z)}\right|\right],\end{aligned} \qquad z = x + \mathrm{i}y.$$

(c)

PROGRAMS

```
%MAV_4C
%
global P;
axis([-9.5 2.5 0 9.5]);
axis('image');
hold on
for P=1:21
  z0=[0;0]; sspan=[0 8*pi];
  [s,z]=ode45(@level,sspan,z0);
  n=size(z);
  in=1;
  while z(in,2) >= 0
    in=in+1;
  end
  t=z(in-1,2)/(z(in-1,2)-z(in,2));
  z(in)=t*z(in)+(1-t)*z(in-1);
  plot(z(1:in,1),z(1:in,2));
end

%LEVEL  The differential equation for MAV_4C
%
function zdot=level(s,z)
global P;
zc=z(1)+i.*z(2);
f=1; fac=1; fd=1;
for ip=1:P
  fac=ip*fac;
  f=f+(zc.^ip)./fac;
  if ip == P-1
    fd=f;
  end
end
q=(f/fd)/abs(f/fd);
zdot=[-imag(q);real(q)];
```

PLOT



5. See the text.

EXERCISES AND MACHINE ASSIGNMENTS TO CHAPTER 6

EXERCISES

1. Describe how Newton's method is applied to solve the system of nonlinear equations

$$\boldsymbol{u}_{n+k} = h\beta_k \boldsymbol{f}(x_{n+k}, \boldsymbol{u}_{n+k}) + \boldsymbol{g}_n, \quad \boldsymbol{g}_n = h\sum_{s=0}^{k-1} \beta_s \boldsymbol{f}_{n+s} - \sum_{s=0}^{k-1} \alpha_s \boldsymbol{u}_{n+s}$$

for the next approximation, $\boldsymbol{u}_{n+k}$.

2. The system of nonlinear equations

$$\boldsymbol{u}_{n+k} = h\beta_k \boldsymbol{f}(x_{n+k}, \boldsymbol{u}_{n+k}) + \boldsymbol{g}_n, \quad \beta_k \neq 0,$$

arising in each step of an implicit multistep method (cf. (6.5)) may be solved by

- Newton's method;
- the modified Newton method (with the Jacobian held fixed at its value at the initial approximation);
- the method of successive approximations (fixed point iteration).

Assume that $\boldsymbol{f}$ has continuous second partial derivatives with respect to the $\boldsymbol{u}$-variables and the initial approximation $\boldsymbol{u}_{n+k}^{[0]}$ satisfies $\boldsymbol{u}_{n+k}^{[0]} = \boldsymbol{u}_{n+k} + O(h^g)$ for some $g > 0$.

   (a) Show that the νth iterate $\boldsymbol{u}_{n+k}^{[\nu]}$ in Newton's method has the property that $\boldsymbol{u}_{n+k}^{[\nu]} - \boldsymbol{u}_{n+k} = O(h^{r_\nu})$, where $r_\nu = 2^\nu(g+1)-1$. Derive analogous statements for the other two methods.

   (b) Show that $g = 1$, if one takes $\boldsymbol{u}_{n+k}^{[0]} = \boldsymbol{u}_{n+k-1}$.

   (c) Suppose one adopts the following stopping criterion: quit the iteration after μ iterations, where μ is the smallest integer ν such that $r_\nu > p+1$, where p is the order of the method. For each of the three iterations, determine μ for $p = 2, 3, \ldots, 10$. (Assume $g = 1$ for simplicity.)

   (d) If $g = p$, what would μ be in (c)?

3. (a) Consider a multistep method of the form

$$\boldsymbol{u}_{n+2} - \boldsymbol{u}_{n-2} + \alpha(\boldsymbol{u}_{n+1} - \boldsymbol{u}_{n-1}) = h[\beta(\boldsymbol{f}_{n+1} + \boldsymbol{f}_{n-1}) + \gamma \boldsymbol{f}_n].$$

Show that the parameters α, β, γ can be chosen uniquely so that the method has order $p = 6$. {*Hint*: to preserve symmetry, and thus algebraic simplicity, define the associated linear functional on the interval $[-2, 2]$ rather than $[0, 4]$ as in Sect. 6.1.2. Why is this permissible?}

(b) Discuss the stability properties of the method obtained in (a).

4. For the local error constants γ_k, γ_k^* of, respectively, the Adams–Bashforth and Adams–Moulton method, prove that

$$|\gamma_k^*| < \frac{1}{k-1}\,\gamma_k \quad \text{for} \quad k \geq 2.$$

5. For the local error constants γ_k, γ_k^* of, respectively, the Adams–Bashforth and Adams–Moulton method, show that, as $k \to \infty$,

$$\gamma_k = \frac{1}{\ln k}\,\left[1 + O\left(\frac{1}{\ln k}\right)\right], \quad \gamma_k^* = -\frac{1}{k \ln^2 k}\,\left[1 + O\left(\frac{1}{\ln k}\right)\right].$$

{*Hint*: express the constants in terms of the gamma function, use

$$\frac{\Gamma(k+t)}{\Gamma(k+1)} = k^{t-1}\left[1 + O\left(\frac{1}{k}\right)\right],\ k \to \infty,$$

and integrate by parts.}

6. Consider the predictor-corrector method using the Adams–Bashforth formula as predictor and the Adams–Moulton formula (once) as corrector, both in difference form:

$$\begin{aligned}
\overset{\circ}{\boldsymbol{u}}_{n+k} &= \boldsymbol{u}_{n+k-1} + h\sum_{s=0}^{k-1} \gamma_s \nabla^s \boldsymbol{f}_{n+k-1},\\
\boldsymbol{u}_{n+k} &= \boldsymbol{u}_{n+k-1} + h\sum_{s=0}^{k-1} \gamma_s^* \nabla^s \overset{\circ}{\boldsymbol{f}}_{n+k},\\
\boldsymbol{f}_{n+k} &= \boldsymbol{f}(x_{n+k}, \boldsymbol{u}_{n+k}),
\end{aligned}$$

where $\overset{\circ}{\boldsymbol{f}}_{n+k} = \boldsymbol{f}(x_{n+k}, \overset{\circ}{\boldsymbol{u}}_{n+k})$, $\nabla \overset{\circ}{\boldsymbol{f}}_{n+k} = \overset{\circ}{\boldsymbol{f}}_{n+k} - \boldsymbol{f}_{n+k-1}$, and so on.

(a) Show that

$$\boldsymbol{u}_{n+k} = \overset{\circ}{\boldsymbol{u}}_{n+k} + h\gamma_{k-1}\nabla^k \overset{\circ}{\boldsymbol{f}}_{n+k}.$$

{*Hint*: first show that $\gamma_s^* = \gamma_s - \gamma_{s-1}$ for $s = 0, 1, 2, \ldots$, where γ_{-1} is defined to be zero.}

(b) Show that

$$\nabla^k \overset{\circ}{\boldsymbol{f}}_{n+k} = \overset{\circ}{\boldsymbol{f}}_{n+k} - \sum_{s=0}^{k-1} \nabla^s \boldsymbol{f}_{n+k-1}.$$

{*Hint*: use the binomial identity $\sum_{\sigma=0}^{m} \binom{\sigma+j}{\sigma} = \binom{m+j+1}{j+1}$.}

7. Prove that the predictor-corrector method

$$\begin{cases} \mathring{\boldsymbol{u}}_{n+k} = -\sum_{s=0}^{k-1} \alpha_s \boldsymbol{u}_{n+s} + h \sum_{s=0}^{k-1} \beta_s \boldsymbol{f}_{n+s}, \\ \boldsymbol{u}_{n+k} = -\sum_{s=1}^{k-1} \alpha_s^* \boldsymbol{u}_{n+s} + h\{\beta_k^* \boldsymbol{f}(x_{n+k}, \mathring{\boldsymbol{u}}_{n+k}) + \sum_{s=1}^{k-1} \beta_s^* \boldsymbol{f}_{n+s}\}, \\ \boldsymbol{f}_{n+k} = \boldsymbol{f}(x_{n+k}, \boldsymbol{u}_{n+k}) \end{cases}$$

is stable for every $\boldsymbol{f} \in \mathcal{F}$ (cf. (6.87), (6.88)), if and only if its characteristic polynomial $\alpha^*(t) = \sum_{s=0}^{k} \alpha_s^* t^s$, $\alpha_0^* = 0$, satisfies the root condition.

8. Let $\alpha(\zeta) = \omega(\zeta)\alpha_0(\zeta)$, $\beta(\zeta) = \omega(\zeta)\beta_0(\zeta)$, and suppose $\{\boldsymbol{u}_n\}$ is a solution of the difference equation (6.2) corresponding to $\{\alpha_0, \beta_0\}$. Show that $\{\boldsymbol{u}_n\}$ also satisfies the difference equation (6.2) corresponding to $\{\alpha, \beta\}$.

9. Construct a pair of four-step methods, one explicit, the other implicit, both having $\alpha(\zeta) = \zeta^4 - \zeta^3$ and order $p = 4$, but global error constants that are equal in modulus and opposite in sign.

10. (a) Compute the zeros of the characteristic polynomial $\alpha(t)$ of the k-step backward differentiation method (6.148) for $k = 1(1)7$ and the modulus of the absolutely largest zero other than 1. Hence, confirm the statement made at the end of Sect. 6.4.1.

    (b) Compare the error constant of the k-step backward differentiation method with that of the k-step Adams–Moulton method for $k = 1(1)7$.

11. (a) Show that the polynomial $b(z)$ in (6.152), for an explicit k-step method, must satisfy $b(1) = 0$.

    (b) Use the proof techniques of Theorem 6.4.2 to show that every stable explicit k-step method has order $p \leq k$. {*Hint*: make use of (6.160).}

12. Determine $\min |C_{k,k+1}|$, where the minimum is taken over all k-step methods of order $k+1$ whose characteristic polynomials have all their zeros ζ_i (except $\zeta_1 = 1$) in the disk $\Gamma_\gamma = \{z \in \mathbb{C} : |z| \leq \gamma\}$, where γ is a prescribed number with $0 \leq \gamma < 1$. {*Hint*: use the theory developed in Sects. 6.4.2 and 6.4.3.}

13. Prove Theorem 6.4.4(b).

MACHINE ASSIGNMENTS

1. This assignment pertains to an initial value problem for a scalar first-order differential equation.

(a) A kth-order Adams–Bashforth predictor step amounts to adding h times a linear combination $\sum_{s=0}^{k-1} \gamma_s \nabla^s f_{n+k-1}$ of k backward differences to the last approximation $u_{\text{last}} = u_{n+k-1}$. Write a Matlab routine `AB.m` implementing this step for $k = 1 : 10$. Use Maple to generate the required coefficients γ_s. Take as input variables the number u_{last}, the k-vector $\boldsymbol{F} = [f_n, f_{n+1}, \ldots, f_{n+k-1}]$ of k successive function values, k, and h.

(b) Do the same as in (a) for the kth-order Adams–Moulton corrector step (in Newton's form), writing a routine `AM.m` whose input is $u_{\text{last}} = u_{n+k-1}$, the vector $\boldsymbol{F0} = [f_{n+1}, f_{n+2}, \ldots, f_{n+k-1}, \mathring{f}_{n+k}]$, k, and h.

(c) Use the routines in (a) and (b) to write a routine `PECE.m` implementing the PECE predictor/corrector scheme (6.64) based on the pair of Adams predictor and corrector formulae:

$$\begin{aligned}
\text{P}: &\quad \mathring{u}_{n+k} = u_{n+k-1} + h \sum_{s=0}^{k-1} \gamma_s \nabla^s f_{n+k-1}, \\
\text{E}: &\quad \mathring{f} = f(x_{n+k}, \mathring{u}_{n+k}), \\
\text{C}: &\quad u_{n+k} = u_{n+k-1} + h \sum_{s=0}^{k-1} \gamma_s^* \nabla^s \mathring{f}_{n+k}, \\
\text{E}: &\quad f_{n+k} = f(x_{n+k}, u_{n+k}),
\end{aligned}$$

where $\nabla \mathring{f}_{n+k} = \mathring{f}_{n+k} - f_{n+k-1}$, $\nabla^2 \mathring{f}_{n+k} = \nabla(\nabla \mathring{f}_{n+k}) = \mathring{f}_{n+k} - 2f_{n+k-1} + f_{n+k-2}$, etc. As input parameters include the function f, the initial and final values of x, the k initial approximations, the order k, the number N of (equally spaced) grid intervals, and the values of $n + k$ at which printout is to occur.

2. (a) Consider the initial value problem

$$\frac{dy}{dx} = \frac{1}{1 - \varepsilon \cos y}, \quad y(0) = 0, \quad 0 \le x \le 2\pi, \quad 0 < \varepsilon < 1.$$

Show that the exact solution $y = y(x)$ is the solution of Kepler's equation $y - \varepsilon \sin y - x = 0$ (cf. Ch. 4, Ex. 21 and Ch. 5, MA 5(a) and (c)). What is $y(\pi)$? What is $y(2\pi)$?

(b) Use the routine `AB.m` of MA 1(a) to solve the initial value problem by the Adams–Bashforth methods of orders $k = 2, 4, 6$, with $N = 40, 160, 640$ integration steps of length $h = 2\pi/N$. At $x = \frac{1}{2}\pi, \pi, \frac{3}{2}\pi, 2\pi$, i.e., for $n+k = \frac{1}{4}N, \frac{1}{2}N, \frac{3}{4}N, N$, print the approximations u_{n+k} obtained, along with the errors $\text{err} = u_{n+k} - y(x)$ and the scaled errors err/h^k. Compute $y(x)$ by applying Newton's method to solve Kepler's equation.

(c) Do the same as in (b) with the `PECE.m` routine of MA 1(c).

In both programs start with "exact" initial values. According to (6.107) and the remarks at the end of Sect. 6.3.4, the scaled errors in the printout should be approximately equal to $C_{k,k}e(x)$ resp. $C^*_{k,k}e(x)$, where $C_{k,k}$, $C^*_{k,k}$ are the global error constants of the kth-order Adams–Bashforth resp. the kth-order Adams predictor/corrector scheme, and $x = \frac{1}{2}\pi, \pi, \frac{3}{2}\pi, 2\pi$. Thus, the errors of the predictor/corrector scheme should be approximately equal to $(C^*_{k,k}/C_{k,k})$ times the errors in the Adams–Bashforth method. Examine to what extent this is confirmed by your numerical results.

3. Use the analytic characterization of order given in Theorem 6.4.1, in conjunction with Maple's series expansion capabilities, to

   (a) determine the coefficients $\{\beta_{k,s}\}_{s=0}^{k-1}$ of the kth-order Adams–Bashforth method (6.48) for $k = 1 : 10$;

   (b) determine the coefficients $\{\beta^*_{k,s}\}_{s=1}^{k}$ of the kth-order Adams–Moulton method (6.56) for $k = 1 : 10$.

4. (a) Write a Matlab routine for plotting the regions $\mathcal{D}_A$ of absolute stability for the kth-order Adams–Moulton methods, $k = 3 : 10$. {*Hint*: seek the boundaries of the regions $\mathcal{D}_A$ in polar coordinates.} In particular, compute the abscissae of absolute stability on the negative real axis.

   (b) Do the same for the Adams (PECE) predictor/corrector method. Compare the stability properties of this predictor/corrector method with those of the corrector alone.

5. Consider the (slightly modified) model problem

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\omega[y - a(x)], \quad 0 \le x \le 1; \quad y(0) = y_0,$$

   where $\omega > 0$ and (i) $a(x) = x^2$, $y_0 = 0$; (ii) $a(x) = \mathrm{e}^{-x}$, $y_0 = 1$; (iii) $a(x) = \mathrm{e}^{x}$, $y_0 = 1$.

   (a) In each of the cases (i) through (iii), obtain the exact solution $y(x)$.

   (b) In each of the cases (i) through (iii), apply the kth-order Adams predictor/corrector method, for $k = 2 : 5$, using exact starting values and step lengths $h = \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}$. Print the exact values y_n and the errors $u_n - y_n$ for $x_n = .25, .5, .75, 1$. Try $\omega = 1$, $\omega = 10$, and $\omega = 50$. Summarize your results.

   (c) Repeat (b), but using the k-step backward differentiation method (6.148) (in Lagrangian form).

6. Consider the nonlinear system

$$\begin{aligned} \frac{\mathrm{d}y_1}{\mathrm{d}x} &= 2y_1(1 - y_2), \quad y_1(0) = 1, \\ & \qquad\qquad 0 \le x \le 10, \\ \frac{\mathrm{d}y_2}{\mathrm{d}x} &= -y_2(1 - y_1), \quad y_2(0) = 3, \end{aligned}$$

of interest in population dynamics.

(a) Use Matlab's `ode45` routine to plot the solution $\boldsymbol{y} = [y_1, y_2]^T$ of the system to get an idea of its behavior. Also plot the norm of the Jacobian matrix $\boldsymbol{f}_{\boldsymbol{y}}$ along the solution curve to check on the stiffness of the system.

(b) Determine a step length h, or the corresponding number N of steps, in the classical Runge–Kutta method that would produce about eight correct decimal digits. {*Hint*: for $N = 10, 20, 40, 80, \ldots$ compute the solution with N steps and $2N$ steps and stop as soon as the two solutions agree to within eight decimal places at all grid points common to both solutions. For the basic Runge–Kutta step, use the routine `RK4` from Ch. 5, MA 1(a).}

(c) Apply $N = 640$ steps of the pair of fourth-order methods constructed in Ex. 9 to obtain asymptotically upper and lower bounds to the solution. Plot suitably scaled errors $\boldsymbol{u}_n - \boldsymbol{y}_n$, $\boldsymbol{u}_n^* - \boldsymbol{y}_n$, $n = 1(1)N$, where $\boldsymbol{y}_n$ is the solution computed in (b) by the Runge–Kutta method. For the required initial approximations, use the classical Runge–Kutta method. Use Newton's method to solve the implicit equation for $\boldsymbol{u}_n^*$.

ANSWERS TO THE EXERCISES AND MACHINE ASSIGNMENTS OF CHAPTER 6

ANSWERS TO EXERCISES

1. The system to be solved is

$$\boldsymbol{g}(\boldsymbol{u}_{n+k}) = \boldsymbol{0}, \quad \boldsymbol{g}(\boldsymbol{y}) := \boldsymbol{y} - h\beta_k \boldsymbol{f}(x_{n+k}, \boldsymbol{y}) - \boldsymbol{g}_n.$$

Since

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} = \boldsymbol{I} - h\beta_k \boldsymbol{f}_{\boldsymbol{y}}(x_{n+k}, \boldsymbol{y}),$$

Newton's method consists of the iteration

$$\boldsymbol{u}_{n+k}^{[i+1]} = \boldsymbol{u}_{n+k}^{[i]} - \boldsymbol{\Delta}^{[i]}, \quad i = 0, 1, 2, \ldots,$$

where $\boldsymbol{\Delta}^{[i]}$ is the solution of the linear system

$$[\boldsymbol{I} - h\beta_k \boldsymbol{f}_{\boldsymbol{y}}(x_{n+k}, \boldsymbol{u}_{n+k}^{[i]})]\boldsymbol{\Delta}^{[i]} = \boldsymbol{u}_{n+k}^{[i]} - h\beta_k \boldsymbol{f}(x_{n+k}, \boldsymbol{u}_{n+k}^{[i]}) - \boldsymbol{g}_n.$$

An explicit multistep method may be used to obtain an initial approximation $\boldsymbol{u}_{n+k}^{[0]}$.

2. See the text.

3. (a) By Theorem 6.1.2, the (algebraic) order of a multistep method is unaffected by a translation (even a scaling) of the independent variable. We may thus take

$$Lu = u(2) - u(-2) + \alpha[u(1) - u(-1)] - \beta[u'(1) + u'(-1)] - \gamma u'(0).$$

Evidently, $Lt^r = 0$ for r even. Thus, to achieve order $p = 6$, it suffices to require

$$\begin{aligned} Lt &= 4 + 2\alpha - 2\beta - \gamma = 0, \\ Lt^3 &= 16 + 2\alpha - 6\beta = 0, \\ Lt^5 &= 64 + 2\alpha - 10\beta = 0. \end{aligned}$$

This gives, uniquely,

$$\alpha = 28, \quad \beta = 12, \quad \gamma = 36.$$

(b) The characteristic polynomial (6.75) is

$$\alpha(t) = t^4 - 1 + \alpha(t^3 - t) = (t^2 - 1)(t^2 + 28t + 1).$$

One of the zeros is $-\frac{1}{2}(28 + \sqrt{780}) = -27.964\ldots$, which is larger than 1 in absolute value. Hence, the method is (strongly) unstable.

