TANA09 Numerical computations in computer science

Problem Collection

Fredrik Berntsson

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

Hint $\pi = 3.1415926535\ldots$.

Exercise 1.4 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 1.5 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 1.6 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 1.7 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 1.8 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.

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

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

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

Exercise 2.4 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 2.5 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 2.6 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 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 3.3 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 3.4 Determine the positive root of the equation $x = 5(1-\mathrm{e}^{-x})$ with five correct decimals.

Exercise 3.5 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 3.6 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 3.7 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 3.8 Show that Newton-Raphsons has a quadratic rate of convergence if x^* is a single root.

Exercise 3.9 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 3.10 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 Basic matrix operations and linear systems of equations

Exercise 4.1 Suppose $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times n}$, $m > n$. How many operations are required to evaluate the formula $z = (A + I)Bx + y$, where x and y are vectors.

Exercise 4.2 Suppose A, B, and C are matrices and b is a vector. How would you implement the formula

$$x = B^{-1}(2A + I)(C^{-1} + A)b.$$

without computing any matrix inverse? Aim for as few arithmetic operations as possible.

Exercise 4.3 Suppose we want to solve the upper triangular system $Rx = y$ by backwards substitution. Clearly show how many floating point operations are needed.

Exercise 4.4 The following table shows the time t, in seconds, needed to solve linear systems of equations with n unknowns on a computer.

n	1024	2048	4096	8192
t	0.36	2.67	19.18	150.17

The LU-decomposition followed by two triangular systems is used. Is the algorithms computational complexity as expected?

Exercise 4.5 The unit circle can be defined as all points $x = (x_1\,,\, x_2)$ such that $\|x\| = 1$. Draw the unit circle when $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ is used.

Exercise 4.6 Explain what is ment by a matrix norm beeing *induced* from a vector norm. Also show that if A and B are matrices then for an induced norm $\|AB\| \le \|A\|\|B\|$.

Exercise 4.7 Prove the inequality $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty$.

Exercise 4.8 Let $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$. Show that

$$\|uv^T\|_2 = \|u\|_2\|v\|_2.$$

Exercise 4.9 Prove that $\|I\| = 1$ and $\|A\|\|A^{-1}\| \ge 1$ for all matrix norms induced by a vector norm.

Exercise 4.10 Let $x = (1\,,\, -3\,,\, 7)^T$. Compute $\|x\|_1$, $\|x\|_2$, and $\|x\|_\infty$.

Exercise 4.11 Let $\bar{x} = (1.23\,,\, 0.37\,,\, -2.6)^T$ and assume that the elements $\bar{x}_k$ are correctly rounded. Compute both the absolute and relative error measured in $\|\cdot\|_\infty$.

Exercise 4.12 Let

$$A = \begin{pmatrix} 1.2 & 0.3 & -2.7 \\ 3.1 & -0.5 & 3.2 \\ 1.6 & -0.8 & -2.3 \end{pmatrix}.$$

Compute $\|A\|_\infty$.

Exercise 4.13 Formulate the following system of equations on matrix form $Ax = b$:

$$\begin{cases} 2x_1 - x_2 + 2x_3 & = 1, \\ 3x_1 + x_2 - x_3 & = 0, \\ -3x_1 + x_2 + 2x_3 & = -3. \end{cases}$$

Exercise 4.14 Find a permutation matrix P such that

$$P \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_4 \\ x_2 \\ x_3 \end{pmatrix}.$$

Exercise 4.15 Find a permutation matrix P such that

$$P \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 0 & 1 & 2 \\ 3 & 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 3 & 4 & 5 & 6 \\ 9 & 0 & 1 & 2 \\ 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}.$$

Exercise 4.16 Find a Gauss transformation M such that

$$M \begin{pmatrix} 2 \\ 3 \\ 0.6 \\ -1.8 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \\ 0 \end{pmatrix}.$$

Exercise 4.17 Suppose we have a linear system $Ax = b$ where

$$A = \begin{pmatrix} 2 & 1 & -2 \\ -1 & 0 & 3 \\ 1 & 2 & -1 \end{pmatrix} \text{ and } b = \begin{pmatrix} 6 \\ 1 \\ -3 \end{pmatrix}.$$

During the first step of Gaussian elimination we multiply the system with a matrix M_1 such that the new system $M_1Ax = M_1b$ is

$$\begin{pmatrix} 2 & 1 & -2 \\ 0 & 0.5 & 2 \\ 0 & 1.5 & 0 \end{pmatrix} x = \begin{pmatrix} 6 \\ 4 \\ -6 \end{pmatrix}.$$

Give the Gausstransformation M_1.

