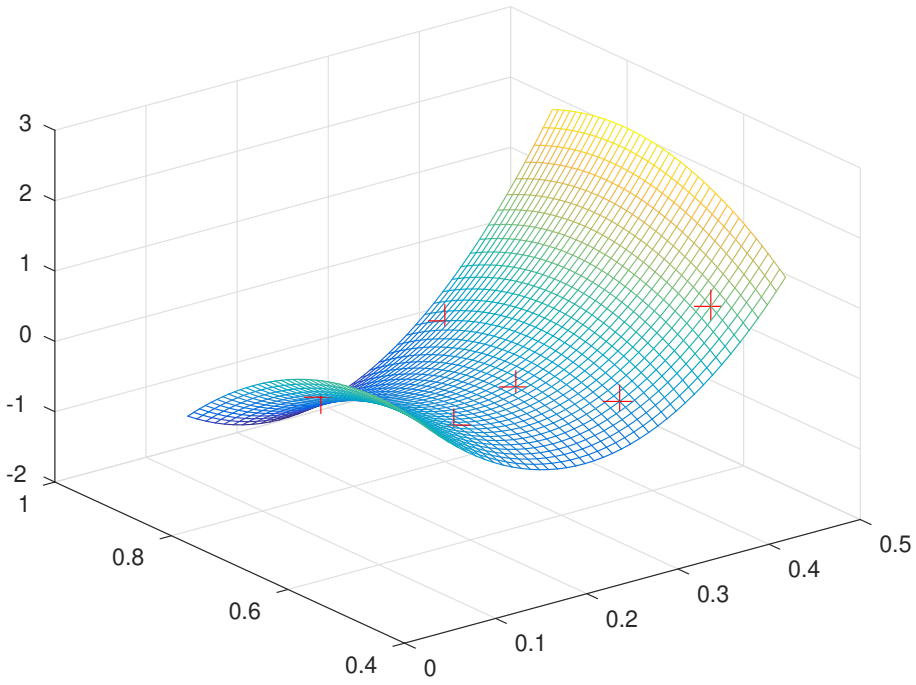
Excercises in Numerical Methods



Fredrik Berntsson (`fredrik.berntsson@liu.se`)

Linköpings universitet

Contents

1 Basic Concepts and Floating Point systems

Exercise 1.1 Let $a = 0.0987 \pm 0.5 \cdot 10^{-4}$ and $b = 20.104 \pm 4 \cdot 10^{-3}$. Determine the number of correct decimals and the number of significant digits for both a and b.

Exercise 1.2 Let $c_0 \approx \bar{c}_0 = 2.99792458 \cdot 10^6$ be correctly rounded. How many correct decimals and significant digits does the approximate value $\bar{c}_0$ have?

Exercise 1.3 Let $a = 22.73531443$. Round the value a correctly to 5 *significant digits* to obtain the approximation $\bar{a}$. Give both the approximate value $\bar{a}$ and an upper bound for the absolute error $|\Delta a|$ in the approximation.

Exercise 1.4 We approximate π by $\bar{\pi} = 3.1415$. How many correct decimals and significant digits do we have?

Hint $\pi = 3.1415926535\ldots$.

Exercise 1.5 Round the mathematical constant of the golden ratio $\phi = 1.61803398875\ldots$ to five correct decimals and also to five significant digits.

Exercise 1.6 Let $\bar{a} = 22.73531$ be an approximate value of a. The error in $\bar{a}$ is $|\Delta a| \leq 4.7 \cdot 10^{-3}$. How many *correct decimlals* and *significant digits* does the approximation $\bar{a}$ have?

Exercise 1.7 Let $y = \mathrm{e}^{\pi}$. Classify the following error sources in the computation of y: The rounding of π to 3.142 and the approximation $\mathrm{e}^x \approx 1 + x + \frac{1}{2}x^2$.

Exercise 1.8 Let $x = -102.232$. Give a bound for the *absolute error* when x is stored on a computer using the floating point system $(10, 3, -10, 10)$.

Exercise 1.9 Let $x = 168.3556541$. What is the closest number to x that exists in the floating point system $(10, 5, -100, 100)$.

Exercise 1.10 A single precision floating point number $x = (-1)^s(1.f)_2 2^{e-127}$ is stored using 32 bits assigned as follows

s (1 bit)	e (8 bits)	f (23 bits)

Clearly show how the numbers $x = 9.25$ and $x = -2.65625$ are stored.

Exercise 1.11 Consider the floating point number system $(10, 5, -9, 9)$, where $\beta = 10$ is the base and $t = 5$ is the number of digits in the fractional part. Let $x = 117.5614$ and $y = 0.01678214$. Find the closest numbers x_r and y_r in the floating point system.

Exercise 1.12 Suppose we represent a number x in the form $x = m \times \beta^e$. What do we call m, β and e? Also, What else must be specified in order to properly define a floating point number system?

Exercise 1.13 Consider the floating point system $(10, 2, -9, 9)$. Let $a = 8.50 \cdot 10^5$ and $b = 5.25 \cdot 10^2$. Compute the floating point results $\mathrm{fl}[a \cdot b]$ and $\mathrm{fl}[a/b]$. In both cases also give a bound for the relative error in the result.

Exercise 1.14 Rewrite the expressions $\sqrt{1+x} - 1$, $(1-x)^{-1} - (1+x)^{-1}$, and $1 - \cos^2(x)$ in such a way that the cancellation is avoided.

Exercise 1.15 Let $\sqrt{101} = 10.04988$ be correctly rounded so that the error is at most $0.5 \cdot 10^{-5}$. What is the resulting absolute error in $x = \sqrt{101} - 10$? Also propose an alternative formula for computing x that leads to a smaller error.

2 Error Analysis

Exercise 2.1 Let $f = (x - y)/z$, where $x = 8.25$, $y = 1.05$ and $z = 4.00$ are correctly rounded. Compute an approximate value for f together with a bound for the absolute error.

Exercise 2.2 Let $f(x) = \mathrm{e}^{bx}$, where $b = 1.70 \pm 0.01$. Compute $f(2)$ and also use the error propagation formula to give an error bound.

Exercise 2.3 The area of a circle is $A = \pi r^2$. Compute the area, and an error bound, for the case when $r = 23.76 \pm 0.02$ and we approximate π by 3.142.

Exercise 2.4 Let $y = a\mathrm{e}^b$, where $a = 1.54 \pm 0.03$ and $b = 3.17 \pm 0.05$. Compute the approximate value $\bar{y}$ and an error bound.

Exercise 2.5 The focal point of a lens can be determined by the formula

$$\frac{1}{f} = \frac{1}{a} + \frac{1}{b}, \text{ where } a = 32 \pm 1 \text{ and } b = 46 \pm 1.3.$$

Determine $f(a, b)$ and an error bound.

Exercise 2.6 We compute the function

$$f(x) = \sqrt{1 + x} - \sqrt{1 - x}$$

for small x values on a computer with unit round off $\mu = 1.11 \cdot 10^{-16}$. We find that the results are quite poor and that the *relative error* in the result tends to grow as $x \to 0$. Explain the poor accuracy by performing an analysis of the computational errors and give a bound for the relative error in the computed result $f(x)$. For the analysis you may assume that all computations are performed with a relative error at most μ.

Exercise 2.7 We want to evaluate a function $f(x)$ on a computer, for small values of x, and have two alternate expressions:

$$f_1(x) = \frac{1 - \cos(x)}{\sin(x)} \text{ or } f_2(x) = \frac{\sin(x)}{1 + \cos(x)}$$

For the case $x = 1.111 \cdot 10^{-8}$ we evaluate both expressions in Matlab and obtain $f_1 = 9.9930\ldots 10^{-9}$ and $f_2 = 5.5550\ldots 10^{-9}$.

Assume that all numerical computations are performed with a relative error of at most the unit round off μ and perform an analysis of the computational errors. Derive a bound for both the absolute error in the results.

Hint The unit round off for Matlab is $\mu = 1.11 \cdot 10^{-16}$.

Exercise 2.8 We compute the function

$$f(x) = \mathrm{e}^x - 3x$$

for small x values on a computer with unit round off $\mu = 1.11 \cdot 10^{-16}$. Preform an analysis of the computational errors to obtain a bound for the relative error in the computed results $f(x)$. For the analysis you may assume that all computations are performed with a relative error at most μ. Also, use the obtained bound to argue if *cancellation* occurs during the computations. In case of cancellation also suggest an alternative formula that can be expected to give better accuracy.

Exercise 2.9 We want to evaluate the function

$$f(x) = \frac{x - \sin(x)}{x^3}$$

for small values of x on a computer with the unit round off $\mu = 1.11 \cdot 10^{-16}$. A Taylor series expansion shows that

$$\lim_{x \to 0} f(x) = \frac{1}{6},$$

but when we compute the expression for $x = 10^{-7}$ we get the difference $|\bar{f}(10^{-7}) - \frac{1}{6}| \approx 5.4 \cdot 10^{-3}$. Explain the above result by performing an error analysis that clearly shows how large the error is when $f(x)$ is evaluated on the computer. For the analysis you should assume that all computations are carried out with a relative error of at most μ.

Hits The Taylor series expansion of $\sin(x)$ is $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$

Exercise 2.10 We compute the function

$$f(x) = 1 - 2x\cos(x)$$

for small x values on a computer with unit round off $\mu = 1.11 \cdot 10^{-16}$. Preform an analysis of the computational errors to obtain a bound for the relative error in the computed results $f(x)$. For the analysis you may assume that all computations are performed with a relative error at most μ. Also, use the obtained bound to argue if *cancellation* occurs during the computations. In case of cancellation also suggest an alternative formula that can be expected to give better accuracy.

Exercise 2.11 Assume that we have an approximate value $\bar{x}$ and want to find the resulting error Δf. The general error propagation formula states that

$$|\Delta f| \lesssim |\frac{\partial f}{\partial x}||\Delta x|.$$

If $|f'(\bar{x})| \approx |f'(x)| = 0$ the formula fails. Show that for this case it is more resonable to use

$$|\Delta f| \lesssim |f''(\bar{x})|\frac{|\Delta x|^2}{2}.$$

3 Non-linear equations

Exercise 3.1 Show that the function $f(x) = \mathrm{e}^x - \frac{4}{2+x}$ has a root in the interval $[0, 1]$.

Exercise 3.2 The equation $\mathrm{e}^x = 2(1-x)$ has a root $x^* \approx 0.3$. Does the fixed point iteration $x_{k+1} = 1 - \frac{1}{2}\mathrm{e}^{x_k}$ converge to the root?

Exercise 3.3 We have solved the equation $f(x) = x^3 + x - 7 = 0$ and obtained an approximate root $\bar{x} = 1.7$. Estimate the error in the approximation $\bar{x}$.