4. From (6.59) we have

$$|\gamma_k^*| = \left| \int_{-1}^{0} \binom{t+k-1}{k} \mathrm{d}t \right| = \left| \int_{0}^{1} \binom{t+k-2}{k} \mathrm{d}t \right|.$$

Using

$$\binom{t+k-2}{k} = \frac{t-1}{t+k-1} \binom{t+k-1}{k},$$

we get

$$|\gamma_k^*| = \int_0^1 \frac{1-t}{t+k-1} \binom{t+k-1}{k} \mathrm{d}t, \quad k \geq 1.$$

If $k \geq 2$, then

$$\frac{1-t}{t+k-1} < \frac{1}{k-1} \quad \text{on} \;\; 0 < t < 1.$$

Therefore, by (6.53),

$$|\gamma_k^*| < \frac{1}{k-1} \int_0^1 \binom{t+k-1}{k} \mathrm{d}t = \frac{1}{k-1}\, \gamma_k.$$

5. In terms of the gamma functions, we have, from (6.53), that

$$\gamma_k = \int_0^1 \frac{\Gamma(k+t)}{\Gamma(k+1)} \frac{\mathrm{d}t}{\Gamma(t)}.$$

Using the *Hint*, one gets

$$\gamma_k = \left\{ \int_0^1 \frac{k^{t-1}}{\Gamma(t)} \mathrm{d}t \right\} \left[1 + O\left(\frac{1}{k}\right) \right], \quad k \to \infty.$$

Let

$$h(t) = \frac{1}{\Gamma(t)} = \frac{t}{\Gamma(t+1)}, \quad 0 \leq t \leq 1.$$

Integrating by parts, we find

$$\begin{aligned} \int_0^1 \frac{k^{t-1}}{\Gamma(t)} \mathrm{d}t = \int_0^1 h(t) k^{t-1} \mathrm{d}t &= h(t) \frac{1}{\ln k}\, k^{t-1} \Big|_0^1 - \int_0^1 h'(t) \frac{1}{\ln k}\, k^{t-1} \mathrm{d}t \\ &= \frac{1}{\ln k} \left[h(1) - \frac{h(0)}{k} \right] - \frac{1}{\ln k} \int_0^1 h'(t) k^{t-1} \mathrm{d}t. \end{aligned}$$

Since $h'(t)$ is continuous and bounded on $[0,1]$, the last integral is of order $O\left(\frac{1}{\ln k}\right)$. Moreover, $h(0) = 0$ and $h(1) = 1$. There follows

$$\int_0^1 \frac{k^{t-1}}{\Gamma(t)} \mathrm{d}t = \frac{1}{\ln k} \left[1 + O\left(\frac{1}{\ln k}\right) \right],$$

hence

$$\gamma_k = \frac{1}{\ln k}\left[1 + O\left(\frac{1}{\ln k}\right)\right].$$

Similarly, from (6.59),

$$\gamma_k^* = \int_{-1}^{0} \frac{\Gamma(k+t)}{\Gamma(k+1)} \frac{\mathrm{d}t}{\Gamma(t)} = \left\{\int_{-1}^{0} \frac{k^{t-1}}{\Gamma(t)}\mathrm{d}t\right\}\left[1 + O\left(\frac{1}{k}\right)\right], \quad k \to \infty.$$

With $h(t)$ as defined above, integrating twice by parts,

$$\begin{aligned}\int_{-1}^{0} \frac{k^{t-1}}{\Gamma(t)}\mathrm{d}t &= \int_{-1}^{0} h(t)k^{t-1}\mathrm{d}t = h(t)\,\frac{1}{\ln k}\,k^{t-1}\Big|_{-1}^{0} - \int_{-1}^{0} h'(t)\,\frac{1}{\ln k}\,k^{t-1}\mathrm{d}t \\ &= \frac{1}{k\ln k}\left[h(0) - \frac{h(-1)}{k}\right] - \frac{1}{\ln k}\int_{-1}^{0} h'(t)k^{t-1}\mathrm{d}t \\ &= \frac{1}{k\ln k}\left[h(0) - \frac{h(-1)}{k}\right] - \frac{1}{k\ln^2 k}\left[h'(0) - \frac{h'(-1)}{k}\right] \\ &\qquad + \frac{1}{\ln^2 k}\int_{-1}^{0} h''(t)k^{t-1}\mathrm{d}t.\end{aligned}$$

As before, $h(0) = 0$. Writing $h(t) = \frac{t(t+1)}{\Gamma(t+2)}$, we find that $h(-1) = 0$, $h'(0) = 1$, $h'(-1) = -1$, and $h''(t)$ is bounded on $[-1, 0]$, so that the last integral is of order $O\left(\frac{1}{k\ln k}\right)$. Therefore,

$$\begin{aligned}\int_{-1}^{0} \frac{k^{t-1}}{\Gamma(t)}\mathrm{d}t &= -\frac{1}{k\ln^2 k}\left[1 + \frac{1}{k}\right] + O\left(\frac{1}{k\ln^3 k}\right) \\ &= -\frac{1}{k\ln^2 k}\left[1 + O\left(\frac{1}{\ln k}\right)\right], \quad k \to \infty,\end{aligned}$$

hence

$$\gamma_k^* = -\frac{1}{k\ln^2 k}\left[1 + O\left(\frac{1}{\ln k}\right)\right], \quad k \to \infty.$$

6. (a) We have

$$\begin{aligned}\gamma_s^* &= \int_{-1}^{0} \binom{t+s-1}{s}\mathrm{d}t = \int_{0}^{1} \binom{t+s-2}{s}\mathrm{d}t \\ &= \int_{0}^{1} \frac{(t+s-2)\cdots t(t-1)}{s!}\,\mathrm{d}t = \int_{0}^{1} \frac{(t+s-2)\cdots t(t+s-1-s)}{s!}\,\mathrm{d}t \\ &= \int_{0}^{1} \left[\binom{t+s-1}{s} - \binom{t+s-2}{s-1}\right]\mathrm{d}t \\ &= \gamma_s - \gamma_{s-1}.\end{aligned}$$

Therefore,

$$\begin{aligned}\boldsymbol{u}_{n+k} - \overset{\circ}{\boldsymbol{u}}_{n+k} &= h\sum_{s=0}^{k-1}(\gamma_s^*\nabla^s \overset{\circ}{\boldsymbol{f}}_{n+k} - \gamma_s\nabla^s \boldsymbol{f}_{n+k-1}) \\ &= h\sum_{s=0}^{k-1}[\gamma_s(\nabla^s \overset{\circ}{\boldsymbol{f}}_{n+k} - \nabla^s \boldsymbol{f}_{n+k-1}) - \gamma_{s-1}\nabla^s \overset{\circ}{\boldsymbol{f}}_{n+k}] \\ &= h\sum_{s=0}^{k-1}(\gamma_s\nabla^{s+1} \overset{\circ}{\boldsymbol{f}}_{n+k} - \gamma_{s-1}\nabla^s \overset{\circ}{\boldsymbol{f}}_{n+k}) \\ &= h\gamma_{k-1}\nabla^k \overset{\circ}{\boldsymbol{f}}_{n+k},\end{aligned}$$

as claimed.

(b) We have

$$\overset{\circ}{\boldsymbol{f}}_{n+k} - \sum_{s=0}^{k-1}\nabla^s \boldsymbol{f}_{n+k-1} = \overset{\circ}{\boldsymbol{f}}_{n+k} - \sum_{s=0}^{k-1}\sum_{j=0}^{s}(-1)^j\binom{s}{j}\boldsymbol{f}_{n+k-j-1}.$$

Interchanging the order of summation gives for the right-hand side

$$\overset{\circ}{\boldsymbol{f}}_{n+k} - \sum_{j=0}^{k-1}(-1)^j \boldsymbol{f}_{n+k-j-1}\sum_{s=j}^{k-1}\binom{s}{j}.$$

But

$$\sum_{s=j}^{k-1}\binom{s}{j} = \sum_{\sigma=0}^{k-j-1}\binom{\sigma+j}{j} = \sum_{\sigma=0}^{k-j-1}\binom{\sigma+j}{\sigma} = \binom{k}{j+1},$$

where the *Hint* with $m = k-j-1$ was used in the last equation. Therefore, the right-hand side above becomes

$$\overset{\circ}{\boldsymbol{f}}_{n+k} - \sum_{j=0}^{k-1}(-1)^j\binom{k}{j+1}\boldsymbol{f}_{n+k-j-1} = \overset{\circ}{\boldsymbol{f}}_{n+k} + \sum_{s=1}^{k}(-1)^s\binom{k}{s}\boldsymbol{f}_{n+k-s} = \nabla^k \overset{\circ}{\boldsymbol{f}}_{n+k},$$

as was to be shown.

7. The necessity of the root condition follows as in the proof of Theorem 6.3.3. For the sufficiency, one uses the definition (6.100) of R_h^{PECE} to obtain

$$\sum_{s=0}^{k}\alpha_s^*(\boldsymbol{v}_{n+s} - \boldsymbol{w}_{n+s}) = \boldsymbol{\psi}_{n+k}, \quad n = 0, 1, \ldots, N-k,$$

where

$$\begin{aligned}\boldsymbol{\psi}_{n+k} = h\{&\beta_k^* [\boldsymbol{f}(x_{n+k}, \mathring{\boldsymbol{v}}_{n+k}) - \boldsymbol{f}(x_{n+k}, \mathring{\boldsymbol{w}}_{n+k})] \\ &+ \sum_{s=1}^{k-1} \beta_s^* [\boldsymbol{f}(x_{n+s}, \boldsymbol{v}_{n+s}) - \boldsymbol{f}(x_{n+s}, \boldsymbol{w}_{n+s})]\} \\ &+ h[(R_h^{PECE}\boldsymbol{v})_n - (R_h^{PECE}\boldsymbol{w})_n].\end{aligned}$$

From the theory of difference equations,

$$\boldsymbol{v}_n - \boldsymbol{w}_n = \sum_{s=0}^{k-1} h_{n,s}^* (\boldsymbol{v}_s - \boldsymbol{w}_s) + \sum_{m=k}^{n} g_{n,m}^* \boldsymbol{\psi}_m,$$

where $h_{n,s}^*$, $g_{n,m}^*$ are defined as in Sect. 6.3.1, but for the difference equation with coefficients α_s^*. Since its characteristic polynomial $\alpha^*(t)$ is assumed to satisfy the root condition, we have (cf. (6.86))

$$|h_{n,s}^*| \le M^*, \quad |g_{n,m}^*| \le M^*$$

for some $M^* \ge 1$. Therefore,

$$\|\boldsymbol{v}_n - \boldsymbol{w}_n\| \le M^* \{k \max_{0 \le s \le k-1} \|\boldsymbol{v}_s - \boldsymbol{w}_s\| + \sum_{m=k}^{n} \|\psi_m\|\}.$$

We now have

$$\begin{aligned}\|\boldsymbol{\psi}_m\| = \|h\{&\beta_k^* [\boldsymbol{f}(x_m, \mathring{\boldsymbol{v}}_m) - \boldsymbol{f}(x_m, \mathring{\boldsymbol{w}}_m)] \\ &+ \sum_{s=1}^{k-1} \beta_s^* [\boldsymbol{f}(x_{m-k+s}, \boldsymbol{v}_{m-k+s}) - \boldsymbol{f}(x_{m-k+s}, \boldsymbol{w}_{m-k+s})]\} \\ &+ h[(R_h^{PECE}\boldsymbol{v})_{m-k} - (R_h^{PECE}\boldsymbol{w})_{m-k}]\| \\ \le h\beta^* L\{&\|\mathring{\boldsymbol{v}}_m - \mathring{\boldsymbol{w}}_m\| + \sum_{s=1}^{k-1} \|\boldsymbol{v}_{m-k+s} - \boldsymbol{w}_{m-k+s}\|\} \\ &+ h\|R_h^{PECE}\boldsymbol{v} - R_h^{PECE}\boldsymbol{w}\|_\infty,\end{aligned}$$

where

$$\beta^* = \max_{1 \le s \le k} |\beta_s^*|.$$

Furthermore,

$$\begin{aligned}\|\mathring{\boldsymbol{v}}_m - \mathring{\boldsymbol{w}}_m\| = \| &- \sum_{s=0}^{k-1} \alpha_s [\boldsymbol{v}_{m-k+s} - \boldsymbol{w}_{m-k+s}] \\ &+ h \sum_{s=0}^{k-1} \beta_s [\boldsymbol{f}(x_{m-k+s}, \boldsymbol{v}_{m-k+s}) - \boldsymbol{f}(x_{m-k+s}, \boldsymbol{w}_{m-k+s})]\| \\ &\leq \alpha \sum_{s=0}^{k-1} \|\boldsymbol{v}_{m-k+s} - \boldsymbol{w}_{m-k+s}\| + h\beta L \sum_{s=0}^{k-1} \|\boldsymbol{v}_{m-k+s} - \boldsymbol{w}_{m-k+s}\|,\end{aligned}$$

where

$$\alpha = \max_{0 \leq s \leq k-1} |\alpha_s|, \qquad \beta = \max_{0 \leq s \leq k-1} |\beta_s|.$$

There follows

$$\begin{aligned}\|\boldsymbol{\psi}_m\| \leq h\beta^* L(\alpha + 1 + (b-a)\beta L) &\sum_{s=0}^{k-1} \|\boldsymbol{v}_{m-k+s} - \boldsymbol{w}_{m-k+s}\| \\ &+ h\|R_h^{PECE}\boldsymbol{v} - R_h^{PECE}\boldsymbol{w}\|_\infty,\end{aligned}$$

which is the same inequality as (6.93) in the proof of Theorem 6.3.3, only with different constants, namely $h\beta L$ replaced by $h\beta^* L(\alpha + 1 + (b-a)\beta L)$ and $\beta_k = 0$.. The rest of the proof, therefore, is identical to the one given in Sect. 6.3.2 for Theorem 6.3.3.

8. Define the shift operator E, applied to any sequence $\{\boldsymbol{v}_n\}$, by

$$E\boldsymbol{v}_n = \boldsymbol{v}_{n+1}.$$

Then $E^2\boldsymbol{v}_n = E(E\boldsymbol{v}_n) = \boldsymbol{v}_{n+2}$, etc., and by the linearity of E, one can define $p(E)$ for any polynomial p. Moreover, given two polynomials p and q, one has

$$(pq)(E) = p(E)q(E).$$

Suppose now that $\alpha(\zeta) = \omega(\zeta)\alpha_0(\zeta)$ and $\beta(\zeta) = \omega(\zeta)\beta_0(\zeta)$. Then a solution of the difference equation (6.2) corresponding ot $\{\alpha_0, \beta_0\}$ satisfies

$$\alpha_0(E)\boldsymbol{u}_n - h\beta_0(E)\boldsymbol{f}_n = 0,$$

from which, multiplying by $\omega(E)$, one gets

$$\omega(E)\alpha_0(E)\boldsymbol{u}_n - h\omega(E)\beta_0(E)\boldsymbol{f}_n = 0,$$

and thus, since $\omega(E)\alpha_0(E) = (\omega\alpha_0)(E) = \alpha(E)$ and similarly for $\omega(E)\beta_0(E)$,

$$\alpha(E)\boldsymbol{u}_n - h\beta(E)\boldsymbol{f}_n = 0,$$

as claimed.

9. We have

$$\begin{aligned}\alpha(\zeta) &= \zeta^4 - \zeta^3 = (\zeta-1)(\zeta-1+1)^3 \\ &= (\zeta-1)[(\zeta-1)^3 + 3(\zeta-1)^2 + 3(\zeta-1) + 1],\end{aligned}$$

hence

$$\begin{aligned}\frac{\alpha(\zeta)}{\ln\zeta} &= \frac{\alpha(\zeta)}{\ln(1+\zeta-1)} = \frac{1+3(\zeta-1)+3(\zeta-1)^2+(\zeta-1)^3}{1-\frac{1}{2}(\zeta-1)+\frac{1}{3}(\zeta-1)^2-\cdots} \\ &= 1+\tfrac{7}{2}(\zeta-1)+\tfrac{53}{12}(\zeta-1)^2+\tfrac{55}{24}(\zeta-1)^3+\tfrac{251}{720}(\zeta-1)^4+\cdots\end{aligned}$$

by an elementary computation or using Maple. The β-polynomial of the explicit method therefore is

$$\beta(\zeta) = 1+\tfrac{7}{2}(\zeta-1)+\tfrac{53}{12}(\zeta-1)^2+\tfrac{55}{24}(\zeta-1)^3 = \tfrac{55}{24}\zeta^3 - \tfrac{59}{24}\zeta^2 + \tfrac{37}{24}\zeta - \tfrac{3}{8},$$

and the global error constant (cf. (6.134), (6.135))

$$C_{4,4} = \tfrac{251}{720}.$$

The β-polynomial of the implicit method is sought in the form

$$\beta^*(\zeta) = 1+\tfrac{7}{2}(\zeta-1)+\tfrac{53}{12}(\zeta-1)^2+\tfrac{55}{24}(\zeta-1)^3 + b(\zeta-1)^4,$$

for which

$$C^*_{4,4} = \tfrac{251}{720} - b.$$

The requirement $C^*_{4,4} = -C_{4,4}$ gives

$$\tfrac{251}{720} - b = -\tfrac{251}{720}, \quad b = \tfrac{251}{360},$$

hence

$$\beta^*(\zeta) = \tfrac{251}{360}\zeta^4 - \tfrac{179}{360}\zeta^3 + \tfrac{69}{40}\zeta^2 - \tfrac{449}{360}\zeta + \tfrac{29}{90}.$$

The pair of methods, therefore, is

$$\begin{aligned}\boldsymbol{u}_{n+4} &= \boldsymbol{u}_{n+3} + \frac{h}{24}(55\boldsymbol{f}_{n+3} - 59\boldsymbol{f}_{n+2} + 37\boldsymbol{f}_{n+1} - 9\boldsymbol{f}_n), \\ \boldsymbol{u}^*_{n+4} &= \boldsymbol{u}^*_{n+3} + \frac{h}{360}(251\boldsymbol{f}^*_{n+4} - 179\boldsymbol{f}^*_{n+3} + 621\boldsymbol{f}^*_{n+2} - 449\boldsymbol{f}^*_{n+1} + 116\boldsymbol{f}^*_n).\end{aligned}$$

10. See the text.

11. (a) Let

$$b(z) = b_0 + b_1 z + b_2 z^2 + \cdots + b_{k-1}z^{k-1} + b_k z^k,$$

where $b_k = 0$ for explicit methods. Using the mapping (cf. (6.150))

$$\zeta = \frac{1+z}{1-z}, \quad z = \frac{\zeta - 1}{\zeta + 1}$$

we have

$$\begin{aligned}
\beta(\zeta) &= (\zeta+1)^k b\left(\frac{\zeta-1}{\zeta+1}\right) \\
&= (\zeta+1)^k \left[b_0 + b_1\frac{\zeta-1}{\zeta+1} + \cdots + b_k\left(\frac{\zeta-1}{\zeta+1}\right)^k\right] \\
&= b_0(\zeta+1)^k + b_1(\zeta-1)(\zeta+1)^{k-1} + \cdots + b_k(\zeta-1)^k \\
&= (b_0 + b_1 + \cdots + b_k)\zeta^k + \cdots \ .
\end{aligned}$$

Since $\beta(\zeta)$ is a polynomial of degree $k-1$, we must have $b_0+b_1+\cdots+b_k = 0$, i.e., $b(1) = 0$.

(b) We must show that $p = k+1$ is impossible. Assume, by way of contradiction, that $p = k+1$. Then from (6.155) it follows that

$$b_0 + b_1 + \cdots + b_k = \lambda_0(a_1 + \cdots + a_k) + \lambda_2 s_1 + \cdots + \lambda_{2k'} s_{k'},$$

where $k' = \lfloor \frac{k}{2} \rfloor$ and the s are partial sums of the $a_1, a_2, \ldots, a_k$. Since, by stability, $a_i \geq 0$, and $\sum_{i=1}^k a_i = 1$ (cf. Properties (i) and (iii) in the proof of Theorem 6.4.2), we have $0 \leq s_i \leq 1$. Therefore, using $\lambda_0 = \frac{1}{2}$ and the negativity of the $\lambda_{2\nu}$, $\nu > 0$, we obtain

$$b_0 + b_1 + \cdots + b_k = \tfrac{1}{2} + \lambda_2 s_1 + \cdots + \lambda_{2k'} s_{k'} \geq \tfrac{1}{2} + \lambda_2 + \cdots + \lambda_{2k'}.$$

By (6.160), we have

$$\sum_{\nu=1}^{k'} \lambda_{2\nu} = -\tfrac{1}{2} - \sum_{\nu=k'+1}^{\infty} \lambda_{2\nu} > -\tfrac{1}{2},$$

so that $b_0 + b_1 + \cdots + b_k > 0$, contradicting the assumption that the k-step method is explicit (cf. (a)). Thus, the order cannot be $k+1$ and hence $p \leq k$ as claimed.

12. The change of variables (cf. (6.150))

$$\zeta = \frac{1+z}{1-z}, \quad z = \frac{\zeta - 1}{\zeta + 1}$$

maps the unit disk $|\zeta| \leq 1$ into the left half-plane $\operatorname{Re} z \leq 0$, and the disk Γ_γ into the disk C_γ in the left half-plane, which is symmetric with respect

to the real axis and intersects it in the points $-\frac{1}{\omega}$ and $-\omega$, where $\omega = \frac{1-\gamma}{1+\gamma}$. According to (6.164), we have

$$C_{k,k+1} = 2^{-(k+1)} \frac{b_{k+1}}{b_0},$$

where b_0, b_{k+1} are the coefficients of the powers z^0 and z^{k+1} in the expansion of (6.153). In particular, $b_0 = \frac{1}{2} a_1$. Furthermore, by (6.155) and $2b_0 = a_1$,

$$2^k |C_{k,k+1}| = \begin{cases} |\lambda_{k+1}| + |\lambda_{k-1}| \frac{a_3}{a_1} + \cdots + |\lambda_2| \frac{a_k}{a_1}, & k \text{ odd}, \\ |\lambda_k| \frac{a_2}{a_1} + |\lambda_{k-2}| \frac{a_4}{a_1} + \cdots + |\lambda_2| \frac{a_k}{a_1}, & k \text{ even}. \end{cases}$$

Here, a_i are the (nonnegative) coefficients of the polynomial $a(z)$ in (6.152),

$$a(z) = \left(\frac{1-z}{2}\right)^k \alpha\left(\frac{1+z}{1-z}\right) = a_1 z + a_2 z^2 + \cdots + a_k z^k,$$

and the λ_j the coefficients in the expansion of (6.157). By the assumptions on the roots of α, we have $a_k > 0$. Also, $a(1) = 1$ (cf. Sect. 6.4.2(i)).

Let now

$$z_1 = 0, \quad z_s \in C_\gamma, \quad s = 2, 3, \ldots, k,$$

be the zeros of $a(z)$, and put $u_j = -z_j$ $(j \geq 2)$. Then

$$a(z) = z \frac{\prod_{j=2}^k (z + u_j)}{\prod_{j=2}^k (1 + u_j)}, \quad u_j \in -C_\gamma.$$

We have

$$\prod_{j=2}^k (z + u_j) = \sigma_0(u) z^{k-1} + \sigma_1(u) z^{k-2} + \cdots + \sigma_{k-s}(u) z^{s-1} + \cdots + \sigma_{k-1}(u),$$

where

$$\sigma_0(u) = 1, \ \sigma_1 = u_2 + u_3 + \cdots + u_k, \ \ldots \ , \sigma_{k-1} = u_2 u_3 \cdots u_k$$

are the elementary symmetric functions of $u_2, u_3, \ldots, u_k$. Note that

$$a_1 = \frac{\prod_{j=2}^k u_j}{\prod_{j=2}^k (1 + u_j)} = \frac{\sigma_{k-1}(u)}{\prod_{j=2}^k (1 + u_j)},$$

so that

$$\frac{a_s}{a_1} = \frac{\sigma_{k-s}(u) / \prod_{j=2}^k (1 + u_j)}{\sigma_{k-1}(u) / \prod_{j=2}^k (1 + u_j)} = \frac{\sigma_{k-s}(u)}{\sigma_{k-1}(u)}, \quad s = 1, 2, \ldots, k.$$

Now define $v_j = \frac{1}{u_j},\ j = 2, 3, \ldots, k$. Since the disk $C_\gamma = \{z \in \mathbb{C} : \left|\frac{1+z}{1-z}\right| \leq 1\}$ is invariant with respect to the transformation $z \mapsto \frac{1}{z}$, so is the disk $-C_\gamma$, and hence with $u_j \in -C_\gamma$, also $v_j \in -C_\gamma$ and vice versa. Moreover, one computes

$$\sigma_{s-1}(v) = \frac{\sigma_{k-s}(u)}{\sigma_{k-1}(u)}, \quad s = 1, 2, \ldots, k,$$

so that

$$\frac{a_s}{a_1} = \sigma_{s-1}(v), \quad s = 1, 2, \ldots, k.$$

We have to minimize this ratio over all $v_j \in -C_\gamma$. If among the v_j there are complex conjugate pairs $v_\mu = \xi_\mu \pm \mathrm{i}\eta_\mu$, $\xi_\mu > 0$, $\eta_\mu > 0$, then from the identity

$$\prod_\lambda (z + v_\lambda) \prod_\mu [(z + \xi_\mu)^2 + \eta_\mu^2] = \sum_{s=1}^{k} \sigma_{s-1}(v) z^{k-s},$$

where all v_λ are positive, it follows that each $\sigma_{s-1}(v)$ is a nondecreasing function of the η_μ. We may therefore restrict ourselves to real (positive) v_j. The minimum of $\sigma_{s-1}(v)$ over all $v_j \in -C_\gamma$ is then clearly attained when $v_2 = v_3 = \cdots = v_k = \omega$, independently of s. Moreover, since each $\sigma_{s-1}(v)$ consists of an aggregate of $\binom{k-1}{s-1}$ products of $s-1$ factors v_j, we have

$$\min_{v_j \in -C_\gamma} \sigma_{s-1}(v) = \binom{k-1}{s-1} \omega^{s-1}.$$

It thus follows that

$$\min \frac{a_s}{a_1} = \binom{k-1}{s-1} \omega^{s-1},$$

and therefore, if k is odd,

$$\min |C_{k,k+1}| = 2^{-k} \left\{ |\lambda_{k+1}| + \binom{k-1}{2} \omega^2 |\lambda_{k-1}| + \binom{k-1}{4} \omega^4 |\lambda_{k-3}| + \cdots + \omega^{k-1} |\lambda_2| \right\},$$

and if k is even,

$$\min |C_{k,k+1}| = 2^{-k} \left\{ \binom{k-1}{1} \omega |\lambda_k| + \binom{k-1}{3} \omega^3 |\lambda_{k-2}| + \cdots + \omega^{k-1} |\lambda_2| \right\}.$$

The polynomial $a(z)$ realizing the minimum is

$$a(z) = z \left(\frac{z + \frac{1}{\omega}}{1 + \frac{1}{\omega}} \right)^{k-1},$$

to which there corresponds the characteristic polynomial

$$\alpha(\zeta) = (\zeta - 1)(\zeta + \gamma)^{k-1}.$$

All its zeros other than $\zeta_1 = 1$ are thus concentrated at the point $\zeta = -\gamma$ furthest away from ζ_1.