Exercise 4.18 A Gauss transformation M_1 that eliminates the non-zeros from the first column of a matrix has the form

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{pmatrix}.$$

Show that the matrix can be written as $M_1 = I - me_1^T$. Also give the elements of the vector m in terms of the elements of the matrix $A = (a_{ij})$ and show that $M_1^{-1} = I + me_1^T$.

Exercise 4.19 The Gauss transformation M_2 that is used during the second step of Gaussian elimination has the structure

$$M_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{pmatrix}.$$

Show that M_2 can be written as $M_2 = I - me_2^T$ where $m = (0, 0, m_{32}, m_{42})$. Also show that $M_2^{-1} = I + me_2^T$.

Exercise 4.20 Let

$$A = \begin{pmatrix} 2 & 1 & -2 \\ -3 & 0.5 & -2 \\ 1 & 2.5 & 0 \end{pmatrix}.$$

Is pivoting required during the first step of Gaussian elimination? If so give the appropriate permutation matrix to use.

Exercise 4.21 After one step of Guassian elimination we have

$$\begin{pmatrix} 2 & 1 & -2 \\ 0 & -0.5 & 1.6 \\ 0 & 1.7 & 0.3 \end{pmatrix}.$$

Give the permutation matrix P_2 and the Gausstransformation M_2 to use the the next step of the Guassian elimination. Also carry out the step and given the resulting upper triangular matrix U.

Exercise 4.22 We need to solve a system of equations $Ax = b$, where

$$A = \begin{pmatrix} 1 & 1.2 & -2.3 \\ 2 & 0 & -1 \\ -1 & 1.5 & 2.1 \end{pmatrix} \text{ and } b = \begin{pmatrix} -0.5 \\ 1.2 \\ 1.27 \end{pmatrix}.$$

We compute the decomposition $PA = LU$ where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.5 & 0.8 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 2 & 0 & -1 \\ 0 & 1.5 & 1.6 \\ 0 & 0 & -3.08 \end{pmatrix} \quad P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Take advantage of the decomposition to compute the solution x.

Exercise 4.23 A computer program has computed the decomposition $PA = LU$ and the output is

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -0.7 & 1 & 0 \\ 0.3 & 1.8 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 1.7 & -2.3 & -1.4 \\ 0 & 1.2 & -0.5 \\ 0 & 0 & 3.1 \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Determine if pivoting was used correctly during the computations.

Exercise 4.24 Let a matrix A and a vector $\bar{b}$ be given as

$$A = \begin{pmatrix} 1 & -2 & 0 \\ 2 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \qquad \bar{b} = \begin{pmatrix} -1.03 \\ 1.34 \\ 0.78 \end{pmatrix}$$

where the elements of $\bar{b}$ are correctly rounded. The system of equations $Ax = b$ was solved using Gaussian elimination and the solution $x = (0.1867\,,\, -0.4217\,,\, 0.9667)^T$. Find an upper bound for the relative error in the solution x caused by the rounding errors in the right hand side.

Hint You may use that $\|A^{-1}\|_\infty = 1$.

Exercise 4.25 Let

$$A = \begin{pmatrix} 3.1 & -1.2 & 2.6 \\ 1.5 & -0.7 & 3.6 \\ -4.1 & 1.1 & 0.7 \end{pmatrix}.$$

If we solve the system $Ax = b_1$, where $b = (0.654\,,\, 0.765\,,\, -1.042)^T$ we obtain the solution $x = (1.2595\,,\, 3.5113\,,\, 0.3705)^T$. Give an upper bound for the change in the solution x if we instead use the approximate right hand side $\bar{b} = (0.657\,,\, 0.761\,,\, -1.039)^T$.

Hint Use that $\|A^{-1}\|_\infty = 17.9052$.

Exercise 4.26 Let $PA = LU$ be the LU decomposition. Prove the formula

$$\det(A) = (-1)^k \Pi_{i=1}^n u_{ii}.$$

What is k here?