Exercise 3.4 We have solved $f(x) = 0$ using Newtons method and obtained an approximate root $\bar{x}$ such that $f(\bar{x}) = 0$ when we evaluate the function on the computer. Suppose we can compute $f(x)$ with an absolute error at most 10^{-8} and that $1.7 \leq f'(x) \leq 2.2$ near the root x^*. Estimate the error in $\bar{x}$ for this case.

Exercise 3.5 Determine the positive root of the equation $x = 5(1-\mathrm{e}^{-x})$ with five correct decimals.

Exercise 3.6 We are intressted in solving the equation $x = 2\sin(x)$. Two possible fixed point iterations are obtained by using the iteration functions $\phi_1(x) = 2\sin(x)$ or $\phi_2(x) = \frac{x}{2} + \sin(x)$. Which of these fixed point iterations would have the fastest convergence if $x^* \approx 1.9$?

Exercise 3.7 Consider the equation $f(x) = 2\cos(x) - 4x$ which has a positive root $x^* \approx 0.45$. Show theoretically that the fixed point iteration $x_{k+1} = \phi(x_n) = \frac{1}{2}\cos(x_n)$ is convergent for any starting value x_0. Also, if we use $x_0 = 0.45$ then we get

k	x_k
0	0.4502236
1	0.4501749
2	0.4501855
3	0.4501832

Estimate the error in the approximation $\bar{x} = 0.4501832$.

Exercise 3.8 The equation $f(x) = x - 3e^{-x} = 0$ has a solution $x^* \approx 1.05$.

a) Estimate the error in the approximation $\bar{x} = 1.05$.

b) Investigate the following fixed point iterations theoretically

$$(i)\ x_{n+1} = 3e^{-x_n}, \quad (ii)\ x_{n+1} = (2x_n + 3e^{-x_n})/3,$$

$$(iii)\ x_{n+1} = 1.05x_n + 3e^{-x_n}, \quad (iv)\ x_{n+1} = (x_n + 3e^{-x_n})/2.$$

Determine if they converge towards x^* and also find out which of the methods that have the fastest convergence.

c) Estimate the number of iterations that would be needed for the fastest of the methods from **b)** if $x_0 = 1.05$ and we want to find the root with an absolute error of at most 10^{-10}.

Exercise 3.9 Show that Newtons method is convergent when applied to the equation $f(x) = x^2 = 0$ and if the starting guess is $x_0 = 1$. Also show that the order of convergence is $p = 1$.

Exercise 3.10 Show that Newton-Raphsons has a quadratic rate of convergence if x^* is a single root.

Exercise 3.11 The equation $f(x) = x^3 - 7.5x^2 + 18x - 14 = 0$ has a double root x^*. Using the secant method we have obtained $\bar{x} = 1.99789$. Derive the error estimate

$$|\bar{x} - x^*|^2 = 2\left|\frac{f(\bar{x})}{f''(\xi)}\right|, \quad \xi \in (\bar{x}, x^*),$$

which is valid for double roots. Also compute the error bound for the approximate root $\bar{x} = 1.99789$.

Exercise 3.12 Consider the equation $f(x) = \cos(x) - x\mathrm{e}^x = 0$. We use the Newton-Raphson method for finding the root, with the starting guess $x_0 = 1$, and obtain the following table

k	x_k	$f(x_k)$
0	1.0000000	$-2.2 \cdot 10^0$
1	0.6530794	$-4.6 \cdot 10^{-1}$
2	0.5313434	$-4.2 \cdot 10^{-2}$
3	0.5179099	$-4.6 \cdot 10^{-4}$
4	0.5177574	$-5.9 \cdot 10^{-8}$

We decide to use $\bar{x} = 0.5178$ as an approximation of x^*. Estimate the error in the approximation $\bar{x}$. Also state the definition of the *order of convergence* for an iterative method and use the table above to estimate the order of convergence for the Newton-Raphson method.

Exercise 3.13 The Newton-Raphsson method is used to find a root of the cubic polynomial $f(x) = x^3 - 9x^2 + 24x - 20$. We obtain the following iteration sequence

k	x_k	$\lvert x_k - x^*\rvert$
0	1.8000	0.2000
1	1.8970	0.1030
2	1.9476	0.0524
3	1.9736	0.0264
4	1.9867	0.0133

State the definition of order of convergence p for an iterative method. Also use the table to determine the order of convergence when Newton-Raphson's method is applied to this specific function $f(x)$. Also, use the results and known properties of Newton-Raphson's method to determine if $x^* = 2$ is a double or single root.

Exercise 3.14 We wish to implement the standard function $\sqrt{\cdot}$ on a computer using double precision arithmetic. We compute $x = \sqrt{a}$ by solving the equation $f(x) = x^2 - a$.

a) Derive the iteration formula $x_k = \varphi(x_{k-1})$ obtained by applying Newton-Raphsons method to the above equation.

b) Explain clearly why it is sufficient to consider the case $1 \leq a < 4$, and thus $1 \leq x < 2$.

c) A convergence analysis for the Newton-Raphson method leads to the estimate,

$$|x_k - \sqrt{a}| \lesssim \frac{1}{2}|\varphi''(\sqrt{a})||x_{k-1} - \sqrt{a}|^2.$$

Take advantage of this and determine the number of iterations needed to achive an error bound $|x_k - \sqrt{a}| \leq \mu = 1.1 \cdot 10^{-16}$ if $x_0 = 1.5$ is used.

d) We wish to decrease the number of iterations, by picking a better starting guess, and select evenly spaced numbers $a_j = 1 + 3j/n$, $j = 0, \ldots, n-1$, and compute $\sqrt{a_j}$ exactly. Clearly demonstrate how the values $(\sqrt{a_j}, a_j)$ can be used to obtain a better starting guess. What table size n do we need to reduce the number of iterations by one compared to the result in **c)**?

4 Polynomial and Spline Interpolation

Exercise 4.1 Let $p_n(x)$ be a polynomial of degree n. How many interpolation conditions of the type $p_n(x_i) = f_i$ are needed for $p_n(x)$ to be uniquely determined?

Exercise 4.2 Let the following table with correctly rounded function values be given

x	0.9	1.1	1.2
$f(x)$	0.4710	0.2452	0.2385

Find an approximate value for $f(1.03)$. Also provide a complete error estimate.

Exercise 4.3 In an application we need to implement the function $y = \log(x)$, for $1 \leq x \leq 4$, with a maximum error $\varepsilon \leq 10^{-5}$. We decide to use linear interpolation and create a table $\{x_i, y_i)\}_{i=1}^{n}$, where $x_1 = 1$, $x_n = 4$ and $h = x_{i+1} - x_i$ is the stepsize. In the table we store approximate values $y_i \approx \log(x_i)$, rounded to 6 correct digits. Determine the smallest size n for the table so that the maximum error in the interpolated values is less than 10^{-5}.

Exercise 4.4 Let $p(x)$ be the linear polynomial that interpolates the function $f(x) = \sin(x)$ at the points $x = 0$ and $x = 1$. Show that the truncation error is bounded by $|p(x) - \sin(x)| \leq \frac{1}{8}$, for $0 < x < 1$.

Exercise 4.5 A table with correctly rounded function values is given.

x	0.0	0.5	1.0	1.5
$f(x)$	1.80	2.80	4.10	5.90

Use linear interpolation to find an approximation of $f(0.4)$ and also give an error estimate.

Exercise 4.6 The following table is given

x	0.6	0.7	0.8	0.9
$f(x)$	1.23	1.29	1.32	1.07

Use quadratic interpolation and compute an approximate value for $f(0.74)$. Also estimate the truncation error in the result.

Exercise 4.7 Let x_1, x_2, x_3 and x_4 be given interpolation points. In the Lagrange interpolation formula we use basis functions $\ell_i(x)$ such that $\ell_i(x_j) = 1$ if $i = j$ and zero otherwise. Give an explicit expression for the basis function $\ell_2(x)$ for the case with $n = 4$ interpolation points. What is the degree of the basis polynomial?

Exercise 4.8 Use Lagrange interpolation to find the polynomial of degree 2 that interpolates the table

x	1	2	3
$f(x)$	1.3	0.6	1.9

Exercise 4.9 Let $p(x) = c_0 + c_1x + c_2x^2 + c_3x^3$ be a cubic polynomial. We want to find values for the coefficients so that $p(0) = p(1) = 0$ and $p'(0) = p'(1) = 1$. Show how to derive a linear system of equations such that the solution $c = (c_0, c_1, c_2, c_3)^T$ are the coefficients of a cubic polynomial satisfying these conditions. Also find the specific polynomial satisfying all the above conditions.

Exercise 4.10 Spline interpolation can be used to approximnate a function $y = f(x)$. We have a table

x	-2	-1	0	1	2
$f(x)$	0	1	3	1	0

We attempt to approximate $f(x)$ by a cubic spline $s(x)$. Clearly state the conditions that have to be satisfied for $s(x)$ to be a cubic spline that interpolates the above table. Is the given information sufficient for the spline $s(x)$ to be uniquely determined?

Exercise 4.11 Let

$$s(x) = \begin{cases} x + 1 & 0 \le x < 1, \\ x^3 - 3x^2 + 4x & 1 \le x < 2. \end{cases}$$

Is $s(x)$ a cubic spline?

Exercise 4.12 Let

$$s(x) = \begin{cases} ax + 1 & 0 \le x < 1, \\ bx^3 + cx^2 & 1 \le x < 2. \end{cases}$$

Determine the constants a, b and c so that $s(x)$ is a cubic spline.

Exercise 4.13 Approximate the function $f(x) = x^3+x^2+1$ by a cubic spline $s(x)$ that interpolates $f(x)$ at the nodes $x = 0, 0.3, 0.6, 0.7$ and 1.0. Use correct end point conditions, i.e. $s'(0) = f'(0)$ and $s'(1) = f'(1)$. Give the expression for $s(x)$.

Exercise 4.14 Consider a case where $s(x)$ is defined by two cubic polynomials,

$$s(x) = \begin{cases} s_1(x) = 0.9 + 0.1x + 0.6x^2 + ax^3, & 0 \le x < 1, \\ s_2(x) = 2.0 + b(x-1) + c(x-1)^2 + 0.4(x-1)^3, & 1 \le x \le 2. \end{cases}$$

Find the appropriate values for the constants a, b and c so that $s(x)$ is a cubic spline that interpolates the table tabellen:

x	0	1.0	2
s(x)	0.9	2.0	6.7

with the end point conditions $s'(0) = 0.1$ and $s'(2) = 7.3$.

Exercise 4.15 A function $s(x)$ is given by two cubic polynomials

$$s(x) = \begin{cases} s_1(x) = 0.8 + 0.2x - 0.4x^2, & 0 \le x < 1, \\ s_2(x) = 0.6 - 0.6(x-1) - 0.4(x-1)^2 - 0.4(x-1)^3, & 1 \le x \le 2. \end{cases}$$

Is $s(x)$ a cubic spline? Present the calculations. Also determine if $s(x)$ is a natural cubic spline?