13. For k even, we have from (6.164)

$$C_{k,k+2} = 2^{-(k+2)} \frac{b_{k+2}}{b_0} ,$$

where, from the proof of Theorem 6.4.2 (cf. (6.155) with $2\nu = k+2$),

$$b_{k+2} = \lambda_4 a_{k-1} + \lambda_6 a_{k-3} + \cdots + \lambda_{k+2} a_1 .$$

Therefore, since $b_0 = \frac{1}{2} a_1$ and $a_s \geq 0$,

$$|C_{k,k+2}| = \frac{1}{2^{k+1} a_1} (|\lambda_4| a_{k-1} + |\lambda_6| a_{k-3} + \cdots + |\lambda_{k+2}| a_1) \geq \frac{|\lambda_{k+2}|}{2^{k+1}} .$$

In the same way as in the proof of part (a), one shows that

$$c_k^* = \inf |C_{k,k+2}| = \frac{|\lambda_{k+2}|}{2^{k+1}} .$$

Suppose now that $|C_{k,k+2}| = c_k^*$ for some stable k-step method. If $k \geq 4$, then necessarily $a_{k-1} = a_{k-3} = \cdots = a_3 = 0$, so that

$$a(z) = a_1 z + a_2 z^2 + a_4 z^4 + \cdots + a_{k-2} z^{k-2} .$$

As in part (a) (or from Theorem 6.4.2), all zeros of $a(z)$ are on the imaginary axis, which implies that $a(z)$ is an odd polynomial. But then $a_2 = a_4 = \cdots = a_{k-2} = 0$, contradicting stability (cf. Sect. 6.4.2(ii)). If $k = 2$, then $a(z) = z$ (since $a_1 = 1$ and $a(z)$ is odd), thus $\alpha(\zeta) = \zeta^2 - 1$, which uniquely gives Simpson's rule, with

$$|C_{2,4}| = \frac{|\lambda_4|}{8} = \frac{1}{180}$$

(cf. Sect. 6.4.1, first Example).

ANSWERS TO MACHINE ASSIGNMENTS

1. (a) The coefficients γ_s are most easily obtained symbolically, using the `taylor` command in the Matlab Symbolic Math toolbox. For example, the first ten of them are obtained as follows:

```
%MAVI_1A
%
syms x
pretty(taylor(-x/((1-x)*log(1-x)),10));
```

The results are incorporated in the vector `g` of the following Matlab routine:

```
 PROGRAM

%AB  kth-order Adams-Bashforth step
%
function unext=AB(ulast,F,k,h)
if k>10
  disp('order k too large')
  return
end
g=[1;1/2;5/12;3/8;251/720;95/288;19087/60480;5257/17280; ...
   1070017/3628800;25713/89600];
Fcopy=F; d=zeros(k,1);
d(1)=F(k);
if k>=2
  for i=1:k-1
    for j=k:-1:i+1
      Fcopy(j)=Fcopy(j)-Fcopy(j-1);
      if j==k, d(i+1)=Fcopy(j); end
    end
  end
end
unext=ulast+h*sum(g(1:k).*d);
```

(b) Symbolic Matlab program for generating the coefficients γ_s^*

```
%MAVI_1B
%
syms x
pretty(taylor(-x/log(1-x),10));

 PROGRAM

%AM  kth-order Adams-Moulton step
%
function unext=AM(ulast,F0,k,h)
if k>10
  disp('order k too large')
return
end
gstar=[1;-1/2;-1/12;-1/24;-19/720;-3/160;-863/60480;-275/24192; ...
   -33953/3628800;-8183/1036800];
F0copy=F0; d=zeros(k,1);
d(1)=F0(k);
if k>=2
  for i=1:k-1
```

```
    for j=k:-1:i+1
      F0copy(j)=F0copy(j)-F0copy(j-1);
      if j==k, d(i+1)=F0copy(j); end
    end
  end
end
unext=ulast+h*sum(gstar(1:k).*d);
```

(c)

```
 PROGRAM

%PECE  Adams predictor/corrector scheme
%
function [xp,up,err,errh]=PECE(f,xin,xfin,uin,k,N,nprint)
snp=size(nprint,1); F=zeros(k,1);
up=zeros(snp,1); err=zeros(snp,1); errh=zeros(snp,1);
u=uin;
h=(xfin-xin)/N;
for s=1:k
  F(s)=f(xin+h*(s-1),u(s));
end
ip=0;
for n=0:N-k
  ulast=u(k);
  u0=AB(ulast,F,k,h);
  F0(1:k-1)=F(2:k); F0(k)=f((n+k)*h,u0);
  unext=AM(ulast,F0,k,h);
  u(1:k-1)=u(2:k); F(1:k-1)=F(2:k);
  u(k)=unext; F(k)=f((n+k)*h,u(k));
  if min(abs(n+k-nprint))==0
    ip=ip+1;
    x=(n+k)*h; xp(ip)=x;
    y=exact(x);
    up(ip)=u(k);
    err(ip)=u(k)-y; errh(ip)=err(ip)/h^k;
  end
end

function y=exact(x)
%
% For MAVI_2C
%
global e eps0
y=kepler(x,x,e,eps0);
%
```

```
% For MAVI_5B
%
%global om
%y=yMAVI_5(x,om);
```

2. (a) With $y = y(x)$ (uniquely) defined by $y - \varepsilon \sin y - x = 0$, we clearly have $y(m\pi) = m\pi$, $m = 0, \pm 1, \pm 2, \ldots$. Differentiating with respect to x, we get $y' - \varepsilon y' \cos y - 1 = 0$, that is,

$$y' = \frac{1}{1 - \varepsilon \cos y} \, .$$

Therefore, $y(x)$ is indeed the solution of the given initial value problem.

(b)

```
 PROGRAMS

%MAVI_2B.m
%
f0='%12.6f %15.12f %11.4e %11.8f\n';
global e
e=.2; eps0=.5e-15;
for k=[2 4 6]
  fprintf('                          k=%2.0f\n',k)
  disp('        x            u          error     error/h^k')
  for N=[40 160 640]
    nprint=[N/4;N/2;3*N/4;N];
    h=2*pi/N;
    u=zeros(k,1);
    for s=1:k
      x=(s-1)*h;
      u(s)=kepler(x,x,e,eps0);
    end
    F=1./(1-e*cos(u));
    for n=0:N-k
      ulast=u(k);
      unext=AB(ulast,F,k,h);
      u(1:k-1)=u(2:k); F(1:k-1)=F(2:k);
      u(k)=unext; F(k)=fMAVI_2((n+k)*h,u(k));
      if min(abs(n+k-nprint))==0
        x=(n+k)*h;
        y=kepler(x,x,e,eps0);
        err=u(k)-y; errh=err/h^k;
        fprintf(f0,x,u(k),err,errh)
      end
    end
```

```
    fprintf('\n')
  end
end

%fMAVI_2
%
function  yprime=fMAVI_2(x,y)
global e
yprime=1/(1-e*cos(y));

 OUTPUT

>> MAVI_2B
                      k= 2
     x           u          error     error/h^k
 1.570796  1.768160505644  1.1999e-03  0.04863002
 3.141593  3.140899369109 -6.9328e-04 -0.02809776
 4.712389  4.513492008144 -2.7327e-03 -0.11075180
 6.283185  6.280684885481 -2.5004e-03 -0.10133827

 1.570796  1.767040839013  8.0231e-05  0.05202626
 3.141593  3.141554926408 -3.7727e-05 -0.02446440
 4.712389  4.516058400758 -1.6630e-04 -0.10783715
 6.283185  6.283065698945 -1.1961e-04 -0.07756063

 1.570796  1.766965704340  5.0964e-06  0.05287618
 3.141593  3.141590387234 -2.2664e-06 -0.02351410
 4.712389  4.516214384529 -1.0315e-05 -0.10701766
 6.283185  6.283178408150 -6.8990e-06 -0.07157942

                      k= 4
     x           u          error     error/h^k
 1.570796  1.767046655579  8.6048e-05  0.14133810
 3.141593  3.141688479616  9.5826e-05  0.15739972
 4.712389  4.516327237729  1.0254e-04  0.16842540
 6.283185  6.283455935379  2.7063e-04  0.44452229

 1.570796  1.766960753588  1.4561e-07  0.06122628
 3.141593  3.141592801745  1.4815e-07  0.06229834
 4.712389  4.516224866046  1.6685e-07  0.07015899
 6.283185  6.283185740009  4.3283e-07  0.18200228

 1.570796  1.766960608298  3.1510e-10  0.03391971
 3.141593  3.141592653907  3.1679e-10  0.03410134
 4.712389  4.516224699585  3.8773e-10  0.04173733
```

```
   6.283185  6.283185308119  9.3936e-10  0.10111890

                      k= 6
       x            u           error     error/h^k
   1.570796  1.766925697082 -3.4911e-05 -2.32403035
   3.141593  3.141568364523 -2.4289e-05 -1.61693130
   4.712389  4.516198155470 -2.6544e-05 -1.76702475
   6.283185  6.283118026605 -6.7281e-05 -4.47889032

   1.570796  1.766960603300 -4.6828e-09 -1.27686618
   3.141593  3.141592650575 -3.0146e-09 -0.82199986
   4.712389  4.516224696484 -2.7130e-09 -0.73975176
   6.283185  6.283185298320 -8.8592e-09 -2.41566930

   1.570796  1.766960607982 -5.0804e-13 -0.56740986
   3.141593  3.141592653590 -2.2382e-13 -0.24997777
   4.712389  4.516224699197 -3.1086e-14 -0.03471913
   6.283185  6.283185307179 -6.5459e-13 -0.73108578
>>
```

The number of Newton iterations for solving Kepler's equation has been observed to be never larger than five.

(c)

```
 PROGRAM

%MAVI_2C
%
f0='%12.6f %15.12f %11.4e %11.8f\n';
global e eps0
%
% Be sure to declare e and eps0 as global variables also
% in the routine PECE.m and its subfunction
%
e=.2; eps0=.5e-15;
xin=0; xfin=2*pi;
for k=[2 4 6]
  fprintf('                          k=%2.0f\n',k)
  disp('        x            u           error     error/h^k')
  for N=[40 160 640]
    h=2*pi/N; uin=zeros(k,1);
    for s=1:k
      x=(s-1)*h;
      uin(s)=kepler(x,x,e,eps0);
    end
    nprint=[N/4;N/2;3*N/4;N];
```

```
    [xp,up,err,errh]=PECE(@fMAVI_2,xin,xfin,uin,k,N,nprint);
    for i=1:4
      fprintf(f0,xp(i),up(i),err(i),errh(i))
    end
    fprintf('\n')
  end
  fprintf('\n')
end
```

OUTPUT

```
>> MAVI_2C
                        k= 2
     x            u          error     error/h^k
 1.570796  1.766802181919 -1.5843e-04 -0.00642077
 3.141593  3.141811415447  2.1876e-04  0.00886608
 4.712389  4.516853298831  6.2860e-04  0.02547618
 6.283185  6.283655827461  4.7052e-04  0.01906947

 1.570796  1.766945965540 -1.4642e-05 -0.00949497
 3.141593  3.141601560605  8.9070e-06  0.00577580
 4.712389  4.516259364793  3.4666e-05  0.02247910
 6.283185  6.283209107804  2.3801e-05  0.01543365

 1.570796  1.766959611165 -9.9682e-07 -0.01034227
 3.141593  3.141593128578  4.7499e-07  0.00492814
 4.712389  4.516226784593  2.0854e-06  0.02163659
 6.283185  6.283186686508  1.3793e-06  0.01431094


                        k= 4
     x            u          error     error/h^k
 1.570796  1.766950098330 -1.0510e-05 -0.01726271
 3.141593  3.141582540508 -1.0113e-05 -0.01661132
 4.712389  4.516213300425 -1.1399e-05 -0.01872314
 6.283185  6.283165832642 -1.9475e-05 -0.03198804

 1.570796  1.766960591730 -1.6253e-08 -0.00683425
 3.141593  3.141592638290 -1.5300e-08 -0.00643351
 4.712389  4.516224681423 -1.7774e-08 -0.00747373
 6.283185  6.283185274793 -3.2387e-08 -0.01361852

 1.570796  1.766960607954 -2.9117e-11 -0.00313439
 3.141593  3.141592653562 -2.8133e-11 -0.00302838
 4.712389  4.516224699162 -3.4588e-11 -0.00372332
```

```
   6.283185  6.283185307109 -7.1005e-11 -0.00764349


                        k= 6
       x            u           error      error/h^k
   1.570796  1.766962249771  1.6418e-06  0.10929439
   3.141593  3.141593789727  1.1361e-06  0.07563301
   4.712389  4.516225928986  1.2298e-06  0.08186748
   6.283185  6.283188374574  3.0674e-06  0.20419746

   1.570796  1.766960608249  2.6655e-10  0.07268181
   3.141593  3.141592653775  1.8487e-10  0.05040765
   4.712389  4.516224699374  1.7740e-10  0.04837284
   6.283185  6.283185307578  3.9805e-10  0.10853774

   1.570796  1.766960607983  2.6645e-14  0.02975926
   3.141593  3.141592653590  1.5099e-14  0.01686358
   4.712389  4.516224699197  5.3291e-15  0.00595185
   6.283185  6.283185307180  3.5527e-14  0.03967901
>>
```

The global error constants for the kth-order Adams predictor and corrector formulae are, respectively,

$$C_{k,k} = \gamma_k, \quad C^*_{k,k} = \gamma^*_k,$$

since $\ell_{k+1} = \gamma_k$ resp. γ^*_k and the sum of the β_s resp. β^*_s in (6.43) is equal to 1 for both formulae (the linear functional L applied to $u(t) = t$ being zero in either case). The errors, scaled or otherwise, of the predictor/corrector scheme, therefore, should be approximately equal to (γ^*_k/γ_k) times the errors of the Adams–Bashforth method. We check this for each $k = 2, 4, 6$ by looking at the third group of results (corresponding to $h = 2\pi/640$) in the outputs for `MAVI_2B` resp. `MAVI_2C`. The result of this check is summarized in the table below.

k	errh(AB)	errh(PECE)	$(\gamma_k^*/\gamma_k)\times$errh(AB)
2	0.0529	−0.0103	−0.0106
	−0.0235	0.0049	0.0047
	−0.1070	0.0216	0.0214
	−0.0716	0.0143	0.0143
4	0.0339	−0.0031	−0.0026
	0.0341	−0.0030	−0.0026
	0.0417	−0.0037	−0.0032
	0.1011	−0.0076	−0.0077
6	−0.5674	0.0298	0.0257
	−0.2500	0.0169	0.0113
	−0.0347	0.0060	0.0016
	−0.7311	0.0397	0.0331

It is seen that agreement between the entries in the last two columns is generally as good as can be expected.

3. (a) According to Theorem 6.4.1 we must determine the coefficients of the polynomial $\beta(\zeta) = \sum_{s=0}^{k-1} \beta_{k,s}\zeta^s$ such that

$$\delta(\zeta) = \frac{\zeta^k - \zeta^{k-1}}{\ln \zeta} - \beta(\zeta)$$

has a zero at $\zeta = 1$ of multiplicity k. This requires us to expand

$$\frac{\zeta^k - \zeta^{k-1}}{\ln \zeta} = c_0 + c_1(\zeta - 1) + c_2(\zeta - 1)^2 + \cdots ,$$

take for $\beta(\zeta)$ the first k terms on the right (up to and including the power $(\zeta - 1)^{k-1}$), and re-express it in terms of powers of ζ. This is accomplished by the following Maple program.

```
 PROGRAM

for k from 1 to 10 do
  Order:=k+1;
  s:=series((x^k-x^(k-1))/log(x),x=1);
  p:=convert(s,polynom);
  AdamsBashforth:=simplify(p);
od;
```

The results are displayed in the table below.

k	$\beta_{k,s},\ s=0,1,\ldots,k-1$					
1	1					
2	$-\frac{1}{2}$	$\frac{3}{2}$				
3	$\frac{5}{12}$	$-\frac{4}{3}$	$\frac{23}{12}$			
4	$-\frac{3}{8}$	$\frac{37}{24}$	$-\frac{59}{24}$	$\frac{55}{24}$		
5	$\frac{251}{720}$	$-\frac{637}{360}$	$\frac{109}{30}$	$-\frac{1387}{360}$	$\frac{1901}{720}$	
6	$-\frac{95}{288}$	$\frac{959}{480}$	$-\frac{3649}{720}$	$\frac{4991}{720}$	$-\frac{2641}{480}$	$\frac{4277}{1440}$
7	$\frac{19087}{60480}$	$-\frac{5603}{2520}$	$\frac{135713}{20160}$	$-\frac{10754}{945}$	$\frac{235183}{20160}$	$-\frac{18637}{2520}$
		$\frac{198721}{60480}$				
8	$-\frac{5257}{17280}$	$\frac{32863}{13440}$	$-\frac{115747}{13440}$	$\frac{2102243}{120960}$	$-\frac{296053}{13440}$	$\frac{242653}{13440}$
		$-\frac{1152169}{120960}$	$\frac{16083}{4480}$			
9	$\frac{1070017}{3628800}$	$-\frac{4832053}{1814400}$	$\frac{19416743}{1814400}$	$-\frac{45586321}{1814400}$	$\frac{862303}{22680}$	$-\frac{69927631}{1814400}$
		$\frac{47738393}{1814400}$	$-\frac{21562603}{1814400}$	$\frac{14097247}{3628800}$		
10	$-\frac{25713}{89600}$	$\frac{20884811}{7257600}$	$-\frac{2357683}{181440}$	$\frac{15788639}{453600}$	$-\frac{222386081}{3628800}$	$\frac{269181919}{3628800}$
		$-\frac{28416361}{453600}$	$\frac{6648317}{181440}$	$-\frac{104995189}{7257600}$	$\frac{4325321}{1036800}$	

(b) Similarly as in (a), but with $\beta^*(\zeta)=\sum_{s=1}^{k}\beta^*_{k,s}\zeta^s$, we must have

$$\frac{\zeta^k-\zeta^{k-1}}{\ln\zeta}-\beta^*(\zeta)=O((\zeta-1)^k),\quad \zeta\to 1,$$

or, equivalently, dividing by ζ,

$$\frac{\zeta^k-\zeta^{k-1}}{\zeta\ln\zeta}-\sum_{s=1}^{k}\beta^*_{k,s}\zeta^{s-1}=O((\zeta-1)^k).$$

Thus, if

$$\frac{\zeta^k-\zeta^{k-1}}{\zeta\ln\zeta}=d_0+d_1(\zeta-1)+d_2(\zeta-1)^2+\cdots\ ,$$

the polynomial $\zeta^{-1}\beta^*(\zeta)$ of degree $k-1$ must be identified with the first k terms on the right of the expansion and then re-expressed in terms of powers of ζ. This is done by the following program in Maple.

```
PROGRAM

for k from 1 to 10
  Order:=k+1;
```

```
    s:=series((x^k-x^(k-1))/(x*log(x)),x=1);
    p:=convert(s,polynom);
    AdamsMoulton:=simplify(p);
od;
```

The results are displayed in the table below.

k	$\beta^*_{k,s},\ s=1,\dots,k$					
1	1					
2	$\frac{1}{2}$	$\frac{1}{2}$				
3	$-\frac{1}{12}$	$\frac{2}{3}$	$\frac{5}{12}$			
4	$\frac{1}{24}$	$-\frac{5}{24}$	$\frac{19}{24}$	$\frac{3}{8}$		
5	$-\frac{19}{720}$	$\frac{53}{360}$	$-\frac{11}{30}$	$\frac{323}{360}$	$\frac{251}{720}$	
6	$\frac{3}{160}$	$-\frac{173}{1440}$	$\frac{241}{720}$	$-\frac{133}{240}$	$\frac{1427}{1440}$	$\frac{95}{288}$
7	$-\frac{863}{60480}$	$\frac{263}{2520}$	$-\frac{6737}{20160}$	$\frac{586}{945}$	$-\frac{15487}{20160}$	$\frac{2713}{2520}$
		$\frac{19087}{60480}$				
8	$\frac{275}{24192}$	$-\frac{11351}{120960}$	$\frac{1537}{4480}$	$-\frac{88547}{120960}$	$\frac{123133}{120960}$	$-\frac{4511}{4480}$
		$\frac{139849}{120960}$	$\frac{5257}{17280}$			
9	$-\frac{33953}{3628800}$	$\frac{156437}{1814400}$	$-\frac{645607}{1814400}$	$\frac{1573169}{1814400}$	$-\frac{31457}{22680}$	$\frac{2797679}{1814400}$
		$-\frac{2302297}{1814400}$	$\frac{2233547}{1814400}$	$\frac{1070017}{3628800}$		
10	$\frac{8183}{1036800}$	$-\frac{116687}{1451520}$	$\frac{335983}{907200}$	$-\frac{462127}{453600}$	$\frac{6755041}{3628800}$	$-\frac{8641823}{3628800}$
		$\frac{200029}{90720}$	$-\frac{1408913}{907200}$	$\frac{9449717}{7257600}$	$\frac{25713}{89600}$	

4. (a) For the model problem $y' = \lambda y$, $\operatorname{Re}\lambda < 0$, the kth-order Adams–Moulton method takes the form (cf. (6.175))

$$u_{n+k} - u_{n+k-1} - \lambda h \sum_{s=1}^{k} \beta^*_{k,s} u_{n+s} = 0.$$

For numerical values of the coefficients $\beta^*_{k,s}$, $1 \le k \le 10$, see MA 3(b). The methods of order 1 and 2 are respectively the implicit Euler method and the trapezoidal rule, both being A-stable (cf. Ch. 5, Sect. 5.9.3(1),(2)). We may assume, therefore, that $k > 2$.

The above equation represents a linear homogeneous kth-order difference equation with constant coefficients,

$$(1 - z\beta^*_{k,k})u_{n+k} - (1 + z\beta^*_{k,k-1})u_{n+k-1} - \sum_{s=1}^{k-2} z\beta^*_{k,s} u_{n+s} = 0,$$

where $z = \lambda h$. Its characteristic polynomial is

$$\tilde{\alpha}(t) = (1 - z\beta^*_{k,k})t^k - (1 + z\beta^*_{k,k-1})t^{k-1} - \sum_{s=1}^{k-2} z\beta^*_{k,s}t^s.$$

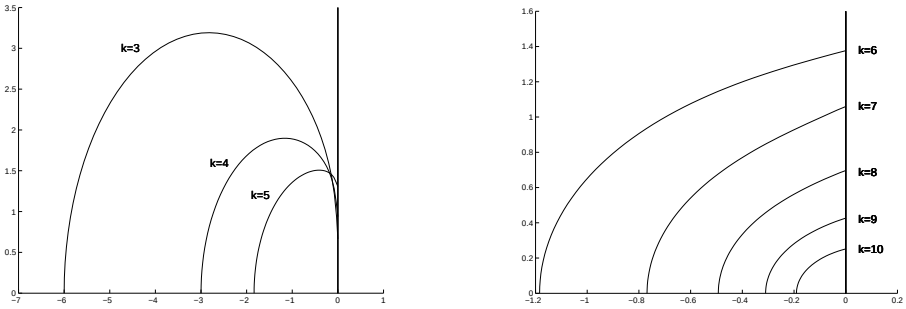
All solutions of the difference equation tend to zero if and only if the modulus of the absolutely largest zero $t_1(z)$ of $\tilde{\alpha}(t)$ is less than 1. This defines the region $\mathcal{D}_A$ of absolute stability of the Adam–Moulton method. The routine `MAVI_4A.m` below computes, for $3 \le k \le 10$, the upper boundary of $\mathcal{D}_A$ using polar coordinates, the method of bisection, and the routine `AMsr.m`, which computes $|t_1(z)|$—the "spectral radius" of the method. The lower boundary is the symmetric (with respect to the real axis) image of the upper boundary. For literature, see Ch. V.1 in E. Hairer and G. Wanner, *Solving ordinary differential equations. II: Stiff and differential-algebraic problems*, 2d rev. ed., Springer Ser. Comput. Math., v. 14, Springer, Berlin, where, however, the Adams corrector is of order $k + 1$ (not k as assumed here).