Exercise 4.27 Let $r = b - A\widehat{x}$ be the residual for an approximate solution to the linear system $Ax = b$. Prove the formula:

$$\|x - \widehat{x}\| \leq \|A^{-1}\|\|r\|.$$

Exercise 4.28 Let

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 3 & -2 \\ 2 & 0 & 1 \end{pmatrix} \text{ och } = \begin{pmatrix} 1.12 \\ -1.07 \\ 2.34 \end{pmatrix}.$$

and the decomposition $PA = LU$, where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 1 & -0.4 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 2.5 & -1.5 \\ 0 & 0 & 1.4 \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

be given. Use the decomposition to compute the determinant $\det(A)$. You may use that $\det(AB) = \det(A)\det(B)$.

5 Least Squares Problems and Orthogonal Decompositions

Exercise 5.1 Suppose we have a series of measurements (x_i, y_i), $i = 1, 2, \ldots, m$, and want to find a function of the type $y = f(x) = c_0 + c_1 x + c_2 \sin(x)$ that best fits the data in the least squares sense. Formulate the problem as an over determined linear system $Ax = b$.

Exercise 5.2 An harmonic wave is described by the amplitude and the phase shift, i.e. $u(t) = A \cdot \sin(t + \phi)$. Suppose we have a sequence of measurements $(t_i, u(t_i))$, $i = 1, 2, \ldots, m$, and want to find A and ϕ using the least squares method. First explain why the least squares method cannot be applied directly. Also rewrite the problem in such a way that the least squares method can be used.

Exercise 5.3 The time needed for a certain algorithm to complete can be described by the formula $t \approx Cn^p$, where n is the problem size, p is the computational complexity, and C is the average time needed for one arithmetic operation. The following table is available

n	100	200	300	400
t	1.213	2.370	3.619	4.875

Formulate the problem of estimating the parameters C and p from the given data as an over determined linear system $Ax = b$.

Exercise 5.4 Suppose Q is an orthogonal matrix. Show that $\|Qx\|_2 = \|x\|_2$, for all vectors x, and thus $\|Q\|_2 = 1$.

Exercise 5.5 If A is both an ortgohonal matrix and an orthogonal projection. What can you conclude about A?

Exercise 5.6 Suppose $A \in \mathbb{R}^{m \times n}$, $m > n$, and that we have the reduced QR decomposition $A = Q_1 R$. Show how the decomposition can be used to find the vector x that minimize $\|Ax - b\|_2$.

Exercise 5.7 Suppose $A \in \mathbb{R}^{m \times n}$, $m > n$, and that we have the QR decomposition

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1 R$$

Show that the linear system $Ax = b$ has an exact solution if $b = Q_1 Q_1^T b$.

Exercise 5.8 Consider the vector a as an $n \times 1$ matrix. Write out its reduced QR decomposition explicitly. Also write down a formula for the solution of the least squares problem $ax \approx b$, where b is a given $n \times 1$ vector.

Exercise 5.9 We are intressted in the least squares problem $\min \|Ax - b\|_2$. Suppose $A = Q_1 R$ is the reduced QR decomposition. Use Q_1 to give a formula for a orthogonal projection P, such that $Pb = r = b - Ax$, where x is the least squares solution.

Exercise 5.10 Compute the reduced QR factorization of the matrix

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

Exercise 5.11 Let $W \in \mathbb{R}^{n\times n}$ be real, symmetric, positive definite, and let $\|\cdot\|_W$ be defined by,

$$\|x\|_W^2 = x^T W x.$$

Verify that $\|x\|_W = 0$ if and only if $x = 0$. Also derive the normal equations for the minimization problem,

$$\min_x \|Ax - b\|_W.$$

Hint Use the Cholesky factorization $W = R^T R$.

Exercise 5.12 Show that $\|A\|_2 = \sigma_1$ and if A^{-1} exists then $\|A^{-1}\|_2 = 1/\sigma_n$.

Exercise 5.13 Let A^T be an $m \times n$ matrix of rank $k < \min(m, n)$. Use the decomposition $A = U\Sigma V^T$ to give an orthogonal basis for $\text{null}(A^T)$.

Exercise 5.14 Show that if $A \in \mathbb{R}^{m\times n}$ has rank n, then $\|A(A^T A)^{-1}A^T\|_2 = 1$.