Exercise 4.16 A function $f(x)$ can be approximated by a piecewise polynomial $s(x)$ on the interval $[a, b]$ by introducing evenly spaced nodes

$$a = x_0 < x_1 < x_2 < \ldots < x_N = b.$$

On each subinterval $[x_k, x_{k+1}]$ we let $s(x)$ be given by a cubic polynomial

$$s_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3, \qquad x_k \le x < x_{k+1}.$$

Clearly formulate the conditions that needs to be satisfied for $s(x)$ to be a cubic spline, defined on $[a, b]$, and that interpolates $f(x)$ in the nodes $\{x_k\}_{k=0}^N$.

Also illustrate the case $N = 3$ and draw a sketch that clearly illustrates the *nodes*, *interpolation points* and *polynomials*.

Exercise 4.17 In order to obtain a unique interpolating spline $s(x)$ we use *correct end point conditions*, i.e. $s'(a) = f'(a)$ and $s'(b) = f'(b)$. We experiment with different number of nodes N and measure the maximnum error $\max |s(x) - f(x)|$ on the interval. This gives us the table

N	5	10	20
$\max \lvert s(x) - f(x) \rvert$	0.0435	0.00269	0.000172

We know that the maximum error should depend on the step size $h = \max |x_{k+1} - x_k|$ as Ch^p, where p is an integer and C is a constant. Use the numbers in the table to determine p.

Exercise 4.18 Construct a linear spline that interpolates the table

x	1	3	4
$f(x)$	1.56	2.31	1.97

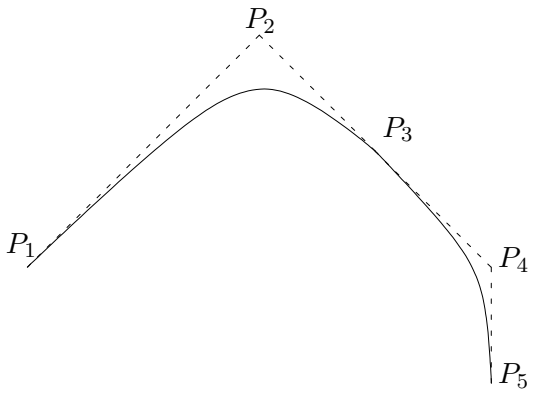
Exercise 4.19 A quadratic Beziér curve is given by the expression

$$p(t) = (1-t)^2 P_1 + 2(1-t)tP_2 + t^2 P_3, \quad 0 < t < 1,$$

where P_1, P_2 and P_3 are control points.

a) Show that the tangent, for $t = 0$ is parallell to the vector $P_2 - P_1$.

b) Suppose we want to put together two quadratic Beziér curves. We select five control points according to the sketch:



The point P_3 is common for both curve segments. We have $P_2 = (2\,,\,6)^T$, $P_3 = (3\,,\,5)^T$ and $P_5 = (6\,,\,1)$. Clearly show how to pick the point P_4 so that the tangent direction of the curve is continuous in at P_3 and so that the tangent direction at P_5 is vertical.

Exercise 4.20 $P_1 = (1\,,\,0)^T$, $P_2 = (1\,,\,3)^T$, $P_3 = (4\,,\,3)^T$ and $P_4 = (4\,,\,2)^T$. Draw a sketch that clearly shows the convex hull formed by these points. Also use the available information to draw the cubic Beziér curve formed by the four points $P_1, \ldots, P_4$ as accurately as possible.

Exercise 4.21 A cubic Beziér curve is given by

$$p(t) = (1-t)^3 P_1 + 3(1-t)^2 tP_2 + 3(1-t)t^2 P_3 + t^3 P_4, \quad 0 < t < 1,$$

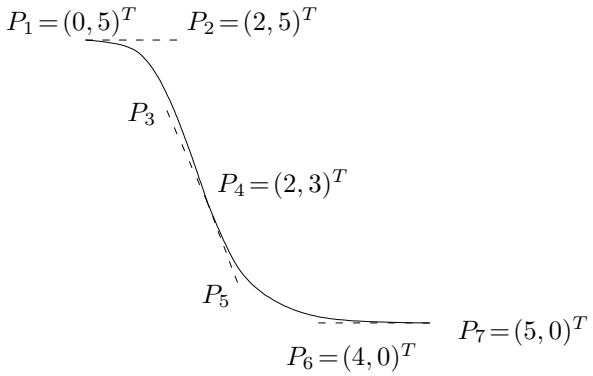
where P_1, P_2, P_3 and P_4 are control points.

a) Show that the tangent of the curve in the starting points $t = 0$ is parallel to the vector $P_2 - P_1$.

b) Give the definition of the *convex hull* formed by the points P_1, P_2, P_3 and P_4. Also show that the cubic Beziér curve is located within the convex hull formed by its control points P_1, P_2, P_3 and P_4.

c) Let $P_1 = (0\,,\,0)^T$, $P_2 = (1\,,\,3)^T$, $P_3 = (4\,,\,2)^T$ and $P_4 = (5\,,\,1)^T$. Also let $s(t)$ be the cubic Beziér curve given by these control points. Compute $s(1/2)$ and use the available information to draw a sketch that, as accurately as possible, shows the shape of the curve $s(t)$.

Exercise 4.22 Create a parametric curve composed of two cubic Beziér curves as shown in the figure



Chose the points P_3 and P_5 so that the curve has a continuous tanget direction at P_4. Also make sure the slope is exactly -2, e.g. the line from P_3 to P_5 can be expressed as $y = -2x + b$ for some constant b.

5 Integration, Differentiation and Extrapolation

Exercise 5.1 Suppose a function $f(x)$ have the required number of continuous derivatives. We want to approximate the derivative $f'(x)$ by the difference formula

$$Df(x) = \frac{1}{h}(f(x+h) - f(x)).$$

Show that the truncation error can be written as $f'(x) - Df(x) \approx Ch$. Give an expression for C.

Exercise 5.2 The the function $f(x)$ be known at the points $\{x_k\}$ by the table

x	0.0	0.25	0.5	0.75	1.0
$f(x)$	1.000	1.015	1.006	0.882	0.670

with correctly rounded function values. Use the Trapezoidal method to compute an approximation of the integral

$$I = \int_0^1 f(x)dx,$$

using the stepsize $h = 0.25$. Estimate the error in the result.

Exercise 5.3 Suppse a function $f(x)$ can be computed with a *relative error* at most ε. Derive a bound for the resulting error when we want to compute an integral

$$I = \int_a^b f(x)dx,$$

using Simpsons rule. Also discuss if there are any cases where the error bound indicates that the the computation may be problematic.

Exercise 5.4 To compute the derivative $f'(2)$ we can use the formula

$$Df(2) = \frac{1}{2h}(-f(x+2h) + 4f(x+h) - 3f(x)).$$

When the formula is applied for a few different h values we obtain the results

h	0.2	0.1	0.05
error	0.342	0.0861	0.0209

Assume that the error is proportional to h^p and use the table to determine p.

Exercise 5.5 Suppose $F_1(h) = a + bh^{p_1} + ch^{p_2}$, for some constants a, b and c and positive integers p_1 and p_2 such that $p_1 < p_2$. Show that

$$F_1(h) + \frac{F_1(h) - F_1(qh)}{q^{p_1} - 1} = a + \mathcal{O}(h^{p_2}).$$

Exercise 5.6 We compute an approximation of an integral using a numerical method $T(h)$ with a truncation error $R_T \approx Ch^4$, where C is a constant. We use two different values for h and obtain

h	0.1	0.05
T(h)	1.6713957	1.6712783

Use the table to estimate the truncation error in the computed value $T(0.05)$.

Exercise 5.7 A numerical method has a truncation error that can be written as $R_T = Ch^p$, where C is a constant, $p > 0$ is an integer, and h is a discretization parameter. The method computes a value $T(h)$ which approximates the exact value T_0. We compute $T(h)$ for a few different h-values to obtain the table

h	0.4	0.2	0.1	0.05
T(h)	3.100	2.701	2.604	2.578

Use the table to determine p.

Exercise 5.8 Compute an approximation of $\int_0^1 f(x)dx$ using the following table with correctly rounded function values

x	0	1/4	1/2	3/4	1
$f(x)$	1.5000	1.2412	1.0713	0.9663	0.9073

by the Trapezoidal method. Use the step size $h = 1/4$. Also estimate the error in the approximate value.

Exercise 5.9 We use a numerical method to compute the derivative of a function $f(x)$ and obtain

h	0.1	0.05	0.025
derivative	0.69280	0.71195	0.72171
error	-0.0388	-0.0196	-0.0099

The dominating source of error is the truncation error R_T which can be assumed to depend on h as $R_T \approx Ch^p$. Use the table to compute both C and p. Also determine the largest possible step size that can be used if we require that $|R_T| \leq 10^{-8}$.

Exercise 5.10 We want to compute an approximation of

$$\int_1^4 f(x)dx, \quad \text{where } f(x) = \cos(x^2)\sqrt{4-x}.$$

Do you expect the Trapezoidal method to work well for this case? Explain your conclusion clearly.

Exercise 5.11 We want to compute the value of the integral

$$I = \int_0^1 \mathrm{e}^{1-x^2} dx,$$

using Simpsons rule. We use a few different step sizes h and obtain

h	1/2	1/4	1/8
S(h)	2.06246	2.03208	2.03021

(a) Suppose $S(h) = I + R_T$, where $R_T \approx c \cdot h^p$ for constants c and p. Clearly demonstrate how the table can be used to determine values for c and p. Answer with both a formula and the resulting values for c and p.

(b) Suppose we want to find an approximate value for the integral with a total error $|R_T| < 10^{-6}$. What step size would be required? Motivate your answer.

6 Differential Equations

Exercise 6.1 Consider the problem

$$\begin{cases} y' = y^2 - t. \\ y(0) = 1, \end{cases}$$

Use the Euler method, and the stepsize $h = 0.5$, to compute an approximation of $y(0.5)$. With the stepsize $h = 0.1$ the Euler method gives an approximation $y(0.5; h = 0.1) \approx 1.668$. Use this information to estimate the truncation error in the approximation $y(0.5; h = 0.5)$.

Exercise 6.2 Consider the ordinary differential equation $y' = t^2 - y$, with $y(0) = 2$. Compute approximations of the solution $y(0.2)$ using the Euler method and stepsizes $h = 0.2$ and $h = 0.1$. Also estimate the error in the approximation $y(0.2; h = 0.1)$.

Exercise 6.3 Consider the problem $y' = -100y + \sin^2(t)$, $y(0) = 0$. Determine the maximum stepsize h such that the Euler method is stable when applied to the problem.

Exercise 6.4 We want to solve the problem $y' = -25y+t+0.04$, $y(0) = 1$, using the Euler method. We require that thre solution is stable. Find the largest possible time step h that can be used.

Exercise 6.5 The Trapezoidal method can be written

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})).$$

Determine if the method is explicit or implicit. Motivate your answer. Also show that the method is stable for $h\lambda < 0$. This means that the trapezoidal method is always stable.