PROGRAMS

```
%MAVI_4A
%
f0='%8.0f %16.12f\n';
N=50;
tol=.5e-12; th=zeros(N+1,1); rr=zeros(N+1,1);
hold on
for k=3:5
%for k=6:10
  for ith=1:N+1
    theta=(pi/2)*(1+(ith-1)/N); th(ith)=theta;
    r=0; rho=0;
    while rho<1
      r=r+.1; z=r*exp(i*theta);
      rho=AMsr(z,k);
    end
    rlow=r-.1; rhigh=r;
    ntol=ceil(log((rhigh-rlow)/tol)/log(2));
    for n=1:ntol
      rs=(rlow+rhigh)/2; zs=rs*exp(i*theta);
      rho=AMsr(zs,k);
      if rho<1
        rlow=rs;
      else
        rhigh=rs;
      end
```

```
    end
    rr(ith)=rs;
    if ith==N+1
      fprintf(f0,k,rr(ith))
    end
  end
  if k==7 | k==8
    rr(1)=2*rr(2)-rr(3);
  end
  plot(rr.*cos(th),rr.*sin(th))
  plot ([0 0],[0 3.5])
%  plot([0 0],[0 1.6])
end
hold off

%AMsr  Adams-Moulton spectral radius
%
function rho=AMsr(z,k)
bstar=[1 0 0 0 0 0 0 0 0 0;1/2 1/2 0 0 0 0 0 0 0 0;
   -1/12 2/3 5/12 0 0 0 0 0 0 0;1/24 -5/24 19/24 3/8 ...
   0 0 0 0 0 0;-19/720 53/360 -11/30 323/360 251/720 ...
   0 0 0 0 0;3/160 -173/1440 241/720 -133/240 1427/1440 ...
   95/288 0 0 0 0;-863/60480 263/2520 -6737/20160 586/945 ...
   -15487/20160 2713/2520 19087/60480 0 0 0;275/24192 ...
   -11351/120960 1537/4480 -88547/120960 123133/120960 ...
   -4511/4480 139849/120960 5257/17280 0 0;-33953/3628800 ...
   156437/1814400 -645607/1814400 1573169/1814400 ...
   -31457/22680 2797679/1814400 -2302297/1814400 ...
   2233547/1814400 1070017/3628800 0;8183/1036800 ...
   -116687/1451520 335983/907200 -462127/453600 ...
   6755041/3628800 -8641823/3628800 200029/90720 ...
   -1408913/907200 9449717/7257600 25713/89600];
if k<=2 | k>10
  disp('k in AMsr not in range')
  return
end
kk=k-2:-1:1;
p=[1-z*bstar(k,k) -1-z*bstar(k,k-1) -z*bstar(k,kk)];
r=roots(p); rho=max(abs(r));
```

PLOTS



On the imaginary axis, the spectral radius $|t_1(\mathrm{i}y)|$ usually decreases from 1 (when $y = 0$) to values less than 1 when y is small positive, but not when $k = 7$ or $k = 8$, in which cases it initially increases, then turns around to values less than 1 before becoming larger than 1 again. In these two cases, our routine simply interpolates linearly from the two preceding points on the boundary curve to determine the point of the curve on the imaginary axis.

OUTPUT

```
>> MAVI_4A
       k   stability abscissa
       3    -6.000000000000
       4    -3.000000000000
       5    -1.836734693878
       6    -1.184210526316
       7    -0.768605124034
       8    -0.492957746479
       9    -0.309961405158
      10    -0.190595923197
>>
```

(b) Similarly as in (a), the kth-order Adams predictor/corrector (PECE) method, applied to the model problem, produces the difference equation

$$u_{n+1} = (1 + z + z^2)u_n \quad \text{if } k = 1,$$

and

$$\begin{aligned} u_{n+k} = {} & [1 + z(\beta^*_{k,k} + \beta^*_{k,k-1}) + z^2\beta^*_{k,k}\beta_{k,k-1}]u_{n+k-1} \\ & + \sum_{s=1}^{k-2}(z\beta^*_{k,s} + z^2\beta^*_{k,k}\beta_{k,s})u_{n+s} + z^2\beta^*_{k,k}\beta_{k,0}u_n \quad \text{if } k > 1, \end{aligned}$$

where the empty sum, when $k = 2$, is to be taken as zero. The characteristic polynomial, therefore, is

$$\tilde{a}(t) = \begin{cases} t - (1 + z + z^2) \quad \text{if } k = 1, \\ t^k - [1 + z(\beta^*_{k,k} + \beta^*_{k,k-1}) + z^2 \beta^*_{k,k} \beta_{k,k-1}] t^{k-1} \\ \qquad - \sum_{s=1}^{k-2} (z\beta^*_{k,s} + z^2 \beta^*_{k,k} \beta_{k,s}) t^{n+s} - z^2 \beta^*_{k,k} \beta_{k,0} \quad \text{if } k > 1. \end{cases}$$

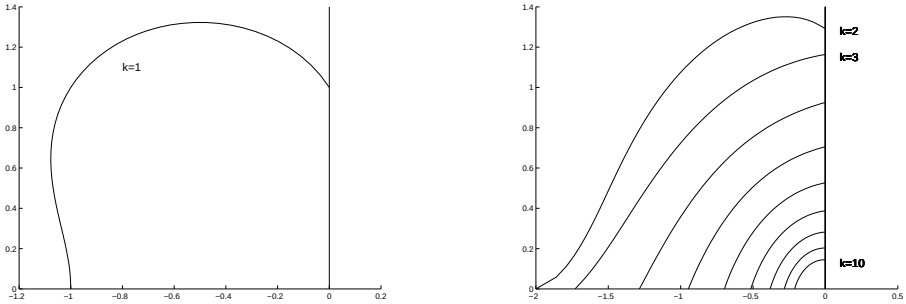
Here, as in (a), $z = h\lambda$ and $\beta^*_{k,s}$ are the coefficients of the kth-order Adams–Moulton formula; the $\beta_{k,s}$ are the coefficients of the kth-order Adams–Bashforth method, for which numerical values are provided in MA 3(a). The upper boundary of the regions $\mathcal{D}_\mathcal{A}$ of absolute stability are now computed for $k = 1 : 10$ by the routine `MAVI_4B.m` calling upon the routine `PECEsr.m` for computing the spectral radius of the Adams PECE method.

```
 PROGRAM

%MAVI_4B
%
f0='%8.0f %18.12f\n';
N=50;
tol=.5e-12; th=zeros(N+1,1); rr=zeros(N+1,1);
hold on
disp('       k   stability interval')
for k=1
%for k=2:10
  for ith=1:N+1
    theta=(pi/2)*(1+(ith-1)/N); th(ith)=theta;
    r=0; rho=0;
    while rho<1
      r=r+.1; z=r*exp(i*theta);
      rho=PECEsr(z,k);
    end
    rlow=r-.1; rhigh=r;
    ntol=ceil(log((rhigh-rlow)/tol)/log(2));
    for n=1:ntol
      rs=(rlow+rhigh)/2; zs=rs*exp(i*theta);
      rho=PECEsr(zs,k);
      if rho<1
        rlow=rs;
      else
        rhigh=rs;
      end
    end
```

```
    rr(ith)=rs;
    if ith==N+1
      fprintf(f0,k,rr(ith))
    end
  end
  if k>1
    rr(1)=2*rr(2)-rr(3);
  end
  plot(rr.*cos(th),rr.*sin(th))
  plot([0 0],[0 1.4])
end
hold off

%PECEsr  PECE spectral radius
%
function rho=PECEsr(z,k)
b=[1 0 0 0 0 0 0 0 0 0;-1/2 3/2 0 0 0 0 0 0 0 0;
   5/12 -4/3 23/12 0 0 0 0 0 0 0;-3/8 37/24 -59/24 ...
   55/24 0 0 0 0 0 0;251/720 -637/360 109/30 -1387/360 ...
   1901/720 0 0 0 0 0;-95/288 959/480 -3649/720 4991/720 ...
   -2641/480 4277/1440 0 0 0 0;19087/60480 -5603/2520 ...
   135713/20160 -10754/945 235183/20160 -18637/2520 ...
   198721/60480 0 0 0;-5257/17280 32863/13440 ...
   -115747/13440 2102243/120960 -296053/13440 ...
   242653/13440 -1152169/120960 16083/4480 0 0; ...
   1070017/3628800 -4832053/1814400 19416743/1814400 ...
   -45586321/1814400 862303/22680 -69927631/1814400 ...
   47738393/1814400 -21562603/1814400 14097247/3628800 ...
   0;-25713/89600 20884811/7257600 -2357683/181440 ...
   15788639/453600 -222386081/3628800 269181919/3628800 ...
   -28416361/453600 6648317/181440 -104995189/7257600 ...
   4325321/1036800];
bstar=[1 0 0 0 0 0 0 0 0 0;1/2 1/2 0 0 0 0 0 0 0 0;
   -1/12 2/3 5/12 0 0 0 0 0 0 0;1/24 -5/24 19/24 3/8 ...
   0 0 0 0 0 0;-19/720 53/360 -11/30 323/360 251/720 ...
   0 0 0 0 0;3/160 -173/1440 241/720 -133/240 1427/1440 ...
   95/288 0 0 0 0;-863/60480 263/2520 -6737/20160 586/945 ...
   -15487/20160 2713/2520 19087/60480 0 0 0;275/24192 ...
   -11351/120960 1537/4480 -88547/120960 123133/120960 ...
   -4511/4480 139849/120960 5257/17280 0 0;-33953/3628800 ...
   156437/1814400 -645607/1814400 1573169/1814400 ...
   -31457/22680 2797679/1814400 -2302297/1814400 ...
   2233547/1814400 1070017/3628800 0;8183/1036800 ...
   -116687/1451520 335983/907200 -462127/453600 ...
   6755041/3628800 -8641823/3628800 200029/90720 ...
```

```
   -1408913/907200 9449717/7257600 25713/89600];
if k>10
  disp('k in PECEsr not in range')
  return
end
if k==1
  p=[1 -(1+z+z^2)];
elseif k==2
  p=[1 -1-z*(bstar(2,2)+bstar(2,1)+z*bstar(2,2)*b(2,2)) ...
     -z^2*bstar(2,2)*b(2,1)];
else
  kk=k-2:-1:1;
  p=[1 -1-z*(bstar(k,k)+bstar(k,k-1)+z*bstar(k,k)*b(k,k)) ...
     -z*(bstar(k,kk)+z*bstar(k,k)*b(k,kk+1)) -z^2*bstar(k,k)*b(k,1)];
end
r=roots(p); rho=max(abs(r));
```

PLOTS



On the imaginary axis, the same corrective measure was taken for all $k > 1$ as mentioned in (a).

OUTPUT

```
>> MAVI_4B
       k   stability interval
       1    -1.000000000000
       2    -1.999999999961
       3    -1.728783568074
       4    -1.284816263107
       5    -0.946917034537
       6    -0.698002629549
       7    -0.515315925510
       8    -0.381569095042
       9    -0.283920041049
      10    -0.212824229008
```

```
>>
```

One could think that the stability properties of the Adams predictor/corrector method are similar to those of the Adams corrector formula. But this is far from the truth, as our plots and stability intervals show when compared with those for the Adams–Moulton method in part (a).

5. (a) Standard rules of integration yield

$$y(x) = y_0 \mathrm{e}^{-\omega x} + \omega \int_0^x \mathrm{e}^{-\omega(x-t)} a(t) \mathrm{d}t.$$

Thus,

$$y(x) = \frac{2}{\omega^2}(1 - \mathrm{e}^{-\omega x}) + x\left(x - \frac{2}{\omega}\right) \qquad \text{in case (i)},$$

$$y(x) = \begin{cases} \dfrac{1}{1-\omega}(\mathrm{e}^{-\omega x} - \omega \mathrm{e}^{-x}) & \text{if } \omega \neq 1 \\ (1+x)\mathrm{e}^{-x} & \text{if } \omega = 1 \end{cases} \qquad \text{in case (ii)},$$

$$y(x) = \frac{1}{1+\omega}(\mathrm{e}^{-\omega x} + \omega \mathrm{e}^{x}) \qquad \text{in case (iii)}.$$

(b)

```
 PROGRAMS

%MAVI_5B
%
f0='%12.6f %20.12e %11.4e %9.4f\n';
f1='%12.6f %20.12e %11.4e\n';
global om
%
% Be sure to declare om as a global variable also
% in the routine PECE.m and its subfunction
%
xs=[1;2;1.728783568074;1.284816263107;0.946917034537; ...
    0.698002629549;0.515315925510;0.381569095042; ...
    0.283920041049;0.212824229008];
om=100;
xin=0; xfin=1;
for k=[1 4 7]
  fprintf('                      k=%1.0f\n',k)
  disp('       x               u             error     stab ratio')
  for N=[40 80 160 320]
    h=1/N; uin=zeros(k,1);
```

```
    stabratio=om*h/xs(k);
    for s=1:k
      x=(s-1)*h;
      uin(s)=yMAVI_5(x,om);
     end
    nprint=[N/4;N/2;3*N/4;N];
    [xp,up,err]=PECE(@fMAVI_5,xin,xfin,uin,k,N,nprint);
    for i=1:4
      if i==1
        fprintf(f0,xp(i),up(i),err(i),stabratio)
      else
        fprintf(f1,xp(i),up(i),err(i))
      end
    end
    fprintf('\n')
  end
  fprintf('\n')
end

%yMAVI_5
%
function y=yMAVI_5(x,om)
y=(2/om^2)*(1-exp(-om*x))+x*(x-2/om);
% if om==1
%   y=(1+x)*exp(-x);
% else
%   y=(exp(-om*x)-om*exp(-x))/(1-om);
% end
% y=(exp(-om*x)+om*exp(x))/(1+om);

%fMAVI_5
%
function yprime=fMAVI_5(x,y)
global om
yprime=-om*(y-a(x));

function y=a(x)
y=x^2;
% y=exp(-x);
% y=exp(x);
```

Reasonable results can be expected only if ωh lies in the interval of absolute stability, $-x_s < x < 0$, of the Adams predictor/corrector method. For each order k, $1 \leq k \leq 10$, this interval is determined in MA 4 (see the output of `MAVI_4B`). The "stability ratio" $\rho_s = \omega h / x_s$

is a good indicator of absolute stability: it has to be less than 1 for the method to be absolutely stable. In the output below, for $\omega = 100$ and case (i), the ratio ρ_s is printed in the last column. It can be seen that the errors are as expected as long as $\rho_s < 1$. Even if ρ_s is quite close to 1, but smaller, the method works well; see, e.g., $k = 4$ and the second group of results (corresponding to $N = 80$). Similar results are observed in cases (ii) and (iii).

```
   OUTPUT (for omega=100 and case (i))

>> MAVI_5B
                    k=1
      x               u              error     stab ratio
   0.250000   2.241422611862e+03  2.2414e+03    2.5000
   0.500000   1.310535465169e+10  1.3105e+10
   0.750000   7.662753839492e+16  7.6628e+16
   1.000000   4.480443144526e+23  4.4804e+23

   0.250000   2.694516240765e-01  2.1175e-01    1.2500
   0.500000   4.922983940885e+01  4.8990e+01
   0.750000   1.127512877855e+04  1.1275e+04
   1.000000   2.594698273522e+06  2.5947e+06

   0.250000   5.797082253115e-02  2.7082e-04    0.6250
   0.500000   2.404708333331e-01  2.7083e-04
   0.750000   5.479708333333e-01  2.7083e-04
   1.000000   9.804708333333e-01  2.7083e-04

   0.250000   5.775965908988e-02  5.9659e-05    0.3125
   0.500000   2.402596590909e-01  5.9659e-05
   0.750000   5.477596590909e-01  5.9659e-05
   1.000000   9.802596590909e-01  5.9659e-05


                    k=4
      x               u              error     stab ratio
   0.250000   5.489782225988e-02 -2.8022e-03    1.9458
   0.500000   1.069335362243e-01 -1.3327e-01
   0.750000  -3.468632312304e+00 -4.0163e+00
   1.000000  -5.054337583214e+01 -5.1524e+01

   0.250000   5.770301912296e-02  3.0191e-06    0.9729
   0.500000   2.402022068408e-01  2.2068e-06
   0.750000   5.477015817582e-01  1.5818e-06
```

```
1.000000   9.802011160529e-01  1.1161e-06

0.250000   5.770000000000e-02  2.2898e-15    0.4865
0.500000   2.402000000000e-01  2.7756e-17
0.750000   5.477000000000e-01  1.1102e-16
1.000000   9.802000000000e-01  0.0000e+00

0.250000   5.770000000000e-02  7.6328e-17    0.2432
0.500000   2.402000000000e-01  2.7756e-17
0.750000   5.477000000000e-01  1.1102e-16
1.000000   9.802000000000e-01  2.2204e-16


                  k=7
   x              u             error     stab ratio
0.250000   5.876489685721e-02  1.0649e-03    4.8514
0.500000   8.510479152234e+00  8.2703e+00
0.750000   6.348404562388e+04  6.3483e+04
1.000000   4.825624880256e+08  4.8256e+08

0.250000   5.672379140550e-02 -9.7621e-04    2.4257
0.500000  -1.319529910546e+02 -1.3219e+02
0.750000  -1.033046220926e+07 -1.0330e+07
1.000000  -5.806582882782e+11 -5.8066e+11

0.250000   5.770053580764e-02  5.3581e-07    1.2128
0.500000   2.402592888599e-01  5.9289e-05
0.750000   5.542115179622e-01  6.5115e-03
1.000000   1.689315197964e+00  7.0912e-01

0.250000   5.770000000000e-02  0.0000e+00    0.6064
0.500000   2.402000000000e-01  0.0000e+00
0.750000   5.477000000000e-01  1.1102e-16
1.000000   9.802000000000e-01  1.1102e-16
>>
```

(c)

```
 PROGRAM

%MAVI_5C
%
f0='%12.6f %20.12e %11.4e\n';
alpha=[-1 0 0 0 0 0 0;1/3 -4/3 0 0 0 0 0; ...
  -2/11 9/11 -18/11 0 0 0 0;3/25 -16/25 36/25 -48/25 ...
  0 0 0;-12/137 75/137 -200/137 300/137 -300/137 0  0; ...
```

```
    10/147 -24/49 75/49 -400/147 150/49 -120/49 0; ...
    -20/363 490/1089 -196/121 1225/363 -4900/1089 490/121 ...
    -980/363];
  om=100;
  for k=[1 4 7]
    fprintf('                        k=%1.0f\n',k)
    disp('       x               u             error')
    r=(1:k)'; u=zeros(k,1);
    alphak=sum(1./r);
    for N=[40 80 160 320]
      nprint=[N/4;N/2;3*N/4;N];
      h=1/N; x=(r-1)*h;
      for s=1:k
        u(s)=yMAVI_5((s-1)*h,om);
      end
      for n=0:N-k
        x=(n+k)*h;
        a=x^2;
  %      a=exp(-x);
  %      a=exp(x);
        unext=(h*om*a/alphak-sum(alpha(k,1:k)'.*u))/(1+h*om/alphak);
        u(1:k-1)=u(2:k); u(k)=unext;
        if min(abs(n+k-nprint))==0
          y=yMAVI_5(x,om); err=u(k)-y;
          fprintf(f0,x,u(k),err)
        end
      end
      fprintf('\n')
    end
  end
```

OUTPUT (for omega=100 and case (i))

```
>> MAVI_5C
                    k=1
       x               u              error
    0.250000   5.794999836871e-02  2.5000e-04
    0.500000   2.404500000000e-01  2.5000e-04
    0.750000   5.479500000000e-01  2.5000e-04
    1.000000   9.804500000000e-01  2.5000e-04

    0.250000   5.782499997061e-02  1.2500e-04
    0.500000   2.403250000000e-01  1.2500e-04
    0.750000   5.478250000000e-01  1.2500e-04
    1.000000   9.803250000000e-01  1.2500e-04
```

```
0.250000   5.776249999903e-02  6.2500e-05
0.500000   2.402625000000e-01  6.2500e-05
0.750000   5.477625000000e-01  6.2500e-05
1.000000   9.802625000000e-01  6.2500e-05

0.250000   5.773124999992e-02  3.1250e-05
0.500000   2.402312500000e-01  3.1250e-05
0.750000   5.477312500000e-01  3.1250e-05
1.000000   9.802312500000e-01  3.1250e-05

                k=4
   x              u              error
0.250000   5.770023904337e-02  2.3904e-07
0.500000   2.401999975294e-01 -2.4706e-09
0.750000   5.476999999851e-01 -1.4901e-11
1.000000   9.802000000002e-01  1.9496e-13

0.250000   5.770000142555e-02  1.4256e-09
0.500000   2.402000000002e-01  1.7694e-13
0.750000   5.477000000000e-01  1.1102e-16
1.000000   9.802000000000e-01  2.2204e-16

0.250000   5.770000000000e-02  7.7646e-15
0.500000   2.402000000000e-01  2.7756e-17
0.750000   5.477000000000e-01  0.0000e+00
1.000000   9.802000000000e-01  0.0000e+00

0.250000   5.770000000000e-02  2.5674e-16
0.500000   2.402000000000e-01  2.7756e-16
0.750000   5.477000000000e-01  5.5511e-16
1.000000   9.802000000000e-01  4.4409e-16

                k=7
   x              u              error
0.250000   5.770218861439e-02  2.1886e-06
0.500000   2.402039955053e-01  3.9955e-06
0.750000   5.476860132559e-01 -1.3987e-05
1.000000   9.801686860942e-01 -3.1314e-05

0.250000   5.769883763523e-02 -1.1624e-06
0.500000   2.401990580251e-01 -9.4197e-07
0.750000   5.477456741724e-01  4.5674e-05
1.000000   9.804936992562e-01  2.9370e-04
```

```
    0.250000   5.769995521342e-02 -4.4787e-08
    0.500000   2.402002140664e-01  2.1407e-07
    0.750000   5.477226579333e-01  2.2658e-05
    1.000000   9.808993567523e-01  6.9936e-04

    0.250000   5.770000134209e-02  1.3421e-09
    0.500000   2.402000900194e-01  9.0019e-08
    0.750000   5.476843474295e-01 -1.5653e-05
    1.000000   9.816479690515e-01  1.4480e-03
>>
```

As expected, the backward differentiation method gives unconditionally accurate results when $1 \leq k \leq 6$, reflecting $A(\alpha)$ stability of the method; cf. Sect. 6.5.2, Table 6.1. For $k = 7$, the accuracy is significantly worse because of lack of strong stability. The errors are not catastrophically large, however, since the root condition is just barely violated, the absolutely largest root having modulus 1.0222; cf. Ex. 10(a). The results in the cases (ii) and (iii) are qualitatively the same.

6. See the text.

EXERCISES AND MACHINE ASSIGNMENTS TO CHAPTER 7

EXERCISES

1. Consider the nonlinear boundary value problem (Blasius equation)

$$y''' + \tfrac{1}{2} y y'' = 0, \quad 0 \le x < \infty,$$

$$y(0) = y'(0) = 0, \quad y'(\infty) = 1.$$

   (a) Letting $y''(0) = \lambda$ and $z(t) = \lambda^{-\frac{1}{3}} y(\lambda^{-\frac{1}{3}} t)$ (assuming $\lambda > 0$), derive an initial value problem for z on $0 \le t < \infty$.

   (b) Explain, and illustrate numerically and graphically, how the solution of the initial value problem in (a) can be used to obtain the solution $y(x)$ of the given boundary value problem.

2. The boundary value problem

$$y'' = -\frac{1}{x}\, y y', \quad 0 < x \le 1; \quad y(0) = 0,\ y(1) = 1,$$

   although it has a singularity at $x = 0$ and certainly does not satisfy (7.112), has the smooth solution $y(x) = 2x/(1 + x)$.