Exercise 5.15 Suppose the matrix $B \in \mathbb{R}^{m\times n}$ has full column rank. Use the decomposition $B = U\Sigma V^T$ to give a formula for the solution to the the problem

$$\min_x \|Bx\|_2, \text{ subject to } \|x\|_2 = 1.$$

Exercise 5.16 Suppose $A \in \mathbb{R}^{m\times n}$, $m > n$, $\text{rank}(A) = n$, and that we have a factorization $A = U\Sigma V^T$. Clearly demonstrate how the matrices U and V provides basis vectors for the spaces $\text{Range}(A)$ and $\text{null}(A)$. What are the dimension of the range and null space respectively.

Exercise 5.17 Let $A \in \mathbb{R}^{m\times n}$, $m > n$, and $\text{rank}(A) = n$. Demonstrate how the decomposition $A = U\Sigma V^T$ can be used for solving the least squares problem

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2.$$

Give formulas for both the solution x and the redisual $r = b - Ax$.

Exercise 5.18 Let $A \in \mathbb{R}^{m\times n}$, $m < n$, and $\text{rank}(A) = m$. Let $b \in \mathbb{R}^m$. Show that the formula

$$x = \sum_{i=1}^{m} \frac{u_i^T b}{\sigma_i} v_i$$

provides a solution to $Ax = b$. Is the solution unique?

Exercise 5.19 Suppose we want to find the solution to a linear system $Ax = b$, where $\text{rank}(A) = k < n$ so that the solution x is not unique. Demonstrate how the solution x can be split into two parts,

$$x = x_1 + x_2, \qquad x_1 \in \text{null}(A)^{\perp}, \quad \text{and}, \quad x_2 \in \text{null}(A),$$

and how the SVD of A can be used to write expressions for the solution components x_1 and x_2.

Exercise 5.20 Consider the Least Squares problem with linear constraints,

$$\min \|Ax - b\|_2, \qquad \text{for all } x \in \mathbb{R}^n \text{ such that } Bx = 0,$$

where A is $m \times n$, $m > n$, and B is $n \times n$.

a) Suppose $\text{rank}(B) = n$. What is the solution of the least squares problem?

b) Suppose $\text{rank}(B) = k < n$. Show how the SVD can be used to derive a formula for the solution of the least squares problem.

Exercise 5.21 Let $A \in \mathbb{R}^{m \times n}$, where $m >> n$, have full column rank. Use the decomposition $A = U\Sigma V^T$ to develop a criteria that ensures that the linear system $Ax = b$ has a solution. Try and make the criteria as inexpensive as possible to check.

Exercise 5.22 Tikhonov regularization means replacing an ill-conditioned linear system $Ax = b$ by the more stable problem,

$$\min_x \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2,$$

where λ is the regularization parameter. Show that the normal equations of the above least squares problem are

$$(A^T A + \lambda^2 I)x = A^T b.$$

Also derive a formula for the singular values of the matrix $(A^T A + \lambda^2 I)$ and use the result to show that the normal equations are not ill-conditioned (provided λ is selected appropriately). Finally derive a formula for the solution x_λ.

6 Polynomial and Spline Interpolation

Exercise 6.1 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 6.2 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 6.3 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 6.4 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 6.5 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 6.6 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 6.7 Let $p(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$ 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 6.8 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 6.9 Let

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

Is $s(x)$ a cubic spline?

Exercise 6.10 Let

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

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

Exercise 6.11 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 6.12 Consider a case where $s(x)$ is defined by two cubic polynomials,

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

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

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

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

Exercise 6.14 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

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

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 6.15 In order to obtain a unique interpolating spline $s(x)$ wew 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
$\max\|s(x) - f(x)\|$	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.