Exercise 6.6 Heun's method computes $y_k \approx y(t_k)$ by the steps

$$\begin{aligned} k_1 &= hf(t_k, y_k), \\ k_2 &= hf(t_k + h, y_k + k_1), \\ y_{k+1} &= y_k + \tfrac{1}{2}(k_1 + k_2). \end{aligned}$$

Derive a condition that guarantees that Heun's method is stable when applied to the test problem $y' = \lambda y$, $y(0) = 1$.

Exercise 6.7 Rewrite the initial value problem $y'' + 3y' - ty + 1 = 0$, $y(0) = 1$, $y'(0) = 0$, as a system of first order.

Exercise 6.8 We have the second order equation

$$y'' = 2y(1 + y^2), \quad y(0) = 1,\ y'(0) = 2.$$

Rewrite the problem as a system of first order and compute an approximation of $y(0.4)$ using the Euler method and $h = 0.2$.

Exercise 6.9 Van der Pools equation is

$$y'' + e(y^2 - a)y' + y = 0, \quad y(0) = a,\ y'(0) = b.$$

Rewrite the equation as a system of first order equations.

1 Basic Concepts and Floating Point systems

Exercise 1.1 We first observe that $|\Delta a| \leq 0.5 \cdot 10^{-4}$ and $|\Delta b| \leq 0.4 \cdot 10^{-2} < 0.5 \cdot 10^{-2}$. Thus a has 4 correct decimals and b has 2. Further a has 3 significant digits and b has 4.

Exercise 1.2 We rewrite $\bar{c}_0$ as 2997924.58 Thus if the approximate value is roundned correctly it has two correct decimals. With 7 digits in the integer part the total number of significant digits is this 9.

Exercise 1.3 In order to obtain 5 significant digits we need 3 correct decimals. Thus $\bar{a} = 22.735$ and $|\Delta a| \leq 0.5 \cdot 10^{-3}$.

Exercise 1.4 The absolute error in the approximation is $|\pi - \bar{\pi}| \leq 9.3 \cdot 10^{-5} = 0.093 \cdot 10^{-3} < 0.5 \cdot 10^{-3}$. Thus the approximation has 3 correct decimals and 4 significant digits.

Exercise 1.5 The rounded value 1.61803 has five correct decimals because $|\phi - 1.61803| = 0.00000398875\ldots = 0.398875\ldots \times 10^{-5} \leq 0.5 \times 10^{-5}$. Further, the the rounded value 1.6180 has five significant digits.

Exercise 1.6 Note that $|\Delta a| \leq 4.7 \cdot 10^{-3} < 0.5 \cdot 10^{-2}$. Hence $\bar{a}$ has 2 correct decimals. Add the two digits before the decimal point and we have 4 significant digits.

Exercise 1.7 The rounding of π is an error in used data R_X, which is propagated to the result. The approximation of the exponential function is a truncation error R_T.

Exercise 1.8 The unit round-off for the number system is $\mu = 0.5 \cdot 10^{-3}$. Thus $|\Delta x| \leq \mu|x| \leq 0.5 \cdot 10^{-3} 102.232 < 0.052$.

Exercise 1.9 First write as a normalized number $x = 1.683556541 \cdot 10^2$. Then round to 5 correct digits in the fractional part and obtain $\bar{x} = 1.68356 \cdot 10^2$.

Exercise 1.10 The numbers are stored as follows

0	10000010	00101000...00

and

1	1000000	01010100...00

Exercise 1.11 We rewrite the numbers in normalized form and obtain $x = 1.175614 \cdot 10^2$ and $y = 1.678214 \cdot 10^{-2}$. If we round the fractional parts to 5 digits we find $x_r = 1.17561 \cdot 10^2$ and $y_r = 1.67821 \cdot 10^{-2}$.

Exercise 1.12 We say that m is the *mantissa*, β is the *base* (or radix) of the number system, and e is the exponent. To fully define a floating point number system we would also need to define the precision t to prescribe the number of digits in the mantissa $m = \pm d_0.d_1 d_2 \ldots d_t$. Also, the floating point system needs lower L and upper U bounds for the value of the exponent, i.e. $L \leq e \leq U$

Exercise 1.13 First do the calculations exactly

$$a \cdot b = 4.4625 \cdot 10^8 \text{ and } a/b = 1.6190476\ldots \cdot 10^3.$$

Then round the results to two digits in the fractional part to obtain

$$\mathrm{fl}[a \cdot b] = 4.46 \cdot 10^8 \text{ and } \mathrm{fl}[a/b] = 1.62\ldots \cdot 10^3.$$

In both cases a bound for the relative error is the unit round off for the floating point system, i.e. $\mu = 0.5 \cdot 10^{-2}$.

Exercise 1.14 For the first expression we rewrite

$$\sqrt{1+x} - 1 = \frac{(\sqrt{1+x}-1)(\sqrt{1+x}+1)}{\sqrt{1+x}+1} = \frac{x}{\sqrt{1+x}+1}.$$

For the second

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

which avoids the cancellation if x is small. For the last expression we use

$$1 - \cos^2(x) = \sin^2(x).$$

Exercise 1.15 Since there is no error in the number 10 and absolute errors are added during a minus operation the error in the computed value $\bar{x}$ is also at most $0.5 \cdot 10^{-5}$. A smaller error is achived by avoiding the cancellation, i.e. use $x = 1/(\sqrt{101}+10)$

2 Error Analysis

Exercise 2.1 First compute the approximate value $\bar{f} = 1.8$ ($|R_B| = 0$). Since x, y and z are correctly rounded we have error bounds $|\Delta x|, |\Delta y|, |\Delta z| \leq 0.5 \cdot 10^{-2}$. The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial x}||\Delta x| + |\frac{\partial f}{\partial y}||\Delta y| + |\frac{\partial f}{\partial z}||\Delta z| = |\frac{1}{z}||\Delta x| + |\frac{-1}{z}||\Delta y| + |-\frac{x-y}{z^2}||\Delta z| \leq 0.5 \cdot 10^{-2}.$$

Thus $f = 1.80 \pm 0.5 \cdot 10^{-2}$.

Exercise 2.2 The error propagation formula gives

$$|\Delta f(2)| \leq |\frac{\partial f(2)}{\partial b}||\Delta b| = |2\mathrm{e}^{2b}||\Delta b| = |2\mathrm{e}^{2 \cdot 1.70}|0.01 < 0.6.$$

The approximate value is $\bar{f(2)} = 29.9641\ldots \approx 30.0$, with $|R_B| \leq 0.5 \cdot 10^{-1}$. Add both errors to obtain $f(2) = 30.0 \pm 0.7$.

Exercise 2.3 The approximate area is $\bar{A} = \bar{\pi}\bar{r}^2 = 1773.77713920 \approx 1774$, $|R_B| \leq 0.3$. Note that the error in $\bar{\pi}$ is at most $0.5 \cdot 10^{-3}$. The error propagation formula gives

$$|\Delta A| \lesssim |\frac{\partial A}{\partial \pi}||\Delta \pi| + |\frac{\partial A}{\partial r}||\Delta r| = |r^2||\Delta \pi| + |2\pi r||\Delta r| \approx 3.2684 < 3.3.$$

The total error is $|R_{TOT}| \leq 3.3 + 0.3 < 4$. Thus $A = 1774 \pm 4$.

Exercise 2.4 The approximate value is $\bar{y} = 1.54\mathrm{e}^{3.17} = 36.7$, $|R_B| \leq 0.5 \cdot 10^{-1}$. The error propagation formula gives

$$|\Delta y| \lesssim |\frac{\partial y}{\partial a}||\Delta a| + |\frac{\partial y}{\partial b}||\Delta b| = |\mathrm{e}^b||\Delta a| + |a\mathrm{e}^b||\Delta b| < 2.55.$$

The total error is $|R_{TOT}| \leq 2.55 + 0.5 \cdot 10^{-1} < 2.6$. Thus $y = 36.7 \pm 2.6$.

In hindsight it would probably have been better to round to 37 and use $|R_B| = |37 - 36.6635\ldots| < 0.34$ to obtain the answer $y = 37 \pm 3$. Either works fine.

Exercise 2.5 Rewrite the expression to read

$$f(a, b) = \frac{ab}{a+b}.$$

The approximate value is $\bar{f} = 18.9$, $|R_B| \leq 0.05$, and the error propagation formula gives

$$|\Delta f| \lesssim |\frac{\partial f}{\partial a}||\Delta a| + |\frac{\partial f}{\partial b}||\Delta b| = |\frac{b^2}{(a+b)^2}||\Delta a| + |\frac{a^2}{(a+b)^2}||\Delta b| \leq 0.57.$$

The total error is $|R_{TOT}| \leq 0.05 + 0.57 < 0.7$. Thus $f = 18.9 \pm 0.7$.

Exercise 2.6 We first determine the computational order as

$$f(x) = \sqrt{1+x} - \sqrt{1-x} = \sqrt{a} - \sqrt{b} = c - d = e.$$

The relative errors in the intermediate results, e.g. $|\Delta a|/|a|$, are boudned by μ. The error propagation formula gives

$$|\Delta f| \lesssim |\frac{1}{2\sqrt{a}}||\Delta a| + |\frac{1}{2\sqrt{b}}||\Delta b| + |\Delta c| + |\Delta d| + |\Delta e|.$$

In order to simplify the result we use $a \approx b \approx c \approx d \approx 1$ for small x. Also

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

for small x. We obtain

$$|\Delta f| \lesssim \mu(\frac{1}{2} + \frac{1}{2} + 1 + 1 + |x|) \approx 3\mu.$$

Since $f(x) \approx x$ for small x the bound for the relative error is $|\Delta f|/|f| \leq 3|x|^{-1}\mu$.

Exercise 2.7 In the first case we have the computational order

$$f_1(x) = \frac{1 - \cos(x)}{\sin(x)} = \frac{1-c}{s} = \frac{d}{s} = e.$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial c}||\Delta c| + |\frac{\partial f}{\partial s}||\Delta s| + |\frac{\partial f}{\partial d}||\Delta d| + |\frac{\partial f}{\partial e}||\Delta e| = |\frac{1}{s}||\Delta c| + |\frac{1-c}{s^2}||\Delta s| + |\frac{1}{s}||\Delta d| + |\Delta e| \leq$$

$$\mu(|\frac{c}{s}| + |\frac{1-c}{s}| + |\frac{d}{s}| + |e|) \approx \frac{\mu}{x}$$

where we have used $c \approx 1$, $s \approx x$ and $d/s = e = f \approx x/2$. Similarly for the second expression we use the computational order

$$f_2(x) = \frac{\sin(x)}{1 + \cos(x)} = \frac{s}{1+c} = \frac{s}{d} = e.$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{s}{(1+c)^2}||\Delta c| + |\frac{1}{1+c}||\Delta s| + |\frac{s}{d^2}||\Delta d| + |\Delta e| \leq \mu(|\frac{cs}{(1+c)^2}| + |\frac{s}{1+c}| + |\frac{s}{d}| + |e|) \approx 1.75x\mu.$$

These are the absolute errors. Insert $x = 1.111 \cdot 10^{-8}$ in and $|\Delta f_1| \leq 10^{-8}$ which is on the same order of magnitude as the actual difference between f_1 and f_2.