   (a) Determine analytically the s-interval for which the initial value problem

$$u'' = -\frac{1}{x}\, u u', \quad 0 < x \le 1; \quad u(0) = 0,\ u'(0) = s$$

   has a smooth solution $u(x; s)$ on $0 \le x \le 1$.

   (b) Determine the s-interval for which Newton's method applied to $u(1; s) - 1 = 0$ converges.

3. Use Matlab to reproduce the results in Table 7.2 and to prepare plots of the four solution components.

4. Derive (7.56) and (7.62).

5. Let

$$\phi(s) = \frac{s}{s \cosh 1 - \sinh 1} - e$$

   and s^0 be the solution of

$$s^0 - \frac{\phi(s^0)}{\phi'(s^0)} = \tanh 1$$

   (cf. the Example of Sect. 7.2.2, (7.58) and (7.62)).

(a) Show that $s^0 < \coth 1$. {*Hint*: consider what $s^0 < t^0$ means in terms of one Newton step at t^0 for the equation $\phi(s) = 0$.}

(b) Use the bisection method to compute s^0 to six decimal places. What are appropriate initial approximations?

6. Generalizing the Example of Sect. 7.2.2, let $\nu > 0$ and consider the boundary value problem

$$\left.\begin{aligned} \frac{\mathrm{d}y_1}{\mathrm{d}x} &= \frac{y_1^{\nu+1}}{y_2^{\nu}} \\ \frac{\mathrm{d}y_2}{\mathrm{d}x} &= \frac{y_2^{\nu+1}}{y_1^{\nu}} \end{aligned}\right\} \quad 0 \le x \le 1,$$

$$y_1(0) = 1, \quad y_1(1) = \mathrm{e}.$$

(a) Determine the exact solution.

(b) Solve the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} u_1^{\nu+1}/u_2^{\nu} \\ u_2^{\nu+1}/u_1^{\nu} \end{bmatrix}, \quad 0 \le x \le 1; \quad \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}(0) = \begin{bmatrix} 1 \\ s \end{bmatrix}$$

in closed form.

(c) Find the conditions on $s > 0$ guaranteeing that $u_1(x)$, $u_2(x)$ both remain positive and finite on $[0, 1]$. In particular, show that, as $\nu \to \infty$, the interval in which s must lie shrinks to the point $s = 1$. What happens when $\nu \to 0$?

7. Suppose the Example of Sect. 7.2.2 is modified by multiplying the right-hand sides of the differential equation by λ, and by replacing the second boundary condition by $y_1(1) = \mathrm{e}^{-\lambda}$, where $\lambda > 0$ is a large parameter.

(a) What is the exact solution?

(b) What are the conditions on s for the associated initial value problem to have positive and bounded solutions? What happens as $\lambda \to \infty$? As $\lambda \to 0$?

8. The Jacobian elliptic functions sn and cn are defined by

$$\mathrm{sn}(u|k) = \sin\varphi, \quad \mathrm{cn}(u|k) = \cos\varphi, \quad 0 < k < 1,$$

where φ is uniquely determined by

$$u = \int_0^{\varphi} \frac{\mathrm{d}\theta}{(1 - k^2 \sin^2\theta)^{\frac{1}{2}}} \ .$$

(a) Show that $K(k) := \int_0^{\frac{1}{2}\pi} (1 - k^2 \sin^2\theta)^{-\frac{1}{2}} \mathrm{d}\theta$ is the smallest positive zero of cn.

(b) Show that

$$\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{sn}(u|k) = \mathrm{cn}(u|k)\sqrt{1-k^2[\mathrm{sn}(u|k)]^2}\ ,$$
$$\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{cn}(u|k) = -\mathrm{sn}(u|k)\sqrt{1-k^2[\mathrm{sn}(u|k)]^2}\ .$$

(c) Show that the initial value problem

$$y'' = \lambda\sinh(\lambda y),\quad y(0)=0,\ y'(0)=s\quad (|s|<2)$$

has the exact solution

$$y(x;s) = \frac{2}{\lambda}\sinh^{-1}\left(\frac{s}{2}\,\frac{\mathrm{sn}(\lambda x|k)}{\mathrm{cn}(\lambda x|k)}\right),\quad k^2 = 1-\frac{s^2}{4}\ .$$

Hence show that $y(x;s)$, $x>0$, becomes singular for the first time when $x = x_\infty$, where

$$x_\infty = \frac{K(k)}{\lambda}\ .$$

(d) From the known expansion (see Radon [1950])

$$K(k) = \ln\frac{4}{\sqrt{1-k^2}} + \frac{1}{4}\left(\ln\frac{4}{\sqrt{1-k^2}} - 1\right)(1-k^2) + \cdots\ ,\quad k\to 1,$$

conclude that

$$x_\infty \sim \frac{1}{\lambda}\ln\frac{8}{|s|}\quad \text{as}\quad s\to 0.$$

9. It has been shown in the first Example of Sect. 7.2.1 that the boundary value problem

$$y'' + \mathrm{e}^{-y} = 0,\quad 0\le x\le 1,\quad y(0)=y(1)=0$$

has a unique solution that is nonnegative on $[0,1]$.

(a) Set up a finite difference method for solving the problem numerically. (Use a uniform grid $x_n = \frac{n}{N+1}$, $n = 0,1,\ldots,N+1$, and the simplest of finite difference approximations to y''.)

(b) Write the equations for the approximate vector $\boldsymbol{u}^T = [u_1, u_2, \ldots, u_N]$ in fixed point form $\boldsymbol{u} = \boldsymbol{\varphi}(\boldsymbol{u})$ and find a compact domain $\mathcal{D}\subset\mathbb{R}^N$ such that $\boldsymbol{\varphi}:\ \mathbb{R}^N\to\mathbb{R}^N$ maps $\mathcal{D}$ into $\mathcal{D}$ and is contractive in $\mathcal{D}$. {*Hint*: use the fact that the tridiagonal matrix $\boldsymbol{A} = \mathrm{tri}[1,-2,1]$ has a nonpositive inverse $\boldsymbol{A}^{-1}$ satisfying $\|\boldsymbol{A}^{-1}\|_\infty \le \frac{1}{8}(N+1)^2$.}

(c) Discuss the convergence of the fixed point iteration applied to the system of finite difference equations.

10. Given the grid $\{x_n\}_{n=0}^{N+1}$ of (7.79), construct finite difference approximations Θ_n for $\theta(x_n)$, where $\theta(x) = \frac{1}{12}[y^{(4)}(x) - 2p(x)y'''(x)]$, such that $\Theta_n - \theta(x_n) = O(h^2)$ (cf. (7.109)). {*Hint*: distinguish the cases $2\le n\le N-1$ and $n=1$ resp. $n=N$.}

11. Prove Theorem 7.3.4.

12. Describe the application of Newton's method for solving the nonlinear finite difference equations $\mathcal{K}_h u = 0$ of (7.117).

13. Let Δ be the subdivision

$$a = x_0 < x_1 < \cdots < x_n < x_{n+1} = b$$

and $S = \{s \in \mathbb{S}_1^0(\Delta) : \ s(a) = s(b) = 0\}$.

(a) With $[\,\cdot\,,\,\cdot\,]$ the inner product defined in (7.130), find an expression for $[u_\nu, u_\mu]$ in terms of the basis of hat functions (cf. Ch. 2, Ex. 72, but note the difference in notation) and in terms of the integrals involved; do this in a similar manner for $\rho_\nu = (r, u_\nu)$, where $(\,\cdot\,,\,\cdot\,)$ is the inner product defined in (7.128).

(b) Suppose that each integral is split into a sum of integrals over each subinterval of Δ and the trapezoidal rule is employed to approximate the values of the integrals. Obtain the resulting approximations for the stiffness matrix $\boldsymbol{U}$ and the load vector $\boldsymbol{\rho}$. Interpret the linear system (7.145) thus obtained as a finite difference method.

14. Apply the approximate variational method of Sect. 7.4.3 to the boundary value problem

$$-y'' = r(x), \quad 0 \le x \le 1; \quad y(0) = y(1) = 0,$$

using for S a space of continuous piecewise quadratic functions. Specifically, take a uniform subdivision

$$\Delta: \quad 0 = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = 1, \quad x_\nu = \nu h,$$

of $[0,1]$ into n subintervals of length $h = 1/n$ and let $S = \{s \in \mathbb{S}_2^0 : \ s(0) = s(1) = 0\}$.

(a) How many basis functions is S expected to have? Explain.

(b) Construct a basis for S. {*Hint*: for $\nu = 1, 2, \ldots, n$ take $u_\nu = A_{\nu-1}$ to be the quadratic function on $[x_{\nu-1}, x_\nu]$ having values $u_\nu(x_{\nu-1}) = u_\nu(x_\nu) = 0$, $u_\nu(x_{\nu-\frac{1}{2}}) = 1$ and define $A_{\nu-1}$ to be zero outside of $[x_{\nu-1}, x_\nu]$. Add to these functions the basis of hat functions B_ν for $\mathbb{S}_1^0(\Delta)$.}

(c) Compute the stiffness matrix $\boldsymbol{U}$ (in (7.142)) for the basis constructed in (b).

(d) Interpret the resulting system $\boldsymbol{U\xi} = \boldsymbol{\rho}$ as a finite difference method applied to the given boundary value problem. What are the meanings of the components of $\boldsymbol{\xi}$?

15. (a) Show that the solution u_S of (7.141) is the orthogonal projection of the exact solution y of (7.138) onto the space S relative to the inner product $[\,\cdot\,,\,\cdot\,]$; that is,
$$[y - u_S, v] = 0 \quad \text{for all } v \in S.$$
(b) With $\|u\|_E$ denoting the energy norm of u (i.e., $\|u\|_E^2 = [u,u]$), show that
$$\|y - u_S\|_E^2 \,=\, \|y\|_E^2 - \|u_S\|_E^2.$$

16. Consider the boundary value problem (7.127) and (7.124). Define the energy norm by $\|u\|_E^2 = [u,u]$. Let Δ_1 and Δ_2 be two subdivisions of $[a,b]$ and $S_i = \{s \in \mathbb{S}_m^k(\Delta_i),\ s(a) = s(b) = 0\},\ i = 1,2$, for some integers m, k with $0 \le k < m$.
(a) With y denoting the exact solution of the boundary value problem, and Δ_1 being a refinement of Δ_2, show that
$$\|y - u_{S_1}\|_E \,\le\, \|y - u_{S_2}\|_E.$$
(b) Let Δ_2 be an arbitrary subdivision of $[a,b]$ with all grid points (including the endpoints) being rational numbers. Prove that there exists a uniform subdivision Δ_1 of $[a,b]$, with $|\Delta_1| = h$ sufficiently small, such that
$$\|u - u_{S_1}\|_E \,\le\, \|y - u_{S_2}\|_E,$$
where S_i are as defined at the beginning of the exercise.

17. Apply the variational method to the boundary value problem
$$\mathcal{L}y := -py'' + qy = r(x), \quad 0 \le x \le 1;$$
$$y(0) = y(1) = 0,$$
where p and q are constants with $p > 0$, $q \ge 0$. Use approximants from the space $S = \operatorname{span}\{u_\nu(x) = \sin(\nu\pi x),\ \nu = 1,2,\dots,n\}$, and interpret $\mathcal{L}u_S$. Find an explicit form for u_S in the case of constant r.

18. Let y be the exact solution of the boundary value problem (7.123)–(7.125) and u_S the approximate solution of the associated extremal problem with $S = \{s \in \mathbb{S}_1^0(\Delta) :\ s(a) = s(b) = 0\}$ and $\Delta :\ a = x_0 < x_1 < \cdots < x_n < x_{n+1} = b$. Prove that
$$\|y - u_S\|_\infty \le \frac{1}{2}\sqrt{\frac{\overline{c}}{\underline{c}}}\ \max_{0\le\nu\le n} \operatorname{osc}_{[x_\nu, x_{\nu+1}]}(y'),$$
where $\operatorname{osc}_{[c,d]}(f) := \max_{[c,d]} f - \min_{[c,d]} f$ and $\overline{c}$, $\underline{c}$ are the constants defined in (7.134). In particular, show that
$$\|y - u_S\|_\infty \le \tfrac{1}{2}\sqrt{\tfrac{\overline{c}}{\underline{c}}}\ |\Delta| \|y''\|_\infty.$$
{*Hint*: apply Theorem 7.4.4, in particular, (7.150).}

19. Consider the boundary value problem (7.123) and (7.127) with $p(x)$ and $q(x)$ being positive constants,

$$p(x) = p > 0, \quad q(x) = q > 0.$$

Let $S = \{s \in \mathbb{S}_1^0(\Delta) : \ s(a) = s(b) = 0\}$, where the subdivision Δ: $a = x_0 < x_1 < x_2 < \cdots < x_n < x_{n+1} = b$ is assumed to be *quasi-uniform*; that is,

$$\Delta x_\nu := x_{\nu+1} - x_\nu \geq \beta|\Delta|, \quad \nu = 0, 1, \ldots, n,$$

for some positive constant β. (Recall that $|\Delta| := \max_{0\leq\nu\leq n} \Delta x_\nu$.) Let $\boldsymbol{U}$ be the stiffness matrix (cf. (7.142)) for the basis $u_\nu = B_\nu$, $\nu = 1, 2, \ldots, n$, of hat functions (cf. Ch. 2, Ex. 72, but note the difference in notation). Write $u(x) = \sum_{\nu=1}^n \xi_\nu u_\nu(x)$ for any $u \in S$, and $\boldsymbol{\xi}^T = [\xi_1, \xi_2, \ldots, \xi_n]$.

(a) Show that $\boldsymbol{\xi}^T \boldsymbol{U} \boldsymbol{\xi} = [u, u]$.

(b) Show that $\|u'\|_{L_2}^2 = \boldsymbol{\xi}^T \boldsymbol{T}_1 \boldsymbol{\xi}$, where $\boldsymbol{T}_1$ is a symmetric tridiagonal matrix with

$$(\boldsymbol{T}_1)_{\nu,\nu} = \frac{1}{\Delta x_{\nu-1}} + \frac{1}{\Delta x_\nu}\ , \quad \nu = 1, 2, \ldots, n;$$
$$(\boldsymbol{T}_1)_{\nu+1,\nu} = (\boldsymbol{T}_1)_{\nu,\nu+1} = -\frac{1}{\Delta x_\nu}\ , \quad \nu = 1, \ldots, n-1.$$

{*Hint*: use integration by parts, being careful to observe that u' is only piecewise continuous.}

(c) Show that $\|u\|_{L_2}^2 = \boldsymbol{\xi}^T \boldsymbol{T}_0 \boldsymbol{\xi}$, where $\boldsymbol{T}_0$ is a symmetric tridiagonal matrix with

$$(\boldsymbol{T}_0)_{\nu,\nu} = \tfrac{1}{3}(\Delta x_{\nu-1} + \Delta x_\nu), \quad \nu = 1, 2, \ldots, n;$$
$$(\boldsymbol{T}_0)_{\nu+1,\nu} = (\boldsymbol{T}_0)_{\nu,\nu+1} = \tfrac{1}{6}\Delta x_\nu, \quad \nu = 1, \ldots, n-1.$$

(d) Combine (a)–(c) to compute $[u, u]$ and hence to estimate the Euclidean condition number $\operatorname{cond}_2 \boldsymbol{U}$. {*Hint*: use Gershgorin's theorem to estimate the eigenvalues of $\boldsymbol{U}$.}

(e) The analysis in (d) fails if $q = 0$. Show, however, in the case of a *uniform* grid, that when $q = 0$ then $\operatorname{cond}_2 \boldsymbol{U} \leq 1/\sin^2 \frac{\pi}{4n}$.

(f) Indicate how the argument in (d) can be extended to variable $p(x)$, $q(x)$ satisfying $0 < p(x) \leq \bar{p},\ 0 < \underline{q} \leq q(x) \leq \bar{q}$ on $[a, b]$.

20. The *method of collocation* for solving a boundary value problem

$$\mathcal{L}y = r(x), \quad 0 \leq x \leq 1; \quad y(0) = y(1) = 0,$$

consists of selecting an n-dimensional subspace $S \subset V_0$ and determining $u_S \in S$ such that $(\mathcal{L}u_S)(x_\mu) = r(x_\mu)$ for a discrete set of points $0 < x_1 < x_2 <$

$\cdots < x_n < 1$. Apply this method to the problem of Ex. 17, with S as defined there. Discuss the solvability of the system of linear equations involved in the method. {*Hint*: use the known fact that the only trigonometric sine polynomial $\sum_{\nu=1}^{n} \xi_\nu \sin(\nu\pi x)$ of degree n that vanishes at n distinct points in $(0,1)$ is the one identically zero.}

MACHINE ASSIGNMENTS

1. The following eigenvalue problem arises in the physics of gas discharges. Determine the smallest positive $\lambda > 0$ such that

$$\varphi'' + \frac{1}{r}\varphi' + \lambda^2\varphi(1-\varphi) = 0, \quad 0 < r \le 1,$$

$$\varphi(0) = a, \quad \varphi'(0) = 0, \quad \varphi(1) = 0,$$

where a is given, $0 < a < 1$.

   (a) Explain why $\lambda = 0$ cannot be an eigenvalue.

   (b) Reduce the problem to an initial value problem. {*Hint*: make a change of variables, $x = \lambda r$, $y(x) = \varphi(x/\lambda)$.}

   (c) Use Maple to determine the Taylor expansion up to the power x^8 of the solution $y(x,a)$ to the initial value problem of (b).

   (d) Integrate the initial value problem starting at $x = .1$, using the Taylor expansion of (c) to determine the initial data $y(.1,a)$ and $\frac{\mathrm{d}y}{\mathrm{d}x}(.1,a)$. Use the classical Runge–Kutta method (for example the Matlab routine of Ch. 5, MA 1(a)) and integrate until the solution y becomes negative. Then apply interpolation to compute an approximation to λ, the solution of $y(\,\cdot\,,a) = 0$, to an accuracy of about five decimal digits. Prepare a table of the λ so obtained for $a = .1 : .1 : .9$, including the values of the integration step h required.

   (e) For $a = .1 : .1 : .9$ use Matlab to produce graphs of the solutions $y(x,a)$ on intervals from $x = 0$ to $x = \lambda$, the zero of y. (Determine the endpoints of these intervals from the results of (d).) Use the Matlab routine `ode45` to do the integration from .1 to λ and connect the points $(0,a)$ and $(.1, y(.1,a))$ by a straight line segment. (Compute $y(.1,a)$) by the Taylor expansion of (c).)

2. The shape of an ideal flexible chain of length L, hung from two points $(0,0)$ and $(1,1)$, is determined by the solution of the eigenvalue problem

$$y'' = \lambda\sqrt{1+(y')^2}, \quad y(0) = 0,\ y(1) = 1,\ \int_0^1 \sqrt{1+(y')^2}\mathrm{d}x = L.$$

Strictly speaking, this is not a problem of the form (7.3), (7.4), but nevertheless can be solved analytically as well as numerically.

(a) On physical grounds, what condition must L satisfy for the problem to have a solution?

(b) Derive three equations in three unknowns: the two constants of integration and the eigenvalue λ. Obtain a transcendental equation for λ by eliminating the other two unknowns. Solve the equation numerically and thus find, and plot, the solution for $L = 2, 4, 8$, and 16.

(c) If one attempts to solve the problem by a finite difference method over a uniform grid $x_0 = 0 < x_1 < x_2 < \cdots < x_N < x_{N+1} = 1$, $x_n = \frac{n}{N+1}$, approximating the integral in the third boundary condition by the composite trapezoidal rule, a system of $N + 1$ nonlinear equations in $N + 1$ unknowns results. Solve the system by a homotopy method, using L as the homotopy parameter. Since for $L = \sqrt{2}$ the solution is trivial, select a sequence of parameter values $L_0 = \sqrt{2} < L_1 < \cdots < L_m$ and solve the finite difference equations for L_i using the solution for L_{i-1} as the initial approximation. Implement this for the values of L given in (b), taking a sequence $\{L_i\}$ which contains these values. Compare the numerical results for the eigenvalues with the analytic ones for $N = 10,\ 20,\ 40$. (Use the routine `fsolve` from the Matlab optimization toolbox to solve the system of nonlinear equations.)

3. Change the boundary value problem of the first Example of Sect. 7.2.1 to

$$y'' = -\mathrm{e}^y, \quad 0 \le x \le 1, \quad y(0) = y(1) = 0.$$

Then Theorem 7.1.2 no longer applies (why not?). In fact, it is known that the problem has two solutions. Use Matlab to compute the respective initial slopes $y'(0)$ to 12 significant digits by Newton's method, as in the Example of Sect. 7.2.1. {*Hint*: use approximations $s^{(0)} = 1$ and $s^{(0)} = 15$ to the initial slopes.}

4. Consider the boundary value problem

$$\text{(BVP)} \qquad y'' = y^2, \quad 0 \le x \le b; \quad y(0) = 0,\ y(b) = \beta,$$

and the associated initial value problem

$$\text{(IVP)} \qquad u'' = u^2, \quad u(0) = 0,\ u'(0) = s.$$

Denote the solution of (IVP) by $u(x) = u(x; s)$.

(a) Let $v(x) = u(x; -1)$. Show that

$$v'(x) = -\sqrt{\tfrac{2}{3}v^3(x) + 1},$$

and thus the function v, being convex (i.e., $v'' > 0$), has a minimum at some $x_0 > 0$ with value $v_{\min} = -(3/2)^{1/3} = -1.1447142\ldots$. Show that v is symmetric with respect to the line $x = x_0$.

(b) Compute x_0 numerically in terms of the beta integral. {*Point of information*: The beta integral is $B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\mathrm{d}t$ and has the value $\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$.}

(c) Use Matlab to compute v by solving the initial value problem

$$v'' = v^2, \quad x_0 \le x \le 3x_0; \quad v(x_0) = v_{\min},\ v'(x_0) = 0.$$

Plot the solution (and its symmetric part) on $-x_0 \le x \le 3x_0$.

(d) In terms of the function v defined in (a), show that

$$u(x; -s^3) = s^2 v(sx), \quad \text{all } s \in \mathbb{R}.$$

As s ranges over all the reals, the solution manifold $\{s^2 v(sx)\}$ thus encompasses all the solutions of (IVP). Prepare a plot of this solution manifold. Note that there exists an envelope of the manifold, located in the lower half-plane. Explain why, in principle, this envelope must be the solution of a first-order differential equation.

(e) Based on the plot obtained in (d), discuss the number of possible solutions to the original boundary value problem (BVP). In particular, determine for what values of b and β there does not exist any solution.

(f) Use the method of finite differences on a uniform grid to compute the two solutions of (BVP) for $b = 3x_0$, $\beta = v_0$, where $v_0 = v(3x_0)$ is a quantity already computed in (c). Solve the systems of nonlinear difference equations by Newton's method. In trying to get the first solution, approximate the solution v of (a) on $0 \le x \le 3x_0$ by a quadratic function $\tilde{v}$ satisfying $\tilde{v}(0) = 0$, $\tilde{v}'(0) = -1$, $\tilde{v}(3x_0) = v_0$, and then use its restriction to the grid as the initial approximation to Newton's method. For the second solution, try the initial approximation obtained from the linear approximation $\overline{v}(x) = v_0 x/(3x_0)$. In both cases, plot initial approximations as well as the solutions to the difference equations. What happens if Newton's method is replaced by the method of successive approximations?

5. The following boundary value problem occurs in soil engineering. Determine $y(r)$, $1 \le r < \infty$, such that

$$\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(ry\frac{\mathrm{d}y}{\mathrm{d}r}\right) + \rho(1-y) = 0, \quad y(1) = \eta,\ y(\infty) = 1,$$

where ρ, η are parameters satisfying $\rho > 0$, $0 < \eta < 1$. The quantity of interest is $\sigma = \left.\frac{\mathrm{d}y}{\mathrm{d}r}\right|_{r=1}$.

(a) Let $z(x) = [y(\mathrm{e}^x)]^2$. Derive the boundary value problem and the quantity of interest in terms of z.