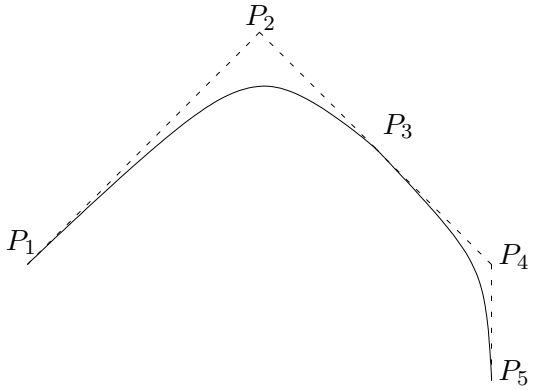
Exercise 6.16 A quadratic Beziér curve is given by the expression

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

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 6.17 $P_1 = (1\,,\,0)^T$, $P_2 = (1\,,\,3)^T$, $P_3 = (4\,,\,3)^T$ 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 6.18 A cubic Beziér curve is given by

$$p(t) = (1-t)^3 P_1 + 3(1-t)^2 t P_2 + 3(1-t)t^2 P_3 + t^3 P_4, \quad 0 < t < 1,$$

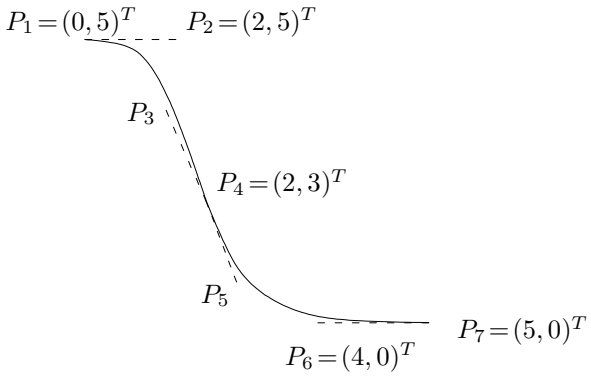
where P_1, P_2, P_3 and P_4 are control points.

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 6.19 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.

7 Taylor expansions and extrapolation

Exercise 7.1 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 7.2 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 7.3 We compute an approximation of an integral using a numerical method $T(h)$ with a truncationb 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 7.4 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 7.5 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 7.6 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 $|R_T| \leq 10^{-8}$.

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 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 1.4 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 1.5 The numbers are stored as follows

0	10000010	00101000...00	and	0	1000000	01010100...00

Exercise 1.6 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 1.7 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

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

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 1.8 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

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

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 $|\Delta x|, |\Delta y|, |\Delta < | \leq 0.5 \cdot 10^{-2}$. 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 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 2.3 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 2.4 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

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

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

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 2.5 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 2.6 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 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 3.3 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 3.4 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 3.5 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 3.6 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 3.7 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 3.8 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 3.9 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 3.10 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 Basic matrix operations and linear systems of equations

Exercise 4.1 We evaluate the expression using the following operations

$$z = (A + I)Bx + y = (A + I)x_1 + y = Ax_1 + x_1 + y = x_2 + x_1 + y = x_3 + y = x_4$$

Computing the matrix vector product $x_1 = Bx$ requires mn multiplications and additions so a total of $2mn$ operations and the product $x_2 = Ax_1$ requires $2m^2$ mult/adds. The remaining three vector additons require m additions (as $y, x_1 \in \mathbb{R}^m$). So the operation count is $m(2m + 2n + 2)$.

Exercise 4.2 We aim to keep intermediate results small. Multiplication by an inverse is dealt with by solving the corresponding linear system, i.e. compute $z = A^{-1}x$ by solving $Az = x$. The order of computation is

$$z_1 = Ab, \quad Cz_2 = b, \quad z_3 = z_1 + z_2, \quad z_4 = Az_3, \quad z_5 = 2z_4 + z_3, \text{ and finally } Bx = z_5.$$

All intermediate results are vectors.

Exercise 4.3 An upper triangular system $Rx = y$ can be solved using,

$$x_i = \left(b_i - \sum_{j=i+1}^{n} r_{ij}x_j \right) / r_{ii}, \quad i = n, n-1, \ldots, 1.$$

Thus in step i exactly $(i - 1)$ multiplications and additions are needed. Also exactly one division. Thus the total amount of work is

$$\sum_{i=1}^{n}(i - 1) \approx \frac{n^2}{2},$$

multiplications and additions or n^2 arithmetic operations.