Exercise 2.8 The computational order is

$$f(x) = \mathrm{e}^x - 3x = a - 3x = a - b = c$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial a}||\Delta a| + |\frac{\partial f}{\partial b}||\Delta b| + |\frac{\partial f}{\partial c}||\Delta c| = |1||\Delta a| + |1||\Delta b| + |1||\Delta c| \lesssim$$

$$\mu(|a| + |b| + |c|) \approx \mu(|1| + |3x| + |1|) \approx 2\mu,$$

where we have used $e^x \approx 1$, $f(x) = c \approx 1$ since x is small. There is no cancellation present in these calculations. Everything turns out fine and both the absolute and relative errors are bounded by 2μ (since the function value $f(x) \approx 1$).

Exercise 2.9 The computational order and the intermediate results are

$$f = \frac{x - \sin(x)}{x^3} = \frac{x - a}{b} = \frac{c}{b} = d.$$

where the relative error in all the intermediate results are boudned by μ. The error propagation formula gives

$$|\Delta f| \lesssim |\frac{\partial f}{\partial a}||\Delta a| + |\frac{\partial f}{\partial b}||\Delta b| + |\frac{\partial f}{\partial c}||\Delta c| + |\frac{\partial f}{\partial d}||\Delta d| =$$

$$| - \frac{1}{b}||\Delta a| + | - \frac{c}{b^2}||\Delta b| + |\frac{1}{b}||\Delta c| + |1||\Delta d| \leq \mu(|\frac{a}{b}| + |\frac{c}{b}||\frac{c}{b}| + |d|) \approx \mu(|\frac{1}{x^2}| + \frac{3}{6}).$$

If we insert $x = 10^{-7}$ and $\mu \approx 1.1 \cdot 10^{-16}$ we obtain $\|\Delta f| \approx 0.0110$. Thus the actual error $5.4 \cdot 10^{-3}$ is within the error bound.

Exercise 2.10 The computational order is

$$f(x) = 1 - 2x\cos(x) = 1 - 2xa + 1 - b = c.$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial a}||\Delta a| + |\frac{\partial f}{\partial b}||\Delta b| + |\frac{\partial f}{\partial c}||\Delta c| = |2x||\Delta a| + |1||\Delta b| + |1||\Delta c| \lesssim$$

$$\mu(|2xa| + |b| + |c|) \approx \mu(|2x| + |2x| + 1) \approx \mu,$$

where we have used $\cos(x) \approx 1$, $f(x) = c \approx 1$ and that x is small. There is no cancellation present in these calculations. Everything turns out fine and both the absolute and relative errors are bounded by μ (since the function value $f(x) \approx 1$).

Exercise 2.11 We do a Taylor series expansion of $f(\bar{x})$ around x to obtain

$$f(\bar{x}) = f(x + \Delta x) = f(x) + f'(x)\Delta x + f''(\eta)\frac{(\Delta x)^2}{2}.$$

where $\eta \in (x, \bar{x})$ and $f'(x) = 0$. If $\Delta x|$ is small then $\eta \approx \bar{x}$ and we obtain

$$|\Delta f| \lesssim |f''(\bar{x})|\frac{|\Delta x|^2}{2}.$$

3 Non-linear equations

Exercise 3.1 Since the function is continuous and $f(0) = -1$ and $f(1) = 1.3849$ there has to be a root in the interval $[0, 1]$.

Exercise 3.2 $\phi'(x) = \frac{1}{2}e^x$ and $|\phi'(0.3)| \approx 0.68 < 1$.

Exercise 3.3 With $\bar{x} = 1.7$ we get $f(\bar{x}) = -0.387$. Also $f'(x) = 3x^2 + 1$ so $f'(\xi) \approx f'(1.7) = 9.67$. Thus

$$|1.7 - x^*| \leq \frac{|f(1.7)|}{|f'(1.7)|} < \frac{0.39}{9.6} < 0.0406 < 0.5 \cdot 10^{-1}.$$

Exercise 3.4 Since $f(\bar{x})$ is evaluated as zero we need to include the computational errors and we actually have $\bar{f}(\bar{x}) = 0$ and $|\bar{f}(\bar{x}) - f(\bar{x})| \leq 10^{-8}$. The error estimate is

$$|\bar{x} - x^*| \leq \frac{|f(\bar{x})|}{|f'(\xi)|} \leq \frac{|\bar{f}(\bar{x})| + |\bar{f}(\bar{x}) - f(\bar{x})|}{|f'(\xi)|} \leq \frac{0 + 10^{-8}}{1.7} < 6 \cdot 10^{-9}.$$

Exercise 3.5 Let $f(x) = 5(1 - e^{-x}) - x$ and find a root of the equation $f(x) = 0$. We use the Newton-Raphson method. If $x_0 = 5$ we get $x_2 = \bar{x} = 4.96511$. The error estimate is

$$|\bar{x} - x^*| \leq \frac{|f(\bar{x})|}{|f'(\bar{x})|} \leq \frac{4.1 \cdot 10^{-6}}{0.96} \leq 0.5 \cdot 10^{-5}.$$

Thus $x^* = 4.96511 \pm 0.5 \cdot 10^{-5}$ has five correct decimals.

Exercise 3.6 The rate of convergence is determined by the derivative $|\phi'(x^*)|$. For the two methods, and $x^* \approx 1.9$, we get $\phi_1'(1.9) \approx 0.647$ and $\phi_2'(1.9) \approx 0.177$. Thus the second method has the fastest convergence. Also note that a fixed point satisfies $x^* = \frac{x^*}{2} + \sin(x^*)$ which can be written as $x^* = \sin(x^*)$. Thus a fixed point satisfies the original equation and we really have convergence to a root.

Exercise 3.7 Let x^* be the fixed point. Then

$$|x_n - x^*| = |\phi(x_{n-1}) - \phi(x^*)| \leq |\phi'(\xi_n)||x_{n-1} - x^*| \leq C|x_{n-1} - x^*|,$$

where the constant C satisfies $|\phi'(\xi_n)| \leq C$. In this case $\phi(x) = \frac{1}{2}\cos(x)$ and $\phi'(x) = -\frac{1}{2}\sin(x)$. Thus $|\phi'(x)| \leq \frac{1}{2}$. Thus we can pick $C = \frac{1}{2} < 1$ and we have proved that the error $|x_n - x^*|$ is reduced by a factor of 2 in each step. Thus the method is convergent regardless of x_0.

In order to obtain the error estimate, for $\bar{x} = 0.4501832$, we comute the derivative $f'(x) = -2\sin(x) - 4$ and use the estimate $= |-4.87026\ldots| > 4.8 = M$. Thus

$$|\bar{x} - x^*| \leq \frac{|f(\bar{x})|}{|f'(\bar{x})|} \leq \frac{2.01 \cdot 10^{-6}}{4.87} < 0.42 \cdot 10^{-6}.$$

We conclude that $x^* = 0.4501832 \pm 0.5 \cdot 10^{-6}$.

Exercise 3.8 First the error in the approximation $\bar{x} = 1.05$ is estimated by

$$|\bar{x} - x^*| \leq \frac{|f(\bar{x})|}{|f'(\bar{x}|} \approx \frac{|f(1.05)|}{|f'(1.05|} < \frac{1.9 \cdot 10^{-4}}{2.0} < 10^{-4},$$

where $f'(x) = 1 + 3\mathrm{e}^{-x}$.

Now let x^* be a fixed point to the iterations. In the case *(i)* the fixed point obviously satisfies the equation $f(x) = 0$. In the case *(ii)* we get $3x^* = 2x^* + 3\mathrm{e}^{-x^*}$ or $x^* = 3\mathrm{e}^{-x^*}$ or $f(x^*) = 0$. A similar resoning holds in the case *(iv)*. However, a fixed point to *(iii)* does not satisfy the equation.

Now we consider the convergence speed of the iterations *(i)*, *(ii)* and *(iv)*. In all cases we compute the derivative of the iteration function and obtain

$$\phi_1(x) = 3\mathrm{e}^{-x}, \text{ so } \phi_1'(1.05) = 1.05 \text{ and divergence,}$$

$$\phi_2(x) = (2x + 3\mathrm{e}^{-x})/3, \text{ so } \phi_2'(1.05) = -0.32 \text{ and convergence,}$$

and

$$\phi_4(x) = (x + 3\mathrm{e}^{-x})/2, \text{ so } \phi_4'(1.05) = -0.025, \text{ and again convergence,}$$

Thus the method *(iv)* should converge the fastest to x^*.

Finally the error in step k is $|x_k - x^*| \approx |\phi_4'(1.05)|^k |1.05 - x^*| \approx (0.025)^k 10^{-4}$. Testing shows that $k = 4$ gives an error of $3.9 \cdot 10^{-11}$ and thus 4 iterations is enough.

Exercise 3.9 If we apply Newton-Raphson to the equation $x^2 = 0$ we get

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2}{2x_k} = \frac{1}{2}x_k,$$

which means that if $x_0 = 1$ then $x_k = 2^{-k} \to 0 = x^*$ as $k \to \infty$. The rate of convergence is verified to be linear by the observation that $e_k = |x_k - x^*| = |\frac{1}{2}x_{k-1} - 0| = \frac{1}{2}|x_{k-1} - 0| = \frac{1}{2}e_{k-1}$.

Exercise 3.10 Newton-Raphsons method is defined by the iteration function

$$\phi(x) = x - \frac{f(x)}{f'(x)}, \text{ and } \phi'(x) = -\frac{f(x)f''(x)}{(f'(x))^2}.$$

Since x^* is a single root, i.e. $f'(x^*) \neq 0$, we see that $\phi'(x^*) = 0$. A Taylor series expansion shows that

$$\phi(x_k) = \phi(x^*) + \phi'(x^*)(x_k - x^*) + \frac{\phi''(\xi)}{2}(x_k - x^*)^2, \xi \in (x_k, x^*).$$

Since $\phi(x_k) = x_{k+1}$, $\phi(x^*) = x^*$ and $\phi'(x^*) = 0$ we obtain

$$x_{k+1} - x^* = \frac{\phi''(\xi)}{2}(x_k - x^*)^2,$$

which shows that the convergence is quadratic.