(b) Consider the initial value problem associated with the boundary value problem in (a), having initial conditions

$$z(0) = \eta^2, \quad z'(0) = s.$$

Discuss the qualitative behavior of its solutions for real values of s. {*Hint*: suggested questions may be: admissible domain in the (x, z)-plane, convexity and concavity of the solutions, the role of the line $z = 1$.}

(c) From your analysis in (b), devise an appropriate shooting procedure for solving the boundary value problem numerically and for computing the quantity σ. Run your procedure on the computer for various values of η and ρ. In particular, prepare a 5-decimal table showing the values of s and σ for $\eta = .1(.1).9$ and $\rho = .5, 1, 2, 5, 10$, and plot σ versus η.

ANSWERS TO THE EXERCISES AND MACHINE ASSIGNMENTS OF CHAPTER 7

ANSWERS TO EXERCISES

1. See the text.

2. (a) Using $xu'' + u' = (xu')'$, one can write the differential equation in the form

$$(xu')' - u' + \tfrac{1}{2}(u^2)' = 0.$$

Integrating this from 0 to x and using the initial condition $u(0) = 0$ gives

$$xu' - u + \tfrac{1}{2}u^2 = 0, \quad u' = \frac{2u - u^2}{2x}\,.$$

Considering x as a function of u, we get

$$\frac{1}{x}\,\frac{\mathrm{d}x}{\mathrm{d}u} = \frac{2}{2u - u^2} = \frac{1}{u} + \frac{1}{2 - u}\,.$$

Integrating, one finds

$$\ln cx = \ln|u| - \ln|2 - u|, \quad cx = \left|\frac{u}{2 - u}\right|,$$

where c is a constant. Assuming $u > 0$, $2 - u > 0$ (all other sign combinations give the same result), one finds

$$u(x) = \frac{2cx}{1 + cx}\,.$$

The condition $u'(0) = s$ then yields $2c = s$, giving

$$u(x; s) = \frac{2sx}{2 + sx}\,.$$

If $s > 0$, this is well defined for all $0 \le x \le 1$. If $s < 0$, the denominator must remain positive for $0 \le x \le 1$, which requires $2 + s > 0$, i.e., $s > -2$. Thus, the s-interval in question is

$$-2 < s < \infty.$$

(b) We have

$$u(1; s) - 1 = \frac{2s}{2 + s} - 1 = \frac{s - 2}{s + 2} =: f(s).$$

The function f is monotonically increasing on $-2 < s < \infty$ and concave, having the unique zero $s = 2$. For Newton's method to converge, we

must make sure that all iterates remain in the interval $s > -2$. This is true for $-2 < s < s^0$, where s^0 is the solution in $s > -2$ of

$$s^0 - \frac{f(s^0)}{f'(s^0)} = -2, \quad (s^0)^2 - 4s^0 - 12 = 0,$$

that is, $s^0 = 6$. Thus, Newton's method converges for all s in

$$-2 < s < 6,$$

which further restricts the interval found in (a).

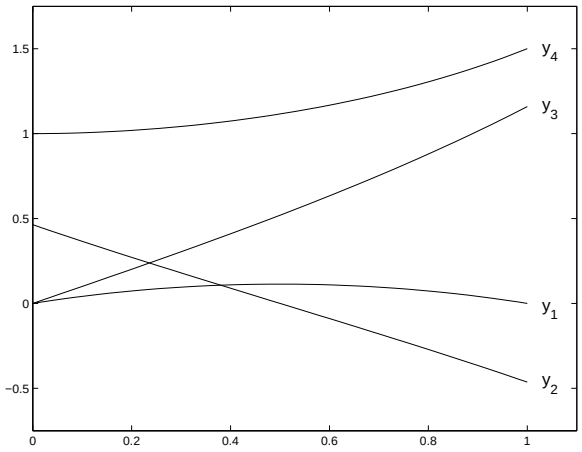
3. PROGRAMS

```
%EXVII_3
%
s1=1; s=0; nu=0;
xspan=[0 1];
disp('      nu       s^{nu}')
options=odeset('AbsTol',.5e-12);
while abs(s1-s)>.5e-12
  s=s1; nu=nu+1;
  y0=[0;s;0;1];
  [x,y]=ode45(@fEXVII_3,xspan,y0,options);
  sy=size(y,1);
  s1=s-y(sy,1)/y(sy,3);
  fprintf('%8.0f %16.12f\n',nu,s1)
end
plot(x,y)

%FEXVII_3
%
function yprime=fEXVII_3(x,y)
yprime=[y(2);-exp(-y(1));y(4);exp(-y(1))*y(3)];
```

OUTPUT

```
>> EXVII_3
      nu       s^{nu}
       1   0.454743906534
       2   0.463629562756
       3   0.463632593670
       4   0.463632593670
>>
```

PLOTS



4. As mentioned in the Example of Sect. 7.2.2, the system (7.55) (and therefore also the system (7.57)) can be linearized by putting $\eta_1 = y_1^{-1}$, $\eta_2 = y_2^{-1}$ resp. $\eta_1 = u_1^{-1}$, $\eta_2 = u_2^{-1}$:

$$\frac{\mathrm{d}\eta_1}{\mathrm{d}x} = -\eta_2,$$
$$\frac{\mathrm{d}\eta_2}{\mathrm{d}x} = -\eta_1.$$

Differentiating the first equation and using the second, one gets $\eta_1'' = \eta_1$, the general solution of which is $c_1\mathrm{e}^x + c_2\mathrm{e}^{-x}$. The boundary conditions in (7.55) imply $c_1 = 0$, $c_2 = 1$, that is, $\eta_1(x) = \mathrm{e}^{-x}$. From the first equation, therefore, $\eta_2(x) = \mathrm{e}^{-x}$, thus proving (7.56). Likewise, the initial conditions in (7.57) imply $c_1 = \frac{1}{2}\left(1 - \frac{1}{s}\right)$, $c_2 = \frac{1}{2}\left(1 + \frac{1}{s}\right)$, giving

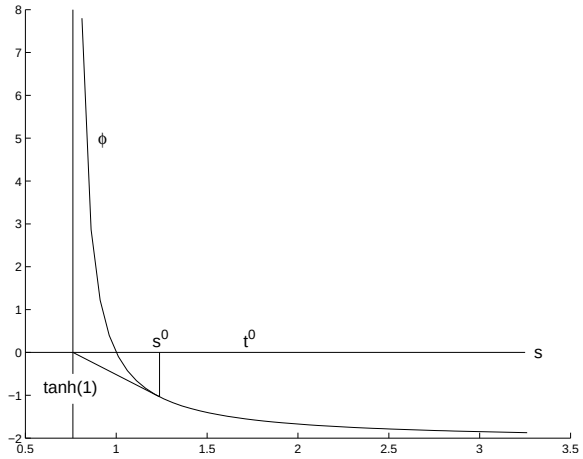
$$\eta_1(x) = \frac{1}{2}\left(1 - \frac{1}{s}\right)\mathrm{e}^x + \frac{1}{2}\left(1 + \frac{1}{s}\right)\mathrm{e}^{-x} = \cosh x - \frac{1}{s}\sinh x,$$
$$\eta_2(x) = -\eta_1'(x) = -\sinh x + \frac{1}{s}\cosh x.$$

Therefore,

$$u_1(x) = \frac{1}{\eta_1(x)} = \frac{s}{s\cosh x - \sinh x}, \quad u_2(x) = \frac{1}{\eta_2(x)} = \frac{s}{\cosh x - s\sinh x},$$

as claimed in (7.62).

5. (a) By the convexity of ϕ (see the figure on the next page) one has



$s^0 < t^0$ precisely if

$$t^0 - \frac{\phi(t^0)}{\phi'(t^0)} < \tanh 1.$$

With $t^0 = \coth 1$, straightforward calculation shows that

$$\phi(t^0) = \cosh 1 - \mathrm{e} = -\sinh 1,$$
$$\phi'(t^0) = \frac{-\sinh 1}{(s\cosh 1 - \sinh 1)^2}\bigg|_{s=\coth 1} = -\sinh^3 1,$$

so that the inequality to be verified is

$$\coth 1 - \frac{1}{\sinh^2 1} < \tanh 1.$$

Numerical calculation gives for the left-hand side .5889... and for the right-hand side .7615... , which confirms the inequality.

(b) A lower initial approximation for s^0 is 1, the zero of ϕ (cf. the figure in (a)), while an upper approximation is $\coth 1$, as shown in (a). The equation for s^0 is $f(s) = 0$, where $f(s) = -s + \frac{\phi(s)}{\phi'(s)} + \tanh 1$. At the lower approximation, f is negative, and at the upper approximation, f is positive. Applying the (nonsymbolic version of the) bisection routine of Ch. 4, Sect. 4.3.1, with `eps` $= .5 \times 10^{-12}$, one obtains $s^0 = 1.238405844044$.

```
 PROGRAMS

%EXVII_5B
%
a=1; b=coth(1); tol=.5e-12;
[ntol,s0]=bisec(a,b,tol)
```

```
%BISEC  Bisection method
%
function [ntol,x]=bisec(a,b,tol)
ntol=ceil(log((b-a)/tol)/log(2));
for n=1:ntol
  x=(a+b)/2;
  fx=f(x);
  if fx<0
    a=x;
  else
    b=x;
  end
end

function y=f(s)
temp=s*cosh(1)-sinh(1);
y=-s-temp*(s-temp*exp(1))/sinh(1)+tanh(1);
```

6. (a) Put $z_1 = y_1^{-\nu}$, $z_2 = y_2^{-\nu}$. Then

$$\frac{\mathrm{d}z_1}{\mathrm{d}x} = -\frac{\nu}{y_1^{\nu+1}}\frac{\mathrm{d}y_1}{\mathrm{d}x} = -\frac{\nu}{y_1^{\nu+1}}\frac{y_1^{\nu+1}}{y_2^{\nu}} = -\frac{\nu}{y_2^{\nu}} = -\nu z_2,$$
$$\frac{\mathrm{d}z_2}{\mathrm{d}x} = -\frac{\nu}{y_2^{\nu+1}}\frac{\mathrm{d}y_2}{\mathrm{d}x} = -\frac{\nu}{y_2^{\nu+1}}\frac{y_2^{\nu+1}}{y_1^{\nu}} = -\frac{\nu}{y_1^{\nu}} = -\nu z_1,$$

so that

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = -\nu \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix};\quad z_1(0) = 1,\ z_1(1) = \mathrm{e}^{-\nu}.$$

The general solution of this system is

$$z(x) = c_1\begin{bmatrix} 1 \\ -1 \end{bmatrix}\mathrm{e}^{\nu x} + c_2\begin{bmatrix} 1 \\ 1 \end{bmatrix}\mathrm{e}^{-\nu x},$$

while the boundary conditions give

$$c_1 + c_2 = 1,$$
$$c_1\mathrm{e}^{\nu} + c_2\mathrm{e}^{-\nu} = \mathrm{e}^{-\nu},$$

that is,

$$c_1 + c_2 = 1,$$
$$\mathrm{e}^{2\nu}c_1 + c_2 = 1.$$

There follows $c_1 = 0$, $c_2 = 1$, and so

$$z(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathrm{e}^{-\nu x}, \quad y(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathrm{e}^{x}.$$

(b) Letting $z_1 = u_1^{-\nu}$, $z_2 = u_2^{-\nu}$, we have, from the work in (a),

$$z(x) = c_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} \mathrm{e}^{\nu x} + c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathrm{e}^{-\nu x}.$$

The given initial conditions yield $z_1(0) = 1$, $z_2(0) = s^{-\nu}$, thus

$$\begin{aligned} c_1 + c_2 &= 1, \\ -c_1 + c_2 &= s^{-\nu}. \end{aligned}$$

The solution is $c_1 = \frac{1}{2}(1 - s^{-\nu})$, $c_2 = \frac{1}{2}(1 + s^{-\nu})$, which yields

$$\begin{aligned} z_1(x) &= \cosh \nu x - s^{-\nu} \sinh \nu x, \\ z_2(x) &= -\sinh \nu x + s^{-\nu} \cosh \nu x, \end{aligned}$$

and thus

$$u_1(x) = \frac{s}{(s^\nu \cosh \nu x - \sinh \nu x)^{1/\nu}}, \quad u_2(x) = \frac{s}{(\cosh \nu x - s^\nu \sinh \nu x)^{1/\nu}}.$$

(c) In order that the denominators of $u_1(x)$ and $u_2(x)$ remain positive on $[0, 1]$ requires

$$s^\nu \cosh \nu - \sinh \nu > 0, \quad \cosh \nu - s^\nu \sinh \nu > 0,$$

hence

$$[\tanh \nu]^{1/\nu} < s < [\coth \nu]^{1/\nu}.$$

Clearly, as $\nu \to \infty$, both bounds tend to 1, in fact, the lower one monotonically increasing, and the upper one monotonically decreasing. When $\nu \to 0$, the interval approaches $(0, \infty)$, i.e., there are no restrictions on s. This makes sense, since in the limit $\nu \to 0$, the given system becomes linear.

7. (a) By considering the linear boundary value problem for $z_1 = y_1^{-1}$, $z_2 = y_2^{-1}$, one finds $z_1(x) = z_2(x) = \mathrm{e}^{-\lambda x}$, hence $y_1(x) = y_2(x) = \mathrm{e}^{\lambda x}$.

(b) The solution of the initial value problem, via the linear initial value problem, is found to be

$$u_1(x; s) = \frac{s}{s \cosh \lambda x - \sinh \lambda x}, \quad u_2(x; s) = \frac{s}{\cosh \lambda x - s \sinh \lambda x},$$

from which one obtains the condition

$$\tanh\lambda < s < \coth\lambda$$

for the solutions to remain positive and bounded on $[0,1]$. As $\lambda \to \infty$, both bounds tend to 1, i.e., the interval shrinks to a point. In the limit as $\lambda \to 0$, there is no restriction on s; the solution in the limit is trivially $y(x) \equiv \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

8. (a) By definition, u is a monotonically increasing function of φ for $\varphi \in \mathbb{R}$. Hence, φ is also monotonically increasing as a function of u, and therefore $\mathrm{cn}(u|k)$, which is 1 at $u = 0$, vanishes for the first time when $\varphi = \frac{1}{2}\pi$, that is, $u = \int_0^{\frac{1}{2}\pi} (1 - k^2 \sin^2\theta)^{-\frac{1}{2}} \mathrm{d}\theta =: K(k)$.

   (b) We have

$$\frac{\mathrm{d}}{\mathrm{d}u}\,\mathrm{sn}(u|k) = \frac{\mathrm{d}\,\mathrm{sn}(u|k)}{\mathrm{d}\varphi}\,\frac{\mathrm{d}\varphi}{\mathrm{d}u} = \cos\varphi\,\frac{\mathrm{d}\varphi}{\mathrm{d}u} = \mathrm{cn}(u|k)\frac{\mathrm{d}\varphi}{\mathrm{d}u}\,.$$

   But

$$\frac{\mathrm{d}u}{\mathrm{d}\varphi} = \frac{1}{\sqrt{1 - k^2\sin^2\varphi}}\,, \quad \frac{\mathrm{d}\varphi}{\mathrm{d}u} = \sqrt{1 - k^2\sin^2\varphi} = \sqrt{1 - k^2[\mathrm{sn}(u|k)]^2},$$

   so that

$$\frac{\mathrm{d}}{\mathrm{d}u}\,\mathrm{sn}(u|k) = \mathrm{cn}(u|k)\sqrt{1 - k^2[\mathrm{sn}(u|k)]^2}.$$

   The second formula follows in the same manner.

   (c) First of all, $y(x;s)$ as defined satisfies the first initial condition, $y(0;s) = 0$. Differentiating with respect to x gives

$$\begin{aligned}\frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{2}{\lambda}\frac{1}{\sqrt{1 + \left(\frac{s}{2}\frac{\mathrm{sn}}{\mathrm{cn}}\right)^2}} \cdot \frac{s}{2}\frac{\mathrm{cn}^2\sqrt{1 - k^2\mathrm{sn}^2} + \mathrm{sn}^2\sqrt{1 - k^2\mathrm{sn}^2}}{\mathrm{cn}^2} \cdot \lambda \\ &= \frac{s}{\sqrt{1 + \frac{s^2}{4}\frac{\mathrm{sn}^2}{\mathrm{cn}^2}}}\frac{\sqrt{1 - k^2\mathrm{sn}^2}}{\mathrm{cn}^2}\,,\end{aligned}$$

   where the formulae in (b) have been used and where $\mathrm{sn} = \mathrm{sn}(\lambda x|k)$, $\mathrm{cn} = \mathrm{cn}(\lambda x|k)$, $\mathrm{sn}^2 + \mathrm{cn}^2 = 1$. Since $k^2 = 1 - \frac{s^2}{4}$, we have

$$\begin{aligned}\sqrt{1 - k^2\mathrm{sn}^2} = \sqrt{1 - (1 - \tfrac{s^2}{4})\mathrm{sn}^2} &= \sqrt{1 - \mathrm{sn}^2 + \tfrac{s^2}{4}\mathrm{sn}^2} = \sqrt{\mathrm{cn}^2 + \tfrac{s^2}{4}\mathrm{sn}^2} \\ &= \mathrm{cn}\sqrt{1 + \tfrac{s^2}{4}\tfrac{\mathrm{sn}^2}{\mathrm{cn}^2}}\,,\end{aligned}$$

   so that

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{s}{\mathrm{cn}(\lambda x|k)}\,.$$

We now see that the second initial condition, $y'(0;s) = s$, is also satisfied. Differentiating once again with respect to x, we get

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = -\frac{s\lambda}{\mathrm{cn}^2}\,(-\mathrm{sn}\sqrt{1-k^2\mathrm{sn}^2}) = \frac{\lambda s}{\mathrm{cn}^2}\,\mathrm{sn}\cdot\mathrm{cn}\sqrt{1+\frac{s^2}{4}\frac{\mathrm{sn}^2}{\mathrm{cn}^2}} = \lambda s\frac{\mathrm{sn}}{\mathrm{cn}}\sqrt{1+\frac{s^2}{4}\frac{\mathrm{sn}^2}{\mathrm{cn}^2}}\,,$$

where, as before, $\mathrm{sn} = \mathrm{sn}(\lambda x|k)$, $\mathrm{cn} = \mathrm{cn}(\lambda x|k)$. On the other hand,

$$\sinh(\lambda y) = \sinh\left(2\sinh^{-1}\left(\frac{s}{2}\frac{\mathrm{sn}}{\mathrm{cn}}\right)\right).$$

Using the duplication formula $\sinh(2t) = 2\sinh t\cosh t$ and the fact that $\cosh t = \sqrt{1+\sinh^2 t}$, we obtain

$$\sinh(\lambda y) = 2\cdot\frac{s}{2}\frac{\mathrm{sn}}{\mathrm{cn}}\cdot\sqrt{1+\left(\frac{s}{2}\frac{\mathrm{sn}}{\mathrm{cn}}\right)^2} = s\,\frac{\mathrm{sn}}{\mathrm{cn}}\sqrt{1+\frac{s^2}{4}\frac{\mathrm{sn}^2}{\mathrm{cn}^2}}\,.$$

Comparing with the formula for $\frac{\mathrm{d}^2y}{\mathrm{d}x^2}$ above, we conclude that

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = \lambda\sinh(\lambda y),$$

as was to be shown.

From the explicit formula for $y(x;s)$ we see that $y(x;s)$ monotonically increases on $0 \le x < \frac{1}{\lambda}K(k)$ (cf. (a)), and tends to ∞ as $x \to x_\infty$, where $x_\infty = \frac{1}{\lambda}K(k)$.

(d) As $s \to 0$, we have $k^2 = 1 - \frac{s^2}{4} \to 1$, and from the given expansion, we obtain

$$K(k) = \ln\frac{8}{|s|} + \frac{1}{4}\left(\ln\frac{8}{|s|} - 1\right)\frac{s^2}{4} + \cdots\,, \quad s \to 0,$$

hence

$$x_\infty = \frac{1}{\lambda}K(k) \sim \frac{1}{\lambda}\ln\frac{8}{|s|}, \quad s \to 0.$$

Thus, for large λ, the initial slope $|s|$ has to be very small to successfully integrate from $x = 0$ to $x = 1$, i.e., to have $x_\infty > 1$. Not only that, it can be shown that the solution of the boundary value problem has a singularity at $x_\infty^* \approx 1 + \frac{1}{\lambda\cosh(\lambda/2)}$, which for large λ is precariously close to the right endpoint $x = 1$. For example, $x_\infty^* \approx 1.0326\ldots$ if $\lambda = 5$.

9. (a) Using the central difference approximation to y'' yields

$$u_{n+1} - 2u_n + u_{n-1} + h^2\mathrm{e}^{-u_n} = 0, \quad n = 1, 2, \ldots, N; \quad u_0 = u_{N+1} = 0,$$

where $h = 1/(N+1)$, or, in matrix form,

$$\boldsymbol{A}\boldsymbol{u} + h^2\boldsymbol{e}(\boldsymbol{u}) = \boldsymbol{0}, \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}, \quad \boldsymbol{e}(\boldsymbol{u}) = \begin{bmatrix} \mathrm{e}^{-u_1} \\ \mathrm{e}^{-u_2} \\ \vdots \\ \mathrm{e}^{-u_N} \end{bmatrix},$$

where $\boldsymbol{A}$ is the tridiagonal matrix defined in the *Hint* to part (b).

(b) We have

$$\boldsymbol{u} = \boldsymbol{\varphi}(\boldsymbol{u}), \quad \text{where } \boldsymbol{\varphi}(\boldsymbol{u}) = -h^2\boldsymbol{A}^{-1}\boldsymbol{e}(\boldsymbol{u}).$$

Since by the *Hint*, and the fact that $\boldsymbol{u} \geq \boldsymbol{0}$ (since $\boldsymbol{A}^{-1} \leq \boldsymbol{0}$),

$$\|\boldsymbol{\varphi}(\boldsymbol{u})\|_\infty \leq \tfrac{1}{8}h^2(N+1)^2\|\boldsymbol{e}(\boldsymbol{u})\|_\infty \leq \tfrac{1}{8}\,,$$

the map $\boldsymbol{\varphi}$ takes the cube $\mathcal{D} = \{\boldsymbol{u} \in \mathbb{R}^N : 0 \leq u_n \leq \frac{1}{8},\ n = 1, 2, \ldots, N\}$ into $\mathcal{D}$. Moreover,

$$\frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{u}} = -h^2\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{u}} = h^2\boldsymbol{A}^{-1}\mathrm{diag}(\mathrm{e}^{-u_1}, \mathrm{e}^{-u_2}, \ldots, \mathrm{e}^{-u_N}),$$

so that

$$\left\| \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{u}} \right\|_\infty \leq h^2\frac{(N+1)^2}{8} \cdot 1 = \frac{1}{8}\,, \quad \text{all } \boldsymbol{u} \in \mathcal{D}.$$

Since, for any $\boldsymbol{u}$, $\boldsymbol{u}^* \in \mathcal{D}$, one has

$$\|\boldsymbol{\varphi}(\boldsymbol{u}) - \boldsymbol{\varphi}(\boldsymbol{u}^*)\|_\infty = \left\| \frac{\partial \boldsymbol{\varphi}}{\partial \boldsymbol{u}}(\overline{\boldsymbol{u}})(\boldsymbol{u} - \boldsymbol{u}^*) \right\|_\infty \leq \frac{1}{8}\|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty,$$

the map $\boldsymbol{\varphi}$ is contractive on $\mathcal{D}$.

(c) By the contraction mapping principle (cf. Chap. 4, Theorem 4.9.1), the equation in (b) has a unique solution, and the fixed point iteration

$$\boldsymbol{u}^{[i+1]} = \boldsymbol{\varphi}(\boldsymbol{u}^{[i]}), \quad i = 0, 1, 2, \ldots\,,$$

converges for every $\boldsymbol{u}^{[0]} \in \mathcal{D}$ geometrically with ratio $\frac{1}{8}$.