Exercise 3.11 We use a Taylor series expansion

$$f(\bar{x}) = f(x^*) + f'(x^*)(\bar{x} - x^*) + \frac{f''(\xi)}{2}(\bar{x} - x^*)^2, \xi \in (\bar{x}, x^*),$$

and since x^* is a double root we have $f(x^*) = f'(x^*) = 0$ and thus

$$f(\bar{x}) = \frac{f''(\xi)}{2}(\bar{x} - x^*)^2.$$

Taking absolute values and approximating $\xi \approx \bar{x}$ gives the desired estimate. In the practical case when $\bar{x} = 1.99789$ we find that $f(\bar{x}) = -6.6875 \cdot 10^{-6}$ and $f''(\bar{x}) = -3.0127$. Thus

$$|\bar{x} - x^*| \leq \sqrt{\frac{2|f(\bar{x})|}{|f''(\bar{x})|}} \leq \sqrt{\frac{2 \cdot 6.7 \cdot 10^{-6}}{3.0}} < 2.2 \cdot 10^{-3}.$$

Note that $x^* = 2$ so the actual error is of the same magnitude as the error estimate in this case.

Exercise 3.12 First we use the error estimate

$$|\bar{x} - x^*| \lesssim \frac{|f(\bar{x})|}{|f'(\bar{x})|} \approx \frac{1.2971 \cdot 10^{-4}}{3.0433} < 4.3 \cdot 10^{-5}.$$

Secondly, we define the order of convergence as the largest integer p such that

$$\lim_{k \to \infty} \frac{|x_k - x^*|}{|x_{k-1} - x^*|^p} = C < \infty.$$

Since x^* is unknown we cannot directly apply the definition. The simplest solution is to assume that the iteraton x_4 has a much smaller error than the other iterations x_1, x_2, x_3. Thus we approximate $x^* = 0.5177574$ and compute the errors $|x_0 - x^*| \approx 4.8 \cdot 10^{-1}$, $|x_1 - x^*| \approx 1.4 \cdot 10^{-1}$, $|x_2 - x^*| \approx 1.4 \cdot 10^{-2}$, and $|x_3 - x^*| \approx 1.5 \cdot 10^{-4}$. Since $(|x_1 - x^*|)^2 \approx (1.4 \cdot 10^{-1})^2 \approx 2 \cdot 10^{-2} \approx |x_2 - x^*|$ and $(|x_2 - x^*|)^2 \approx (1.4 \cdot 10^{-2})^2 \approx |x_3 - x^*|$ we conclude that the table shows that $p = 2$ for Newton-Raphsons method.

Exercise 3.13 The order of convergence is the largest integer p such that

$$\lim_{k \to \infty} \frac{|x_k - x^*|}{|x_{k-1} - x^*|^p} = C < \infty.$$

This means that $|x_k - x^*| \approx C|x_{k-1} - x^*|^p$, for a constant C. In the table we see that the error in step $k+1$ is always about half of of the error at step k. This fits very nicely with $p = 1$ and $C = 0.5$. Thus we have *linear convergence* for this specific polynomial. It is known that Newton-Raphsons method has order of convergence $p = 2$ for single roots and $p = 1$ for double roots and this $x^* = 2$ has to be a *double root*.

Exercise 3.14 First if we apply the Newton-Raphson method to $f(x) = x^2 - a = 0$ we obtain

$$x_{k+1} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{1}{2}(x_k - a/x_k) = \phi(x_k).$$

Since a is a normalized floating point number we can write $a = (-1)^2(1.f)_2 2^k$. We only need to consider positive numbers and if k is even then $\sqrt{a} = \sqrt{(1.f)_2}2^{k/2}$ and if k is odd we instead have $\sqrt{a} = \sqrt{(1.f)_2 2^1}2^{(k-1)/2}$. Thus in worst case we need to compute the square root of a number $1 \leq (1.f)_2 2^1 < 4$. So it is enough to consider the case $1 \leq a < 4$ and $1 \leq x < 2$.

For the convergence analysis we compute $\phi''(x) = -\frac{a}{x^3}$ and therefore $\phi''(\sqrt{a}) = 1/\sqrt{a} \leq 1$. Thus

$$|x_k - \sqrt{a}| \lesssim \frac{1}{2}|x_{k-1} - \sqrt{a}|^2.$$

If $x_0 = 1.5$ the maximum error is $|x_0 - \sqrt{a}| \leq 0.5$. We get $|x_1 - \sqrt{a}| \leq 0.125$, $|x_2 - \sqrt{a}| \leq 0.078$, $|x_3 - \sqrt{a}| \leq 3.1 \cdot 10^{-5}$, $|x_4 - \sqrt{a}| \leq 4.7 \cdot 10^{-10}$, and $|x_5 - \sqrt{a}| \leq 1.1 \cdot 10^{-19}$. We see that 5 iterations are needed if $x_0 = 1.5$.

In order to decrease the number of iterations by one we need an initial guess with an error $|x_0 - \sqrt{a}| < 0.125$. Then the new x_0 is the same as x_1 above. This means dividing the interval $[1, 4]$ into n intervals $[a_0, a_1)$, $[a_1, a_2)$,...,$[a_{n-2}, a_{n-1})$. For a certain starting value a we indentify the index k such that $a_k \leq a < a_{k+1}$. The initial guess x_0 is then given by the middle point $(\sqrt{a_k} + \sqrt{a_{k+1}})/2$ and the initial error is $(\sqrt{a_{+1}} - \sqrt{a_k})/2$. If $n = 4$ then the largest initial error is given by $(\sqrt{1 + 3/n} - \sqrt{1})/2 = 0.1614$. If we continue as above we get $|x_4 - \sqrt{a}| \leq 6.5 \cdot 10^{-18}$ which is below machine precision.

4 Polynomial and Spline Interpolation

Exercise 4.1 A polynomial, $p_n(x)$, of degree n, can be written $p_n(x) = c_0 + c_1 x + \ldots + c_n x^n$. Thus there are $n+1$ parameters to determine and the polynomial can satisfy exactly $n+1$ interpolation conditions.

Exercise 4.2 Use the Ansatz $p(x) = c_0 + c_1(x - 0.9) + c_2(x - 0.9)(x - 1.1)$, where the last term is used to estimate the truncation error. The interpolation conditions give

$$p(0.9) = c_0 = f(0.9) = 0.4710, \text{ and } p(1.1) = c_0 + c_1(1.1 - 0.9) = f(1.1) = 0.2452 \text{ so } c_1 = -1.1290.$$

Thus the linear polynomial is $p(x) = 0.4710 - 1.1290(x - 0.9)$. The truncation error is obtained using

$$p(1.2) = c_0 + c_1(1.2 - 0.9) + c_2(1.2 - 0.9)(1.2 - 1.1) = f(1.2) = 0.2385 \text{ so } c_2 = 3.5400.$$

The truncation error is $R_T \approx 3.45(x - 0.9)(x - 1.1)$. Insert $x = 1.03$ to find $f(1.03) \approx p(1.03) = 0.3242$, $|R_B| \leq 0.5 \cdot 10^{-4}$, and $|R_T(1.03)| \leq 0.033$. The errors in the table also gives an error $|R_{XF}| \leq 0.5 \cdot 10^{-4}$ and we obtain $f(1.03) = 0.3242 \pm 0.034$. Its resonable to round off a bit more to get $f(1.03) = 0.324 \pm 0.04$.

Exercise 4.3 We require that the total error is $R_{TOT} \leq 10^{-5}$. If the values in the table is stored with 6 correct digits then the resulting error is $R_{XF} \leq 0.5 \cdot 10^{-6}$. Thus the truncation error can be at most $R_T \leq 9.5 \cdot 10^{-6}$. For linear interpolation the truncation error is given by

$$R_T \leq \frac{h^2}{8} \max_{x_i \leq \xi \leq x_{i+1}} |f''(\xi)|,$$

and since $f(x) = \log(x)$ we find that $|f''(x)| = |-x^{-2}| \leq 1$ since $1 \leq x \leq 4$. Thus $R_T \leq 9.5 \cdot 10^{-6}$ if $h^2 \leq 8 \cdot 9.5 \cdot 10^{-6}$ which gives $h \leq 0.0087$. Since the interval length is $x_n - x_1 = 4 - 1 = 3$ the required table size is $n = 3/h \approx 3/0.0087 \approx 344.82 < 345$. Thus we need at least 345 function values in our table.

Exercise 4.4 The error estimate for linear interpolation is

$$f(x) - p(x) = \frac{1}{2} f''(\eta(x))(x - x_1)(x - x_2),$$

where, for our case, $x_1 = 0$, $x_2 = 1$, and $0 < \eta(x) < 1$. Since $f(x) = \sin(x)$ we obtain $|f''(\eta)| = |-\sin(\eta)| \leq 1$. We also see that $|x(x-1)| \leq \frac{1}{4}$ since the maximum occurs for $x = \frac{1}{2}$. Thus

$$|f(x) - p(x)| = \frac{1}{2}|f''(\eta(x))(x - 0)(x - 1)| \leq \frac{1}{2} 1 \frac{1}{4} = \frac{1}{8}.$$

Exercise 4.5 Use the Ansatz $p(x) = c_0 + c_1(x - 0.0) + c_2(x - 0.0)(x - 0.5)$, where the last term is used to estimate the truncation error. The interpolation conditions $p(0.0) = 1.80$ and $p(0.5) = 2.80$ gives us the linear interpolating polynomial $p_1(x) = 1.8 + 2x$. The final interpolation condition $p(1.0) = 4.10$ gives us truncation error $R_T \approx 0.6(x - 0.0)(x - 0.5)$. Insert $x = 0.4$ to obtain $f(1.03) \approx p_1(0.4) = 2.6$, where $|R_B| = 0$, and $|R_T(0.4)| \leq 0.024$. The errors in the table also gives an error $|R_{XF}| \leq 0.5 \cdot 10^{-2}$ and we obtain $f(0.4) = 2.6 \pm 0.03$.

Exercise 4.6 Use the Ansatz $p(x) = c_0 + c_1(x-0.7) + c_2(x-0.7)(x-0.8) + c_3(x-0.7)(x-0.8)(x-0.6)$, where the last term is used to estimate the truncation error. The interpolation conditions $p(0.7) = 1.29$, $p(0.8) = 1.32$ and $p(0.6) = 1.23$ gives the coefficents of the quadratic polynomial $c = (1.29\,,\, 0.3\,,\, -1.5)^T$. The truncation error is obtained from the interpolation condition $p(0.9) = 1.07$. We find that $c_3 \approx -41.7$. We find that $p_2(0.74) = 1.306$, with $|R_B| \leq 0.5 \cdot 10^{-3}$ (not asked for in the exercise), and with the truncation error $|R_T| \leq -0.014$.

Exercise 4.7 The basis function satisfies $\ell_2(x_2) = 1$ and $\ell_2(x_i) = 0$, $i \neq 2$. Thus

$$\ell_2(x) = \frac{(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)}.$$

The degree of $\ell_2(x)$ is $n = 3$.

Exercise 4.8 The polynomial is

$$p(x) = 1.3\frac{(x-2)(x-3)}{(1-2)(1-3)} + 0.6\frac{(x-1)(x-3)}{(2-1)(2-3)} + 1.9\frac{(x-1)(x-2)}{(3-1)(3-2)}.$$

There is no reason to simplify the expression further.