10. For $2 \leq n \leq N-1$ we can construct symmetric difference approximations as follows: combine

$$y'''(x_n) = \frac{y''(x_{n+1}) - y''(x_{n-1})}{2h} + O(h^2)$$

with

$$y''(x_{n+1}) = \frac{1}{h^2}(y_{n+2} - 2y_{n+1} + y_n) + O(h^2),$$
$$y''(x_{n-1}) = \frac{1}{h^2}(y_n - 2y_{n-1} + y_{n-2}) + O(h^2)$$

to obtain

$$y'''(x_n) = \frac{1}{2h^3}(y_{n+2} - 2y_{n+1} + 2y_{n-1} - y_{n-2}) + O(h),$$

where the error term is actually $O(h^2)$, the right-hand side being an even function of h. This holds if $y \in C^5[a,b]$. Furthermore,

$$y^{(4)}(x_n) = \frac{1}{h^4}(y_{n+2} - 4y_{n+1} + 6y_n - 4y_{n-1} + y_{n-2}) + O(h^2),$$

provided that $y \in C^6[a,b]$. Therefore,

$$\begin{aligned}\theta(x_n) &= \frac{1}{12}\,[y^{(4)}(x_n) - 2p_n y'''(x_n)] \\ &= \frac{1}{12h^4}[(1-hp_n)y_{n+2} - 2(2-hp_n)y_{n+1} + 6y_n \\ &\qquad -2(2+hp_n)y_{n-1} + (1+hp_n)y_{n-2}] + O(h^2),\end{aligned}$$

where $p_n = p(x_n)$, and $y \in C^6[a,b]$.

For $n = 1$, we seek an approximation

$$y^{(4)}(x_1) = \frac{1}{h^4}(a_0y_0 + a_1y_1 + a_2y_2 + a_3y_3 + a_4y_4 + a_5y_5) + O(h^2).$$

Expanding the right-hand side in a Taylor series about x_1, one finds for the a_i the linear system

$$\begin{array}{rrrrrrl} a_0 & +a_1 & +a_2 & +a_3 & +a_4 & +a_5 & = 0, \\ -a_0 & & +a_2 & +2a_3 & +3a_4 & +4a_5 & = 0, \\ a_0 & & +a_2 & +4a_3 & +9a_4 & +16a_5 & = 0, \\ -a_0 & & +a_2 & +8a_3 & +27a_4 & +64a_5 & = 0, \\ a_0 & & +a_2 & +16a_3 & +81a_4 & +256a_5 & = 24, \\ -a_0 & & +a_2 & +32a_3 & +243a_4 & +1024a_5 & = 0. \end{array}$$

Gauss elimination, or more easily Matlab, yields the solution

$$a_0 = 2, \quad a_1 = -9, \quad a_2 = 16, \quad a_3 = -14, \quad a_4 = 6, \quad a_5 = -1,$$

hence

$$y^{(4)}(x_1) = \frac{1}{h^4}(2y_0 - 9y_1 + 16y_2 - 14y_3 + 6y_4 - y_5) + O(h^2).$$

Similarly we approximate

$$y'''(x_1) = \frac{1}{h^3}(b_0y_0 + b_1y_1 + b_2y_2 + b_3y_3 + b_4y_4) + O(h^2)$$

and obtain by Taylor expansion about x_1 the system

$$\begin{array}{rrrrrr} b_0 & +b_1 & +b_2 & +b_3 & +b_4 & = 0, \\ -b_0 & & +b_2 & +2b_3 & +3b_4 & = 0, \\ b_0 & & +b_2 & +4b_3 & +9b_4 & = 0, \\ -b_0 & & +b_2 & +8b_3 & +27b_4 & = 6, \\ b_0 & & +b_2 & +16b_3 & +81b_4 & = 0, \end{array}$$

from which

$$b_0 = -\tfrac{3}{2}, \quad b_1 = 5, \quad b_2 = -6, \quad b_3 = 3, \quad b_4 = -\tfrac{1}{2}.$$

Thus,

$$y'''(x_1) = \frac{1}{2h^3}(-3y_0 + 10y_1 - 12y_2 + 6y_3 - y_4) + O(h^2).$$

Therefore,

$$\theta(x_1) = \frac{1}{12h^4}[(2 + 3hp_1)y_0 - (9 + 10hp_1)y_1 + 4(4 + 3hp_1)y_2$$
$$-2(7 + 3hp_1)y_3 + (6 + hp_1)y_4 - y_5] + O(h^2).$$

The formulae are similar at x_{N-1}: The system for the a_i is the same as before, and likewise for the system for the b_i, except for a sign change in the right-hand side. Thus, for $y^{(4)}(x_{N-1})$ we get the same formula as before, in which $y_0, y_1, \ldots, y_5$ is replaced by $y_N, y_{N-1}, \ldots, y_{N-5}$, while for $y'''(x_{N-1})$ all coefficients change sign. Consequently,

$$\theta(x_{N-1}) = \frac{1}{12h^4}[(2 - 3hp_1)y_N - (9 - 10hp_1)y_{N-1} + 4(4 - 3hp_1)y_{N-2}$$
$$-2(7 - 3hp_1)y_{N-3} + (6 - hp_1)y_{N-4} - y_5] + O(h^2).$$

11. For any two grid functions $v = \{v_n\}$, $w = \{w_n\}$, we have from (7.113), for $n = 1, 2, \ldots, N$,

$$\begin{aligned} \frac{h^2}{2}[(\mathcal{K}_h v)_n - (\mathcal{K}_h w)_n] &= -\frac{1}{2}(v_{n+1} - w_{n+1}) + v_n - w_n - \frac{1}{2}(v_{n-1} - w_{n-1}) \\ &+ \frac{h^2}{2}\left[f(x_n, v_n, \frac{v_{n+1} - v_{n-1}}{2h}) - f(x_n, w_n, \frac{w_{n+1} - w_{n-1}}{2h})\right] \\ &= -\frac{1}{2}(v_{n+1} - w_{n+1}) + v_n - w_n - \frac{1}{2}(v_{n-1} - w_{n-1}) + \frac{h^2}{2}\times \\ &\left[f_y(x_n, \bar{y}_n, \bar{z}_n)(v_n - w_n) + f_z(x_n, \bar{y}_n, \bar{z}_n)\frac{1}{2h}(v_{n+1} - w_{n+1} - (v_{n-1} - w_{n-1})\right] \\ &= a_n(v_{n-1} - w_{n-1}) + b_n(v_n - w_n) + c_n(v_{n+1} - w_{n+1}), \end{aligned}$$

where

$$a_n = -\tfrac{1}{2}[1 + \tfrac{1}{2}hf_z(x_n, \bar{y}_n, \bar{z}_n)],$$

$$b_n = 1 + \tfrac{1}{2}h^2 f_y(x_n, \bar{y}_n, \bar{z}_n),$$

$$c_n = -\tfrac{1}{2}[1 - \tfrac{1}{2}hf_z(x_n, \bar{y}_n, \bar{z}_n)].$$

By assumption, $\frac{1}{2}h|f_z| \le \frac{1}{2}h\bar{p} \le 1$, so that $a_n \le 0$, $c_n \le 0$, and

$$|a_n| + |c_n| = \tfrac{1}{2}[1 + \tfrac{1}{2}hf_z(x_n, \bar{y}_n, \bar{z}_n)] + \tfrac{1}{2}[1 - \tfrac{1}{2}hf_z(x_n, \bar{y}_n, \bar{z}_n)] = 1.$$

Furthermore,

$$b_n \ge 1 + \tfrac{1}{2}h^2\underline{q}.$$

We now have

$$b_n(v_n - w_n) = -a_n(v_{n-1} - w_{n-1}) - c_n(v_{n+1} - w_{n+1}) + \tfrac{1}{2}h^2[(\mathcal{K}_h v)_n - (\mathcal{K}_h w)_n],$$

hence

$$(1 + \tfrac{1}{2}h^2\underline{q})|v_n - w_n| \le |a_n||v_{n-1} - w_{n-1}| + |c_n||v_{n+1} - w_{n+1}|$$
$$+\tfrac{1}{2}h^2|(\mathcal{K}_h v)_n - (\mathcal{K}_h w)_n|.$$

Therefore, since $|a_n| + |c_n| = 1$,

$$(1 + \tfrac{1}{2}h^2\underline{q})|v_n - w_n| \le \|v - w\|_\infty + \tfrac{1}{2}h^2\|\mathcal{K}_h v - \mathcal{K}_h w\|_\infty, \quad n = 1, 2, \ldots, N.$$

We distinguish two cases:

Case I: $\|v - w\|_\infty = |v_{n_0} - w_{n_0}|,\ 1 \le n_0 \le N$.
Here,

$$(1 + \tfrac{1}{2}h^2\underline{q})|v_{n_0} - w_{n_0}| \le |v_{n_0} - w_{n_0}| + \tfrac{1}{2}h^2\|\mathcal{K}_h v - \mathcal{K}_h w\|_\infty,$$

from which

$$|v_{n_0} - w_{n_0}| \le \frac{1}{\underline{q}}\|\mathcal{K}_h v - \mathcal{K}_h w\|_\infty.$$

This proves (7.116), since $\frac{1}{\underline{q}} \le M$.

Case II: $\|v - w\|_\infty = |v_{n_0} - w_{n_0}|$, $n_0 = 0$ or $n_0 = N + 1$.
In this case, (7.116) is trivial, since $M \ge 1$.

12. By definition (7.113) of $\mathcal{K}_h u$, the system of equations is

$$\boldsymbol{g}(\boldsymbol{u}) = \boldsymbol{0}, \quad \boldsymbol{u} = [u_1, u_2, \ldots, u_N]^T,$$

where

$$g_n(\boldsymbol{u}) = -\frac{1}{2}\,u_{n+1} + u_n - \frac{1}{2}\,u_{n-1} + \frac{1}{2}\,h^2 f(x_n, u_n, \frac{u_{n+1}-u_{n-1}}{2h}),$$
$$n = 1, 2, \ldots, N,$$

$$u_0 = \alpha, \quad u_{N+1} = \beta.$$

Since $g_n(\boldsymbol{u})$ depends only on u_{n-1}, u_n, u_{n+1}, the Jacobian of $\boldsymbol{g}(\boldsymbol{u})$ is tridiagonal. Specifically,

$$\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}\right)_{n,n} = 1 + \frac{1}{2}\,h^2 f_y(x_n, u_n, \frac{u_{n+1}-u_{n-1}}{2h}), \quad n = 1, 2, \ldots, N,$$
$$\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}\right)_{n,n-1} = -\frac{1}{2}\,[1 + \frac{1}{2}hf_z(x_n, u_n, \frac{u_{n+1}-u_{n-1}}{2h})], \quad n = 2, \ldots, N,$$
$$\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}\right)_{n,n+1} = -\frac{1}{2}\,[1 - \frac{1}{2}hf_z(x_n, u_n, \frac{u_{n+1}-u_{n-1}}{2h})], \quad n = 1, \ldots, N-1.$$

Wherever u_0 and u_{N+1} occur on the right, they have to be replaced by α and β, respectively. Newton's method then proceeds as follows:

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}(\boldsymbol{u}^{[i]})\boldsymbol{\Delta}_i = -\boldsymbol{g}(\boldsymbol{u}^{[i]}),$$
$$\boldsymbol{u}^{[i+1]} = \boldsymbol{u}^{[i]} + \boldsymbol{\Delta}_i.$$

13. (a) The basis in question (cf. Ch. 2, Ex. 72, but note the difference in notation) is $u_\nu = B_\nu$, where

$$B_\nu(x) = \begin{cases} \dfrac{x - x_{\nu-1}}{\Delta x_{\nu-1}}, & x_{\nu-1} \le x \le x_\nu, \\ \dfrac{x_{\nu+1} - x}{\Delta x_\nu}, & x_\nu \le x \le x_{\nu+1}, \end{cases} \qquad \nu = 1, 2, \ldots, n,$$

and $\Delta x_\nu = x_{\nu+1} - x_\nu$. We have that $\boldsymbol{U} = ([u_\nu, u_\mu])$ is a symmetric tridiagonal matrix in $\mathbb{R}^{n\times n}$, with

$$[u_\nu, u_\nu] = \int_a^b (p(x)[u_\nu'(x)]^2 + q(x)[u_\nu(x)]^2)\mathrm{d}x$$
$$= \frac{1}{(\Delta x_{\nu-1})^2}\int_{x_{\nu-1}}^{x_\nu} (p(x) + (x - x_{\nu-1})^2 q(x))\mathrm{d}x$$
$$+ \frac{1}{(\Delta x_\nu)^2}\int_{x_\nu}^{x_{\nu+1}} (p(x) + (x_{\nu+1} - x)^2 q(x))\mathrm{d}x, \quad 1 \le \nu \le n,$$

and, since $u_\nu(x)u_{\nu+1}(x) \equiv 0$ on $[x_{\nu-1}, x_\nu] \cup [x_{\nu+1}, x_{\nu+2}]$,

$$[u_\nu, u_{\nu+1}] = \int_a^b [p(x)u'_\nu(x)u'_{\nu+1}(x) + q(x)u_\nu(x)u_{\nu+1}(x)]\mathrm{d}x$$
$$= \frac{1}{(\Delta x_\nu)^2}\int_{x_\nu}^{x_{\nu+1}} [-p(x) + q(x)(x_{\nu+1} - x)(x - x_\nu)]\mathrm{d}x, \quad 1 \le \nu \le n-1.$$

Furthermore, the vector $\boldsymbol{\rho}$ has components

$$\rho_\nu = \int_a^b r(x)u_\nu(x)\mathrm{d}x = \frac{1}{\Delta x_{\nu-1}}\int_{x_{\nu-1}}^{x_\nu} (x - x_{\nu-1})r(x)\mathrm{d}x$$
$$+\frac{1}{\Delta x_\nu}\int_{x_\nu}^{x_{\nu+1}} (x_{\nu+1} - x)r(x)\mathrm{d}x, \quad 1 \le \nu \le n.$$

(b) Using the trapezoidal rule to approximate the integrals in (a), we obtain

$$[u_\nu, u_\nu] \approx \frac{1}{2}\left\{\frac{p_{\nu-1} + p_\nu}{\Delta x_{\nu-1}} + \frac{p_\nu + p_{\nu+1}}{\Delta x_\nu} + (\Delta x_{\nu-1} + \Delta x_\nu)q_\nu\right\},$$
$$[u_\nu, u_{\nu+1}] \approx -\frac{1}{2}\frac{p_\nu + p_{\nu+1}}{\Delta x_\nu},$$
$$\rho_\nu \approx \tfrac{1}{2}(\Delta x_{\nu-1} + \Delta x_\nu)r_\nu,$$

where $p_\nu = p(x_\nu)$, $q_\nu = q(x_\nu)$, etc. The approximate linear system $\boldsymbol{U\xi} = \boldsymbol{\rho}$ then becomes

$$[u_1, u_1]\xi_1 + [u_1, u_2]\xi_2 = \rho_1,$$
$$[u_{\nu-1}, u_\nu]\xi_{\nu-1} + [u_\nu, u_\nu]\xi_\nu + [u_\nu, u_{\nu+1}]\xi_{\nu+1} = \rho_\nu, \quad \nu = 2, 3, \ldots, n-1,$$
$$[u_{n-1}, u_n]\xi_{n-1} + [u_n, u_n]\xi_n = \rho_n.$$

Thus, if we define $\xi_0 = \xi_{n+1} = 0$,

$$-\frac{1}{2}\frac{p_{\nu-1} + p_\nu}{\Delta x_{\nu-1}}\xi_{\nu-1} + \frac{1}{2}\left\{\frac{p_{\nu-1} + p_\nu}{\Delta x_{\nu-1}} + \frac{p_\nu + p_{\nu+1}}{\Delta x_\nu} + (\Delta x_{\nu-1} + \Delta x_\nu)q_\nu\right\}\xi_\nu$$
$$-\frac{1}{2}\frac{p_\nu + p_{\nu+1}}{\Delta x_\nu}\xi_{\nu+1} = \frac{1}{2}(\Delta x_{\nu-1} + \Delta x_\nu)r_\nu, \quad \nu = 1, 2, \ldots, n,$$

that is,

$$-\frac{p_{\nu-1} + p_\nu}{\Delta x_{\nu-1}(\Delta x_{\nu-1} + \Delta x_\nu)}\xi_{\nu-1} + \left\{\left(\frac{p_{\nu-1} + p_\nu}{\Delta x_{\nu-1}} + \frac{p_\nu + p_{\nu+1}}{\Delta x_\nu}\right)\frac{1}{\Delta x_{\nu-1} + \Delta x_\nu} + q_\nu\right\}\xi_\nu$$
$$-\frac{p_\nu + p_{\nu+1}}{\Delta x_\nu(\Delta x_{\nu-1} + \Delta x_\nu)}\xi_{\nu+1} = r_\nu, \quad \nu = 1, 2, \ldots, n.$$

Here, according to (7.140), $\xi_\nu = u(x_\nu)$, where $u(x) = \sum_{\nu=1}^n \xi_\nu u_\nu(x)$.

The same equations are obtained if one uses the finite difference approximations

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}y}{\mathrm{d}x}\right)\Big|_{x_\nu} \approx \frac{\frac{p_\nu+p_{\nu+1}}{2}\frac{y_{\nu+1}-y_\nu}{\Delta x_\nu} - \frac{p_{\nu-1}+p_\nu}{2}\frac{y_\nu-y_{\nu-1}}{\Delta x_{\nu-1}}}{\frac{1}{2}(\Delta x_{\nu-1}+\Delta x_\nu)}$$

in the differential equation (7.126) written down at $x = x_\nu$, $\nu = 1, 2, \ldots, n$.

14. See the text.

15. (a) We have (cf. (7.132))

$$[y, v] = (r, v), \quad \text{all } v \in S.$$

Furthermore (cf. Sect. 7.4.3), with $\hat{\boldsymbol{\xi}}$ denoting the solution of $\boldsymbol{U\xi} = \boldsymbol{\rho}$,

$$\begin{aligned}[u_S, v] &= \left[\sum_{\nu=1}^n \hat{\xi}_\nu u_\nu, \sum_{\mu=1}^n \xi_\mu u_\mu\right] = \sum_{\nu,\mu=1}^n \hat{\xi}_\nu \xi_\mu [u_\nu, u_\mu] \\ &= \sum_{\nu,\mu=1}^n \hat{\xi}_\nu \xi_\mu [u_\mu, u_\nu] = \sum_{\mu=1}^n \xi_\mu \sum_{\nu=1}^n [u_\mu, u_\nu]\hat{\xi}_\nu \\ &= \sum_{\mu=1}^n \xi_\mu \rho_\mu = \sum_{\mu=1}^n \xi_\mu (r, u_\mu) = (r, v).\end{aligned}$$

By subtraction,

$$[y - u_S, v] = 0, \quad \text{all } v \in S.$$

(b) Putting $v = u_S$ in (a), we get

$$[y, u_S] = [u_S, u_S] = \| u_S \|_E^2 .$$

Therefore,

$$\begin{aligned}\|y - u_S\|_E^2 &= [y - u_S, y - u_S] = [y, y] - 2[y, u_S] + [u_S, u_S] \\ &= \|y\|_E^2 - \|u_S\|_E^2.\end{aligned}$$

16. (a) By the optimal approximation property (cf. (7.148)), we have

$$\|y - u_S\|_E = \min_{u \in S} \|y - u\|_E.$$

Therefore, if Δ_1 is a refinement of Δ_2, it is clear that $S_2 \subset S_1$, hence

$$\min_{u \in S_1} \|y - u\|_E \le \min_{u \in S_2} \|y - u\|_E,$$

from which the assertion follows.

(b) By assumption, all Δx_ν for the subdivision Δ_2 are rational numbers, say

$$\Delta x_\nu = \frac{p_\nu}{q_\nu}, \quad \nu = 0, 1, 2, \ldots, n.$$

Then the uniform grid Δ_1 with $|\Delta_1| = h = \frac{1}{q_0 q_1 \cdots q_n}$ is a subgrid of Δ_2, and the assertion follows from (a).

17. We have, by (7.130),

$$\begin{aligned}
[u_\nu, u_\mu] &= \nu\mu\pi^2 p \int_0^1 \cos(\nu\pi x)\cos(\mu\pi x)\mathrm{d}x + q\int_0^1 \sin(\nu\pi x)\sin(\mu\pi x)\mathrm{d}x \\
&= \tfrac{1}{2}(p\pi^2\nu^2 + q)\delta_{\nu,\mu}, \\
(r, u_\nu) &= \int_0^1 r(x)\sin(\nu\pi x)\mathrm{d}x =: \rho_\nu.
\end{aligned}$$

The system $\boldsymbol{U}\boldsymbol{\xi} = \boldsymbol{\rho}$ (cf. (7.145)) is diagonal and has the solution

$$\hat{\xi}_\nu = \frac{2\rho_\nu}{p\pi^2\nu^2 + q}, \quad \nu = 1, 2, \ldots, n.$$

Thus (cf. (7.147)),

$$u_S(x) = 2\sum_{\nu=1}^{n} \frac{\rho_\nu}{p\pi^2\nu^2 + q}\sin(\nu\pi x), \quad 0 \le x \le 1.$$

One computes

$$\mathcal{L}u_S = 2\sum_{\nu=1}^{n} \rho_\nu \sin(\nu\pi x),$$

which is the nth partial sum of the Fourier sine expansion of $r(x)$.

If $r(x) = r = \text{const}$, then

$$\rho_\nu = r\int_0^1 \sin(\nu\pi x)\mathrm{d}x = \frac{r}{\nu\pi}(1 - (-1)^\nu) = \begin{cases} 0, & \nu \text{ even}, \\ \dfrac{2r}{\nu\pi}, & \nu \text{ odd}. \end{cases}$$

Thus,

$$\begin{aligned}
u_S(x) &= 2\sum_{\mu=1}^{\lfloor (n+1)/2 \rfloor} \frac{\rho_{2\mu-1}}{p\pi^2(2\mu-1)^2 + q}\sin((2\mu-1)\pi x) \\
&= \frac{4r}{\pi}\sum_{\mu=1}^{\lfloor (n+1)/2 \rfloor} \frac{\sin((2\mu-1)\pi x)}{(2\mu-1)[p\pi^2(2\mu-1)^2 + q]}.
\end{aligned}$$

18. We have

$$u(x) = \sum_{\mu=1}^{n} \xi_\mu B_\mu(x),$$

with $\{B_\mu\}$ the basis of hat functions (cf. Ch. 2, Ex. 72, but note the difference in notation). Fix the interval $[x_\nu, x_{\nu+1}]$ for some ν with $0 \le \nu \le n$. Then, for $x \in [x_\nu, x_{\nu+1}]$,

$$y'(x) - u'(x) = y'(x) - \xi_\nu B'_\nu(x) - \xi_{\nu+1} B'_{\nu+1}(x) = y'(x) - \frac{\xi_{\nu+1} - \xi_\nu}{\Delta x_\nu},$$

where $\xi_0 = \xi_{n+1} = 0$. Thus,

$$\max_{[x_\nu, x_{\nu+1}]} |y' - u'| = \max_{x_\nu \le x \le x_{\nu+1}} \left| y'(x) - \frac{\xi_{\nu+1} - \xi_\nu}{\Delta x_\nu} \right|.$$

The maximum on the right is minimized by taking

$$\frac{\xi_{\nu+1} - \xi_\nu}{\Delta x_\nu} = \frac{1}{2} \left(\min_{[x_\nu, x_{\nu+1}]} y' + \max_{[x_\nu, x_{\nu+1}]} y' \right),$$

in which case

$$\max_{x_\nu \le x \le x_{\nu+1}} \left| y'(x) - \frac{\xi_{\nu+1} - \xi_\nu}{\Delta x_\nu} \right| = \frac{1}{2} \operatorname{osc}_{[x_\nu, x_{\nu+1}]}(y').$$

Therefore,

$$\inf_{u \in S} \|y' - u'\|_\infty = \max_{0 \le \nu \le n} \max_{[x_\nu, x_{\nu+1}]} |y' - u'| = \frac{1}{2} \max_{0 \le \nu \le n} \operatorname{osc}_{[x_\nu, x_{\nu+1}]}(y'),$$

from which the assertion follows from (7.150). In particular, since by the mean value theorem of calculus,

$$\operatorname{osc}_{[x_\nu, x_{\nu+1}]}(y') = y''(\bar{x}_\nu) \Delta x_\nu$$

for some $\bar{x}_\nu \in (x_\nu, x_{\nu+1})$, we get

$$\|y - u_S\|_\infty \le \frac{1}{2} \sqrt{\frac{\bar{c}}{\underline{c}}}\, |\Delta| \|y''\|_\infty.$$

19. See the text.

20. Letting $u_S(x) = \sum_{\nu=1}^{n} \xi_\nu \sin(\nu\pi x)$, the collocation equations are

$$p \sum_{\nu=1}^{n} \xi_\nu (\nu\pi)^2 \sin(\nu\pi x_\mu) + q \sum_{\nu=1}^{n} \xi_\nu \sin(\nu\pi x_\mu) = r_\mu,$$

where $r_\mu = r(x_\mu)$, that is,

$$\sum_{\nu=1}^{n}(p\pi^2\nu^2 + q)\sin(\nu\pi x_\mu)\xi_\nu = r_\mu, \quad \mu = 1, 2, \ldots, n.$$

If $\boldsymbol{A} = [a_{\mu\nu}]$ is the matrix of this system, then

$$a_{\mu\nu} = (p\pi^2\nu^2 + q)\sin(\nu\pi x_\mu),$$

and $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix with positive entries on the diagonal, and

$$\boldsymbol{X} = [x_{\mu\nu}], \quad x_{\mu\nu} = \sin(\nu\pi x_\mu).$$

Since any trigonometric sine polynomial in S (of degre n) that vanishes at n distinct points in $(0, 1)$ must be identically zero, the homogeneous system with the matrix $\boldsymbol{X}$ has only the trivial solution, hence the matrix $\boldsymbol{X}$ is nonsingular. Therefore, also $\boldsymbol{A}$ is nonsingular. The system of linear equations is uniquely solvable for any set of distinct points $x_\mu \in (0, 1)$.