Exercise 4.9 First $p(0) = c_0 = 0$ and $p(1) = c_0 + c_1 + c_2 + c_3 = 0$ gives two equations. Then $p'(x) = c_1 + 2c_2x + 3c_3x^2$ so we also obtain $p'(0) = c_1 = 1$ and $p'(1) = c_1 + 2c_2 + 3c_3 = 1$. Thus the system of equations is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

We can solve the linear system by noting that $c_0 = 0$ and $c_1 = 1$. Then we are left with two equations for c_2 and c_3. The solution is $p(x) = x - 3x^2 + 2x^3$.

Exercise 4.10 The conditions for $s(x)$ to be a cubic spline are (i) on each sub interval $[x_i, x_{i+1}]$ the spline $s(x)$ should be given by a cubic polynomial, and (ii) $s(x)$, $s'(x)$ and $s''(x)$ should be continuous on the whole interval $[x_1, x_n]$. Also the (iii) the interpolation conditions $s(x_i) = f(x_i)$ needs to be satisfied. The given information is not sufficient since we also need two end point conditions for the spline to be unique.

Exercise 4.11 The function $s(x)$ is a cubic spline since $s(x)$, $s'(x)$ and $s''(x)$ are continuous at $x = 1$.

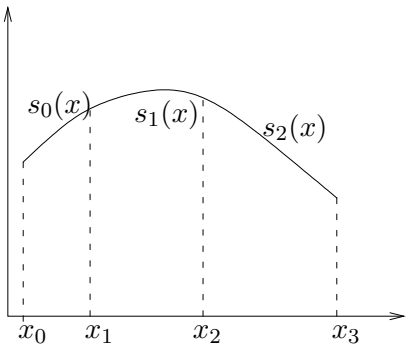
Exercise 4.12 The conditions that has to be satisfied are $s(1) = a + 1 = b + c$, $s'(1) = a = 3b + 2c$ and $s''(1) = 0 = 6b + 2c$. The solution is $a = -3$, $b = 1$ and $c = -3$.

Exercise 4.13 Since $f(x)$ is a cubic polynomial, and correct end point conditions are to be used, then $s(x) = f(x)$. This means the spline $s(x)$ is the same cubic polynomial in each of the sub intervals.

Exercise 4.14 We have more than enough information to determine the parameters. The most efficient way is to first compute $s_2(1) = 2$ and determine a using $s_1(1) = 1.6 + a = 2$, i.e. $a = 0.4$. Now compute $s_1'(1) = 2.5$ and use $s_2'(1) = b = 2.5$. This leaves c which can be computed using $s_2(2) = 4.9 + c = 6.7$ or $c = 1.8$.

Exercise 4.15 Check the continuity requirements by $s_1(1) = 0.6 = s_2(1)$, $s_1'(1) = -0.6 = s_2'(1)$ and finally $s_1''(1) = -0.8 = s_2''(1)$. Thus $s(x)$ is a cubic spline. We also compute $s_1''(0) = -0.8$ and conclude that $s(x)$ is *not* a natural cubic spline.

Exercise 4.16 For $s(x)$ to be a cubic spline we require that $s_{k-1}(x_k) = s_k(x_k)$, $s_{k-1}'(x_k) = s_k'(x_k)$ and $s_{k-1}''(x_k) = s_k''(x_k)$, for $k = 1, 2, \ldots, N-1$. In addition we need the interpolation conditions $s_k(x_k) = f(x_k)$ and $s_k(x_{k+1}) = f(x_{k+1})$, for $k = 0, 1, \ldots, N-1$. A sketch of the case $N = 3$ is given below



Exercise 4.17 Since the error satisfies $E(N) \approx Ch^p = C'(1/N)^p$, where $C' = C(b-a)^p$ is a constant. Thus $E(N)/E(2N) = 2^p$ and if we insert the values from the table we get

$$\frac{E(5)}{E(10)} = \frac{0.0435}{0.00269} \approx 16.2 \text{ and } \frac{E(10)}{E(20)} = \frac{0.00269}{0.000172} \approx 15.6.$$

In both cases the quitient is sufficiently close to $2^4 = 16$ to conclude that $p = 4$.

Exercise 4.18 The function $s(x)$ is a linear spline if it is given by a first degree polynomial on each subinterval. In our case the intervals are $1 < x < 3$ and $3 < x < 4$. Thus we are seeking two linear polynomials $s_1(x)$ and $s_2(x)$. We can write $s_1(x) = 1.56 + c_1(x-1)$ and use $s_1(3) = 2.31$ to obtain $c_1 = 0.375$. Similarily we obtain $s_2(x) = 2.31 - 0.34(x-3)$. The linear spline is thus given by

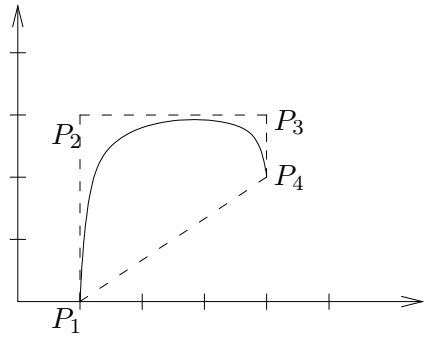
$$s(x) = \begin{cases} 1.56 + 0.375(x-1), & 1 \le x < 3, \\ 2.31 - 0.34(x-3), & 3 \le x < 4. \end{cases}$$

Exercise 4.19 For **a)** differentiate the expression for $p(t)$ to obtain the tangent

$$p'(t) = -2(1-t)P_1 + 2(1-2t)P_2 + 2tP_3, \text{ and } p'(0) = 2(P_2 - P_1).$$

For **b)** we first note that P_4 and P_5 should have the same x-coordinate so $P_4 = (6, , \alpha)^T$. The tangent direction at P_3 should be paralell to $P_3 - P_2 = (3\,,\,5)^T - (2\,,\,6)^T = (1\,,\,-1)^T$. Compute $P_4 - P_3 = (6\,,\,\alpha)^T - (3\,,\,5)^T = (3, \alpha - 5)^T = 3(1\,,\,-1)^T$ if $\alpha = 2$. Thus we have to use $P_4 = (6\,,\,2)^T$.

Exercise 4.20 The sketch is



The convex hull is the area enclosed by the dashed lines. Important features of the Beziér curve is that since both P_1/P_2 and P_3/P_4 have the same x-coordinate the tangent direction of the curve is vertical at both the starting and ending points.

Exercise 4.21 For **a)** we differentiate

$$p'(t) = -3(1-t)^2P_1+3(-2(1-t)t+(1-t)^2)P_2+3(2(1-t)t-t^2)P_3+3t^2P_4, \text{ and } p'(0) = 3(P_2-P_1).$$

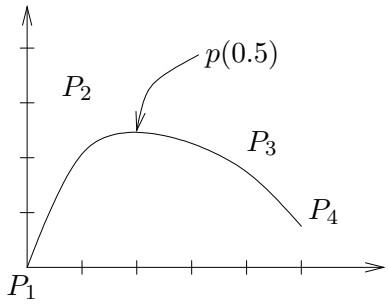
For **b)** the definition of the *convex hull* is the set of all convex linear combinations $\alpha_1P_1 + \alpha_2P_2 + \alpha_3P_3 + \alpha_4P_4$, where $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$ and $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0$. To show that the Beziér curve is located within the convex hull we note that the weights in the expression for $p(t)$ are calculated from the identity

$$1 = 1^3 = (1 - t + t)^3 = (1 - t)^3 + 3(1 - t)^2t + 3(1 - t)t^2 + t^3,$$

and that all terms are positive for $0 \leq t \leq 1$. For **c)** we compute $p(1/2)$ by inserting $t = 1/2$ into the expression (note that $t = 0.5$ means also $1 - t = 0.5$)

$$p(1/2) = (\frac{1}{2})^3 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + 3 \begin{pmatrix} 1 \\ 3 \end{pmatrix} + 3 \begin{pmatrix} 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 5 \\ 1 \end{pmatrix} \right) = \begin{pmatrix} 2.5 \\ 2.0 \end{pmatrix}.$$

The sketch is



Exercise 4.22 For the slope to be -2 the tangent vector at P_4 should be in the direction $(1\,,\,-2)^T$. Thus we can pick

$$P_3 = \begin{pmatrix} 2 \\ 3 \end{pmatrix} - \alpha \begin{pmatrix} 1 \\ -2 \end{pmatrix} \text{ and } P_5 = \begin{pmatrix} 2 \\ 3 \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ -2 \end{pmatrix},$$

where α is a positive number. There is no unique solution to this problem.

5 Integration, Differentiation and Extrapolation

Exercise 5.1 Taylors formula gives

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{6}f^{(3)}(x)h^3 + \dots$$

Thus

$$\frac{1}{h}(f(x+h) - f(x)) = f'(x) + \frac{1}{2}f''(x)h + \frac{1}{6}f^{(3)}(x)h^2 + \dots = f'(x) + Ch + \mathcal{O}(h^2), \quad C = \frac{1}{2}f''(x).$$

Exercise 5.2 First we compute two different approximations using the Trapezoidal method. We have

$$T(0.5) = 0.5\left(\frac{1.000}{2} + 1.006 + \frac{0.670}{2}\right) = 0.9205,$$

and

$$T(0.25) = 0.25\left(\frac{1.000}{2} + 1.015 + 1.006 + 0.882 + \frac{0.670}{2}\right) = 0.9345.$$

Since the Trapezoidal method has a truncation error of the form $R_T = ch^2$ we can estimate the truncation error by

$$R_T(0.25) = \frac{T(0.5) - T(0.25)}{3} = -4.667 \cdot 10^{-3}.$$

Also since the table is given with absolute error $|\Delta f| < \cdot 0.5 \cdot 10^{-3}$ we get an error in the integral that can be estimated $|R_{XF}| \leq (1-0)0.5 \cdot 10^{-3} = 0.5 \cdot 10^{-3}$. If we round the result to 3 correct digits we get $I = 0.934 \pm (0.5 \cdot 10^{-3} + 4.7 \cdot 10^{-3} + 0.5 \cdot 10^{-3}) = 0.934 \pm 6 \cdot 10^{-3}$.

Exercise 5.3 Since $|\Delta f(x)|/|f(x)| \leq \varepsilon$ we obtain the error estimate

$$|R_{XF}| = |\int_a^b f(x)dx - \int_a^b \bar{f}(x)dx| \leq \int_a^b |f(x) - \bar{f}(x)|dx \leq \int_a^b \varepsilon |f(x)|dx \leq \left(\int_a^b |f(x)|dx\right)\varepsilon.$$

This means that the bound is good for the case when $f(x)$ is either positive or negative. If $f(x)$ changes sign in the interval (a, b) then it may happen that

$$\left|\int_a^b f(x)dx\right| << \int_a^b |f(x)|dx,$$

and the result may have a large relative error.

Exercise 5.4 We denote the error by $\epsilon_h \approx Ch^p$. Then

$$\frac{\epsilon_{h_1}}{\epsilon_{h_2}} \approx \frac{Ch_1^p}{Ch_2^p} = (\frac{h_1}{h_2})^p.$$

Insert numbers from the table we obtain

$$2^p = (\frac{0.2}{0.1})^p \approx \frac{\epsilon_{0.2}}{\epsilon_{0.1}} = \frac{0.342}{0.0861} \approx 3.97 \text{ and } 2^p \approx \frac{\epsilon_{0.1}}{\epsilon_{0.05}} = \frac{0.0861}{0.0209} \approx 4.11.$$

We see that $2^p = 4$ which means $p = 2$.

Exercise 5.5 We see that

$$F_1(h) + \frac{F_1(h) - F_1(qh)}{q^{p_1} - 1} = a + bh^{p_1} + ch^{p_2} + \frac{a + bh^{p_1} + ch^{p_2} - (a + bq^{p_1}h^{p_1} + cq^{p_2}h^{p_2})}{q^{p_1} - 1} =$$

$$= a + bh^{p_1} + ch^{p_2} + \frac{b(1 - q^{p_1})h^{p_1} + c(1 - q^{p_2})h^{p_2}}{q^{p_1} - 1} = a + c\left(1 - \frac{1 - q^{p_2}}{1 - q^{p_1}}\right)h^{p_2} = a + \mathcal{O}(h^{p_2})$$

Exercise 5.6 Let $h = 0.05$. From the table we see that

$$T(h) = I + R_T \approx I + Ch^4, \text{ and } T(2h) \approx I + C(2h)^4 = I + 16Ch^4 \approx I + 16R_T$$

where I is the exact value of the integral and R_T is the truncation error for $h = 0.05$. Thus

$$R_T \approx Ch^4 \approx \frac{T(2h) - T(h)}{15} = 7.83 \cdot 10^{-6}.$$

Exercise 5.7 With $T(h) = T_0 + Ch^p$ we form the quotient

$$\frac{T(h) - T(h/2)}{T(h/2) - T(h/4)} = \frac{h^p(1 - 2^{-p})}{(h/2)^p(1 - 2^{-p})} = 2^p.$$

Insert the values for $h = 0.4$ and $h = 0.2$ to obtain $2^p \approx 4.11$ and $2^p \approx 3.73$. We conclude that $p = 2$.

Exercise 5.8 First we compute

$$T(1/2) = 0.5\left(\frac{1.5000}{2} + 1.0713 + \frac{0.9073}{2}\right) = 1.137475,$$

and

$$T(1/4) = 0.25\left(\frac{1.5000}{2} + 1.2412 + 1.0713 + 0.9663 + \frac{0.9073}{2}\right) = 1.1206125.$$

The truncation error in the result for $h = 1/4$ is estimated by $|R_T| \leq |T(1/2) - T(1/4)|/3 < 5.7 \cdot 10^{-3}$ since the truncation error for the Trapezoidal method has the form $R_T \approx ch^2$. Since the function values are correctly rounded we also have $|R_{XF}| \leq (b - a)\varepsilon = (1 - 0)0.5 \cdot 10^{-4}$. This means that the integral can be estimated by $I = 1.121 \pm (0.5 \cdot 10^{-4} + 5.7 \cdot 10^{-3} + 0.5 \cdot 10^{-3}) = 1.121 \pm 7 \cdot 10^{-3}$.

Exercise 5.9 We find that

$$\frac{R_T(h)}{R_T(h/2)} = 2^p.$$

with $h = 0.1$ we get $2^p \approx 1.98$ and with $h = 0.05$ we get $2^p \approx 1.98$. Thus $p = 1$ and for $h = 0.025$ we get $R_T = -0.0099 = C(0.025)^1$ which means $C \approx -0.4$. The truncation error is thus given by $R_T \approx -0.4h^1$ and $R_T < 10^-8$ if $h < 10^-8/0.4 = 2.5 \cdot 10^{-8}$.

Exercise 5.10 No since the Trapezoidal method needs, at least a continuous second derivative $f''(x)$ on the intervall $[1, 4]$ to have a truncation error $R_T \approx ch^2$. In this case the first derivative is

$$f'(x) = -2x\sin(x^2)\sqrt{4 - x} - \cos(x^2)(4 - x)^{-1/2},$$

which tends to $-\infty$ as $x \to 4$. This means that $f(x)$ does not have the nessecary smoothness for the Trapezoidal method to work well.

Exercise 5.11 For **a)** we compute

$$\frac{S(4h) - S(2h)}{S(2h) - S(h)} \approx \frac{c4^p h^p - c2^p h^p}{c2^p h^p - h^p} = 2^p.$$

From the table, with $h = 1/8$, we obtain

$$\frac{S(1/2) - S(1/4)}{S(1/4) - S(1/8)} \approx \frac{2.06246 - 2.03208}{2.03208 - 2.03021} = 16.25 \approx 2^4.$$

We thus conclude that $p = 4$. We can then use

$$S(1/4) - S(1/8) = 2.03208 - 2.03021 \approx c(2^4 - 1)(1/8)^4,$$

to obtain $c = 0.51$. For **b)** we note that $|R_T| \approx 0.51h^4$ is smaller than 10^{-6} if $h \leq (10^{-6}/0.51)^{1/4} \approx 0.03742$. Thus a step size $h < 0.037$ is required.

6 Differential Equations

Exercise 6.1 With $y(0) = y_0 = 1$, and $h = 0.5$ we obtain

$$y(0.5) \approx y_1 = y_0 + hf(t_0, y_0) = y_0 + h(y_0^2 - t_0) = 1 + 0.5(1^2 - 0) = 1.5.$$

Also, since the truncation error for Euler's method is $R_T \approx ch$ we find $y(0.5; 5h) - y(0.5; h) \approx y(0.5) + c5h - y(0.5) - ch = 4ch$. Thus, with $h = 0.1$ we get $R_T \approx c5h \approx 5(y(0.5; 5h) - y(0.5; h))/4 \approx 0.021$.

Exercise 6.2 Using Euler's method we find that, using $h = 0.2$, we have $y(0.2; h = 0.2) = y_1 = y_0 + hf(t_0, y_0) = 1.6$. If we instead use $h = 0.1$ we get

$$y(0.1) \approx y_1 = y_0 + hf(t_0, y_0) = 2 - 0.1 \cdot (0^2 - 2) = 1.8 \text{ and } y(0.2) \approx y_2 = y_1 + hf(t_1, y_1) \approx 1.621$$

Considering that the truncation error for Euler is $R_T = Ch$ we also find $y(0.2; 2h) - y(0.2; h) \approx y(0.2) + c2h - y(0.2) - ch = ch$. Thus, with $h = 0.1$ we get $R_T \approx ch \approx (y(0.2; 2h) - y(0.2; h)) \approx 0.021$. For the complete error estimate we also note that the value $y(0.2; h = 0.1) = 1.621$ is rounded to 4 correct decimals. This $|R_B| \leq 0.5 \cdot 10^{-4}$ and $y(0.2) = 1.621 \pm 0.022$.

Exercise 6.3 Euler's method applied to the problem gives the $y_{k+1} = (1 - 100h)y_k + \sin^2(t_k)$. For stability we only look at the homogeneous part, i.e. $y_{k+1} = (1 - 100h)y_k$. Thus the requirement is $-1 \leq 1 - 100h \leq 1$, or, since h is positive, $h \leq 2/100 = 0.02$.

Exercise 6.4 By applying the Euler method we obtain the difference formula $y_{k+1} = (1 - 100h)y_k + h\sin^2(t_k)$. Thus the homogeneous part is stable if $|1 - 100h| \leq 1$, or $h \leq 2/100 = 0.02$.

Exercise 6.5 The method is *implicit* since the right hand side has the term $f(t_{k+1}, y_{k+1})$ and y_{k+1} is the unknown we seek to compute. Thus we must solve an equation in each step. For the stability we apply the method to the test problem $y' = \lambda y$, $y(0) = 1$, and obtain $y_{k+1} = y_k + \frac{h\lambda}{2}(y_k + y_{k+1})$ or

$$y_{k+1} = \left(\frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}}\right) y_k.$$

Thus the Trapezoidal method is stable if $|1 + \frac{h\lambda}{2}| \leq |1 - \frac{h\lambda}{2}|$. This means that $h\lambda/2$ is closer to -1 than to 1. Since h is positive this means that the method is stable for all $\lambda < 0$. So if the solution $y(t) = \mathrm{e}^{\lambda t}$, to the test problem, should be decreasing, i.e. for $\lambda < 0$, then the numerical solution will also be decreasing.

Exercise 6.6 Since $f(t, y) = \lambda y$ we obtain

$$k_1 = hf(t_k, y_k) = h\lambda y_k, \text{ and } k_2 = hf(t_k, y_k + k1) = h\lambda(y_k + k_1) = h\lambda(y_k + h\lambda y_k) = (h\lambda + (h\lambda)^2)y_k.$$

The next iterate y_{k+1} is thus given by

$$y_{k+1} = y_k + \frac{1}{2}(k_1 + k_2) = y_k + \frac{1}{2}(h\lambda + h\lambda + (h\lambda)^2)y_k = (1 + h\lambda + (h\lambda)^2/2)y_k.$$

Thus the method is stable if $|1 + h\lambda + (h\lambda)^2/2| \leq 1$.

Exercise 6.7 We define $v = y'$. Then $v' = y''$ and

$$\begin{pmatrix} y \\ v \end{pmatrix}' = \begin{pmatrix} v \\ ty - 3v - 1 \end{pmatrix}, \quad \begin{pmatrix} y(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Exercise 6.8 Introduce $v = y\prime$. Then $v' = y'' = 2y(1 + y^2)$. Thus the system is

$$\begin{pmatrix} y \\ v \end{pmatrix}' = \begin{pmatrix} v \\ 2y(1 + y^2) \end{pmatrix}, \quad \begin{pmatrix} y(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

If we apply the Euler method we get, with $h = 0.2$,

$$\begin{pmatrix} y_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ v_0 \end{pmatrix} + h \begin{pmatrix} v_0 \\ 2y_0(1 + y_0^2) \end{pmatrix} = \begin{pmatrix} 1.4 \\ 2.8 \end{pmatrix}.$$

Similarily $(y_2, v_2)^T = (1.96, 4.576)^T$. Thus we obtain the approximation $y(0.4) \approx y_2 = 1.96$.

Exercise 6.9 Introduce $v = y'$ to obtain

$$v' + e(y^2 - a)v + y = 0, \quad y(0) = a,\ v(0) = b.$$

The system is thus

$$\begin{pmatrix} y \\ v \end{pmatrix}' = \begin{pmatrix} v \\ -e(y^2 - a)v - y \end{pmatrix}, \quad \begin{pmatrix} y(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}.$$