ANSWERS TO MACHINE ASSIGNMENTS

1. See the text.

2. (a) The length L must be larger than, or equal to, the distance from the point (0,0) to the point (1,1), that is,

$$L \geq \sqrt{2}.$$

   (b) The differential equation for $z = y'$ is

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \lambda\sqrt{1 + z^2}, \quad \frac{\mathrm{d}x}{\mathrm{d}z} = \frac{1}{\lambda\sqrt{1 + z^2}}\,.$$

   Integrating, one obtains

$$x = \frac{1}{\lambda}\int_0^z \frac{\mathrm{d}t}{\sqrt{1 + t^2}} + c = \frac{1}{\lambda}\sinh^{-1} z + c, \quad c = \text{const},$$

   hence

$$z = y' = \sinh\lambda(x + c_1).$$

   A second integration gives

$$y(x) = \frac{1}{\lambda}\cosh\lambda(x + c_1) + c_2. \tag{0}$$

   The two boundary conditions become

$$\frac{1}{\lambda}\cosh\lambda c_1 + c_2 = 0, \tag{1}$$

$$\frac{1}{\lambda}\cosh\lambda(1+c_1)+c_2=1. \tag{2}$$

The third condition is

$$\int_0^1\sqrt{1+(y')^2}\mathrm{d}x=\int_0^1\sqrt{1+\sinh^2\lambda(x+c_1)}\mathrm{d}x=\int_0^1\cosh\lambda(x+c_1)\mathrm{d}x$$
$$=\frac{1}{\lambda}\sinh\lambda(x+c_1)\Big|_0^1=\frac{1}{\lambda}(\sinh\lambda(1+c_1)-\sinh\lambda c_1)=L,$$

hence

$$\sinh\lambda(1+c_1)-\sinh\lambda c_1=\lambda L. \tag{3}$$

Subtracting (1) from (2), and using the addition theorem for the hyperbolic cosine and the half-angle formula for the hyperbolic sine, gives

$$2\sinh\lambda(c_1+\tfrac{1}{2})\sinh\tfrac{1}{2}\lambda=\lambda. \tag{4}$$

A similar calculation, applied to (3), gives

$$2\cosh\lambda(c_1+\tfrac{1}{2})\sinh\tfrac{1}{2}\lambda=\lambda L. \tag{5}$$

From the last two equations, c_1 can be eliminated by subtracting the square of the first from the square of the second. Using the identity $\cosh^2(\,\cdot\,)-\sinh^2(\,\cdot\,)=1$, we obtain

$$4\sinh^2\tfrac{1}{2}\lambda=\lambda^2(L^2-1).$$

This gives us a transcendental equation for λ,

$$\frac{\sinh\frac{1}{2}\lambda}{\frac{1}{2}\lambda}=\sqrt{L^2-1}. \tag{6}$$

Once this is solved for λ, we obtain from (5)

$$c_1=\frac{1}{\lambda}\cosh^{-1}\left(\frac{L}{\sqrt{L^2-1}}\right)-\frac{1}{2}, \tag{7}$$

and from (1)

$$c_2=-\frac{1}{\lambda}\cosh\lambda c_1. \tag{8}$$

For $L=\sqrt{2}$, we get $\lambda=0$ from (6), and thus, from the differential equation, $y''=0$, hence $y(x)$ is a linear function. The solution in this case is the straight line connecting the points (0,0) and (1,1). For all other values $L>\sqrt{2}$ we must solve (6) numerically. The equation has the form

$$f(x):=\sinh x-ax=0,\quad x=\tfrac{1}{2}\lambda,\ a=\sqrt{L^2-1}>1,$$

which is easily solved by Newton's method. Since the function f is convex and has a negative minimum at $\cosh^{-1} a$, Newton's method converges (essentially) monotonically for any starting value $x_0 > \cosh^{-1} a$. In the program below, we choose $x_0 = 1 + a \cosh a$, for which $f(x_0) > 0$ and x_0 therefore is to the right of the solution.

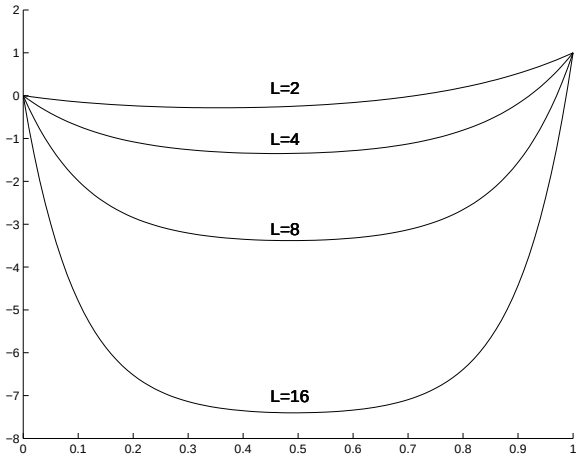
```
 PROGRAM

%MAVII_2B
%
f0='%8.0f %16.12f\n';
disp('       L       lambda')
for L=[2 4 8 16]
  a=sqrt(L^2-1);
  x1=1+acosh(a); err=1;
  while err>.5e-12
    x0=x1;
    x1=x0-(sinh(x0)-a*x0)/(cosh(x0)-a);
    err=abs(x1-x0);
  end
  lambda=2*x1;
  fprintf(f0,L,lambda)
end

 OUTPUT

>> MAVII_2B
       L       lambda
       2   3.826627375012
       4   6.434661004518
       8   8.400062862159
      16  10.182724373718
>>
```

Given λ, the solution $y(x)$ of the boundary value problem can be computed from (0), using the values (7) and (8) for c_1 and c_2. The routine

```
%PLOTMAVII_2B
%
lambda=[3.8266;6.4347;8.4001;10.1827];
L=1;
hold on
for iL=1:4
  L=2*L; lam=lambda(iL);
  c1=acosh(L/sqrt(L^2-1))/lam-1/2;
```

```
  c2=-cosh(lam*c1)/lam;
  x=.01*(0:100)';
  y=cosh(lam*(x+c1))/lam+c2
  plot(x,y)
end
hold off
```

produces the following plots.



(c) Replacing derivatives at the interior points x_n, in the differential equation as well as in the third boundary condition, by symmetric difference approximations, and the integral in the third boundary condition by its trapezoidal approximation, using one-sided difference approximations for the first derivatives at the boundary points, one gets the following $N+1$ nonlinear equations:

$$y_{n+1} - 2y_n + y_{n-1} = \lambda h^2 \sqrt{1 + \left(\frac{y_{n+1} - y_{n-1}}{2h}\right)^2},$$
$$n = 1, 2, \ldots, N,$$

$$\frac{1}{2}h\sqrt{1 + \left(\frac{y_1}{h}\right)^2} + h\sum_{n=1}^{N}\sqrt{1 + \left(\frac{y_{n+1} - y_{n-1}}{2h}\right)^2} + \frac{1}{2}h\sqrt{1 + \left(\frac{1 - y_N}{h}\right)^2} = L.$$

In the first, Nth, and last equation, y_0 and y_{N+1} have to be replaced by 0 and 1, respectively. The $N+1$ unknowns are $y_1, y_2, \ldots, y_N$ and $y_{N+1} = \lambda$.

As suggested, we apply a homotopy procedure to solve this system for given values of N and L, letting L vary from 1.5 in steps of .5 to 16, and for each N using as initial approximation $y_n^{[0]} = nh$, $n = 1, 2, \ldots, N$ and $\lambda^{[0]} = y_{N+1}^{[0]} = .1$. (Recall that $y_{N+1} = \lambda$. The initial approximation $\lambda^{[0]} = 0$ does not work since the third boundary condition with $y_n^{[0]}$ linear, as selected, and $L = \sqrt{2}$, is trivially satisfied, the trapezoidal formula being exact for linear functions.)

PROGRAMS

```
%MAVII_2C
%
global N L
f0='%8.0f %5.2f %16.12f %11.4e\n';
disp('       N   L     lambda approx.   rel. err')
lambda=[3.826627375012;6.434661004518; ...
  8.400062862159;10.182724373718];
LL=(1.5:.5:16)';
for N=[10 20 40]
  h=1/(N+1);
  y=zeros(N+1,1); y0=zeros(N+1,1);
  n=(1:N)'; y=[h*n;0.1];
  k=0;
  for iL=1:30
    y0=y; L=LL(iL);
    y=fsolve(@fMAVII_2C,y0);
    if abs(L-2)<.001 | abs(L-4)<.001 | abs(L-8)<.001 | abs(L-16)<.001
      k=k+1;
      err=abs((abs(y(N+1))-lambda(k))/lambda(k));
      fprintf(f0,N,L,abs(y(N+1)),err)
    end
  end
  fprintf('\n')
end

%FMAVII_2C
%
function F=fMAVII_2C(y)
global N L
F=zeros(N+1,1);
h=1/(N+1);
F(1)=y(2)-2*y(1)-y(N+1)*h^2*sqrt(1+(y(2)/(2*h))^2);
n=(2:N-1)';
F(n)=y(n+1)-2*y(n)+y(n-1)-y(N+1)*h^2*sqrt(1+((y(n+1) ...
         -y(n-1))/(2*h)).^2);
```

```
F(N)=1-2*y(N)+y(N-1)-y(N+1)*h^2*sqrt(1+((1-y(N-1))/(2*h))^2);
n=(2:N-1)';
F(N+1)=(h/2)*sqrt(1+(y(1)/h)^2)+h*sqrt(1+(y(2)/(2*h))^2) ...
  +h*sum(sqrt(1+((y(n+1)-y(n-1))/(2*h)).^2))+h*sqrt( ...
  1+((1-y(N-1))/(2*h))^2)+(h/2)*sqrt(1+((1-y(N))/h)^2)-L;

 OUTPUT

>> MAVII_2C
       N   L     lambda approx.   rel. err
      10  2.00   3.864419045954  9.8760e-03
          4.00   6.413401257132  3.3039e-03
          8.00   8.242762704347  1.8726e-02
         16.00   9.810703658590  3.6534e-02

      20  2.00   3.837100741244  2.7370e-03
          4.00   6.429312689528  8.3117e-04
          8.00   8.357705841555  5.0425e-03
         16.00  10.080684719015  1.0021e-02

      40  2.00   3.829392615403  7.2263e-04
          4.00   6.433294597785  2.1235e-04
          8.00   8.389010630497  1.3157e-03
         16.00  10.155963436331  2.6281e-03
>>
```

Actually, for some reason, the routine produced the negative eigenvalues, which in view of Eq. (6) are mathematically acceptable but have no physical meaning.

3. Condition (2) of Theorem 7.1.2 is violated since $f_{u_1} < 0$.

   The equation to be solved is

$$\phi(s) := u_1(1; s) = 0$$

   and the associated initial value problem (cf. the first Example of Sect. 7.2.1 and Ex. 3)

$$\begin{aligned}
\frac{\mathrm{d}u_1}{\mathrm{d}x} &= u_2, & u_1(0) &= 0,\\
\frac{\mathrm{d}u_2}{\mathrm{d}x} &= -\mathrm{e}^{u_1}, & u_2(0) &= s,\\
\frac{\mathrm{d}u_3}{\mathrm{d}x} &= u_4, & u_3(0) &= 0,\\
\frac{\mathrm{d}u_4}{\mathrm{d}x} &= -\mathrm{e}^{u_1}u_3, & u_4(0) &= 1.
\end{aligned}$$

Thus, Newton's iteration for s is

$$s^{[\nu+1]} = s^{[\nu]} - \frac{u_1(1; s^{[\nu]})}{u_3(1; s^{[\nu]})}\,, \quad \nu = 0, 1, 2, \ldots\,,$$

$$s^{[0]} = \begin{cases} 1 & \text{for first solution,} \\ 15 & \text{for second solution.} \end{cases}$$

PROGRAMS

```
%MAVII_3
%
f0='%8.0f %17.12f\n';
disp('      it        s^{nu}')
s=0;
snew=1;  % first solution
%snew=15;  % second solution
nu=0; fprintf(f0,nu,snew)
options=odeset('AbsTol',.5e-12,'RelTol',.5e-10);
while abs((snew-s)/snew)>.5e-12
  nu=nu+1;
  s=snew;
  y0=[0;s;0;1]; xspan=[0 1];
  [x,y]=ode45(@fMAVII_3,xspan,y0,options);
  r=size(y,1);
  phi=y(r,1); phiprime=y(r,3);
  snew=s-phi/phiprime;
  fprintf(f0,nu,snew)
end

%FMAVII_3
%
function yprime=fMAVII_3(x,y)
yprime=[y(2);-exp(y(1));y(4);-exp(y(1))*y(3)];
```

OUTPUT

```
>> MAVII_3
      it        s^{nu}         it        s^{nu}
       0    1.000000000000      0    15.0000000000
       1    0.537851334914      1    11.1188713195
       2    0.549346538918      2    10.8491143706
       3    0.549352728774      3    10.8468991758
       4    0.549352728776      4    10.8468990194
                                5    10.8468990194
>>
```

Hence, $y'(0) = .549352728776$ for the first solution, and $y'(0) = 10.8468990194$ for the second.

4. See the text.

5. (a) Put $r = \mathrm{e}^x$. Then differentiating $[y(r)]^2 = z(\ln r)$ with respect to r, we get $2y(r)y'(r) = z'(\ln r)/r$, hence

$$\frac{\mathrm{d}}{\mathrm{d}r}(ry(r)y'(r)) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}}{\mathrm{d}x}z(x)\right)\frac{\mathrm{d}x}{\mathrm{d}r} = \frac{1}{2}z''(x)\mathrm{e}^{-x},$$

and the differential equation becomes

$$\mathrm{e}^{-x}\cdot\tfrac{1}{2}z''(x)\mathrm{e}^{-x} + \rho(1-\sqrt{z}) = 0.$$

Thus, in terms of z, the boundary value problem is

$$z'' + 2\mathrm{e}^{2x}\rho(1-\sqrt{z}) = 0, \quad z(0) = \eta^2,\ z(\infty) = 1,$$

and the quantity of interest is

$$\sigma = \frac{z'(0)}{2\eta}.$$

(b) Any solution of the differential equation is defined only in the domain $z \geq 0$ for $x \geq 0$. It is obvious from the differential equation that in the domain $z > 1$ any solution has to be convex (i.e., $z'' > 0$), while in the domain $z < 1$ it must be concave. From this it follows that $s = z'(0)$ must be positive, since otherwise the solution will be concave from the start and decrease, thus eventually hitting the boundary $z = 0$. If $s > 0$, then three things can happen: (i) the solution remains concave all the time, reaching a maximum with $z < 1$, and then turning down to hit the boundary $z = 0$; (ii) the solution $z(x) = z(x; s)$ reaches the line $z = 1$ with a positive slope and at this time changes from concave to convex and grows to infinity; (iii) the solution reaches the line $z = 1$ with zero slope and, by uniqueness, must then remain constant equal to 1. The likelihood of (iii) seems small if not nil. Very likely, the desired solution has the line $z = 1$ as an asymptote as $x \to \infty$.

(c) From the analysis of (b), we must clearly choose $s > 0$. We call a solution $z(\,\cdot\,; s)$, $s > 0$, of type (i), (ii) or (iii), according as the situation (i), (ii) or (iii) of (b) occurs. The following algorithm then suggests itself.

Find a value of s, say $\underline{s}_0$, such that $z(\,\cdot\,; \underline{s}_0)$ is of type (i), and another value, $\overline{s}_0 > \underline{s}_0$ such that $z(\,\cdot\,; \overline{s}_0)$ is of type (ii) (we ignore the possibility of (iii)). Then, for $n = 0, 1, 2, \ldots$ do a kind of bisection procedure,

$$\left[\begin{array}{l}
\text{if } z(\,\cdot\,; \tfrac{1}{2}(\underline{s}_n + \overline{s}_n)) \text{ is of type (i) then} \\
\qquad \underline{s}_{n+1} = \tfrac{1}{2}(\underline{s}_n + \overline{s}_n),\ \overline{s}_{n+1} = \overline{s}_n \\
\quad \text{otherwise} \\
\qquad \underline{s}_{n+1} = \underline{s}_n,\ \overline{s}_{n+1} = \tfrac{1}{2}(\underline{s}_n + \overline{s}_n); \\
\text{if } \overline{s}_{n+1} - \underline{s}_{n+1} < \varepsilon \text{ then stop and compute } \sigma = \frac{\underline{s}_{n+1} + \overline{s}_{n+1}}{4\eta}\,.
\end{array}\right.$$

Here, ε is a tolerance on the desired accuracy. To determine whether a solution is of type (i), it suffices to check whether z strictly decreases for some x. The solution will be of type (ii) if $z > 1$ for some x. In the program below, the Runge-Kutta routine `RK4` of Ch. 5, MA 1 is used to carry out the integrations.

```
 PROGRAMS

%MAVII_5C
%
global rho
f0='%10.2f %6.2f %9.5f %9.5f\n';
f1='%17.2f %9.5f %9.5f\n';
disp('      rho    eta       s       sigma')
rrho=[.5;1;2;5;10]; sigma=zeros(9,5);
eps0=.5e-5;
h=.01; hs=.02;
for ir=1:5
  rho=rrho(ir);
  for ie=1:9
    eta=ie*.1;
%
% Determining a low and a high value of s
%
    s=0;
    while s<10
      s=s+hs;
      x=0; znext=[eta^2;s]; z=[0;0];
      while znext(1)>=z(1) & znext(1)<=1.1
        z=znext;
        znext=RK4(@fMAVII_5,x,z,h);
        x=x+h;
      end
      if znext(1)<1.1
        slow=s;
      else
        shigh=s;
        break
      end
    end
%
% Solving the boundary value problem by a bisection procedure
%
    while shigh-slow>eps0
      s=(slow+shigh)/2;
```

```
      x=0; znext=[eta^2;s]; z=[0;0];
      while znext(1)>=z(1) & znext(1)<=1.1
        z=znext;
        znext=RK4(@fMAVII_5,x,z,h);
        x=x+h;
      end
      if znext(1)<1.1
        slow=s;
      else
        shigh=s;
      end
    end
    sigma(ie,ir)=s/(2*eta);
    if ie==1
      fprintf(f0,rho,eta,s,sigma(ie,ir))
    else
      fprintf(f1,eta,s,sigma(ie,ir))
    end
  end
  fprintf('\n')
end
%
% Plot sigma vs eta
%
etp=.1:.01:.9;
hold on
et=(.1:.1:.9)';
for ir=1:5
  sigp=interp1(et,sigma(:,ir),etp,'spline');
  plot(etp,sigp);
  axis([0 1 0 23])
end
hold off

%FMAVII_5
%
function zprime=fMAVII_5(x,z)
global rho
zprime=[z(2);-2*exp(2*x)*rho*(1-sqrt(z(1)))];
```

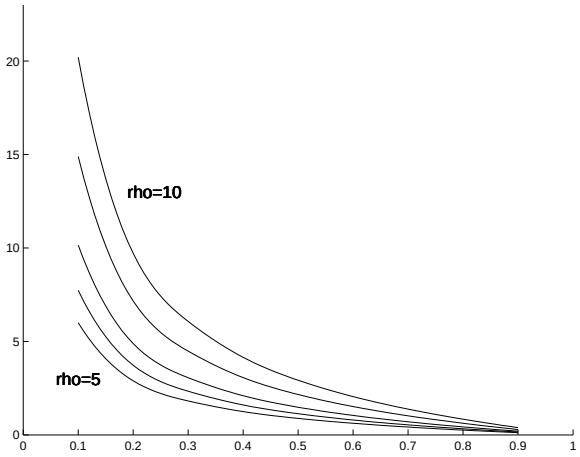
```
 OUTPUT

>> MAVII_5C
      rho    eta       s       sigma
      0.50   0.10   1.20039   6.00193
             0.20   1.15779   2.89447
             0.30   1.08952   1.81586
             0.40   0.99723   1.24654
             0.50   0.88226   0.88226
             0.60   0.74570   0.62142
             0.70   0.58847   0.42034
             0.80   0.41134   0.25709
             0.90   0.21499   0.11944

      1.00   0.10   1.54730   7.73650
             0.20   1.49122   3.72804
             0.30   1.40182   2.33637
             0.40   1.28159   1.60198
             0.50   1.13245   1.13245
             0.60   0.95596   0.79663
             0.70   0.75345   0.53818
             0.80   0.52601   0.32876
             0.90   0.27458   0.15255

      2.00   0.10   2.03059  10.15295
             0.20   1.95554   4.88885
             0.30   1.83655   3.06091
             0.40   1.67720   2.09650
             0.50   1.48033   1.48033
             0.60   1.24817   1.04014
             0.70   0.98261   0.70187
             0.80   0.68520   0.42825
             0.90   0.35728   0.19849

      5.00   0.10   2.97897  14.89485
             0.20   2.86642   7.16605
             0.30   2.68904   4.48173
             0.40   2.45268   3.06585
             0.50   2.16194   2.16194
             0.60   1.82044   1.51704
             0.70   1.43120   1.02228
             0.80   0.99667   0.62292
             0.90   0.51901   0.28834

     10.00   0.10   4.04104  20.20520
```

```
0.20   3.88631   9.71578
0.30   3.64333   6.07221
0.40   3.32054   4.15068
0.50   2.92455   2.92455
0.60   2.46054   2.05045
0.70   1.93280   1.38057
0.80   1.34487   0.84054
0.90   0.69974   0.38875
>>
```

Almost identical result are obtained with twice the step length (i.e., with $h = .02$), except for $\eta = .1$ when there are discrepancies in as many as three trailing digits.

The plots σ vs. η (see below) are obtained by interpolating the data above, using the Matlab routine `interp1` with the cubic spline option.

PLOTS



http://www.springer.com/978-0-8176-8258-3

Numerical Analysis
Gautschi, W.
2012, XXVI, 588p. 59 illus., Hardcover
ISBN: 978-0-8176-8258-3
A product of Birkhäuser Basel