Exercise 4.4 The LU-decomposition and two triangular systems require $2n^3/3 + 2n^2$ arithmetic operations. For large n we have $t(n) \approx c \cdot n^3$, where c is the average time for one operation, and thus

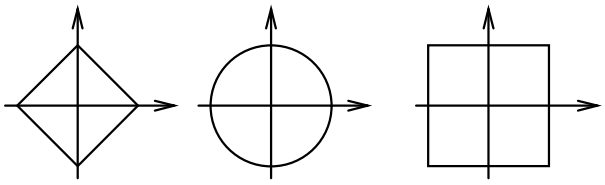
$$\frac{t(2n)}{t(n)} \approx \frac{c2^3n^3}{cn^3} = 8.$$

If we compute the same quotients using the table we obtain

n	1024	2048	4096
$t(2n)/t(n)$	7.4166	7.1835	7.8295

The results look promising and fits quite well with the idea that the computational complexity is $\mathcal{O}(n^3)$.

Exercise 4.5 The unit circles are



Exercise 4.6 A matrix norm is *induced* if its definition is basede on a vector norm, i.e.

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$$

For such norms we have

$$\|AB\| = \max_{x\neq 0} \frac{\|ABx\|}{\|x\|} = \max_{x\neq 0} \frac{\|ABx\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \leq \left(\max_{x\neq 0} \frac{\|ABx\|}{\|Bx\|}\right)\left(\max_{x\neq 0} \frac{\|Bx\|}{\|x\|}\right) \leq \max_{y\neq 0} \frac{\|Ay\|}{\|y\|}\|B\| \leq \|A\|\|B\|.$$

Exercise 4.7 Demonstrate the first inequality by

$$\|x\|_\infty^2 = \max_{1\leq i\leq n} |x_i|^2 \leq \sum_{i=1}^n |x_i|^2 = \|x\|_2^2.$$

Also, since $|x_i| \leq \|x\|_\infty$, we have

$$\|x\|_2^2 = \sum_{i=1}^n |x_i|^2 \leq \sum_{i=1}^n \|x\|_\infty^2 = n\|x\|_\infty.$$

Exercise 4.8 Recall the definition

$$\|uv^T\|_2 = \max_{x\in\mathbb{R}^n} \frac{\|uv^Tx\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^n} \frac{|v^Tx|\|u\|_2}{\|x\|_2}.$$

The Cauchy-Schwarz inequality is $|v^Tx| \leq \|v\|_2\|x\|_2$ with equality for $x = v$. So

$$\|uv^T\|_2 = \frac{|v^Tv|\|u\|_2}{\|v\|_2} = \|v\|_2\|u\|_2.$$

Exercise 4.9 First from the definition of the matrix norm, and since $Ix = x$ we have

$$\|I\| = \max_{x\neq 0} \frac{\|Ix\|}{\|x\|} = \max_{x\neq 0} \frac{\|x\|}{\|x\|} = 1, \text{ so } 1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|.$$

Exercise 4.10 If $x = (1, -3, 7)^T$ then $\|x\|_\infty = 7$, $\|x\|_2 = \sqrt{59}$ and $\|x\|_1 = 11$.

Exercise 4.11 If $\bar{x} = (1.23\,,\ 0.37\,,\ -2.6)^T$ is correctly rounded then the error vector satisfies $|\delta x| \leq (0.005\,,\ 0.005\,,\ 0.05)^T$. Thus $\|x - \bar{x}\|_\infty \leq 0.5 \cdot 10^{-1}$ is the absolute error and $\|x - \bar{x}\|_\infty / \|x\|_\infty \leq 0.05/2.6 < 0.02$ is the relative error.

Exercise 4.12 From the second row we get the largest sum and $\|A\|_\infty = 3.1 + 0.5 + 3.2 = 6.8$.

Exercise 4.13 The linear system is $Ax = b$ where

$$A = \begin{pmatrix} 2 & -1 & 2 \\ 3 & 1 & -1 \\ -3 & 1 & 2 \end{pmatrix} \text{ and } b = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}.$$

Exercise 4.14 The rows should appear in the order $1 - 4 - 2 - 3$ and thus

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Exercise 4.15 The rows should appear in the order $4 - 3 - 1 - 2$ which gives

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Exercise 4.16 The multipliers are $m_3 = 0.6/3 = 0.2$ and $m_4 = -1.8/3 = -0.6$. Therefore the Gauss transformation is

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.2 & 1 & 0 \\ 0 & -0.6 & 0 & 1 \end{pmatrix}.$$

Exercise 4.17 The multipliers are $m_{21} = -1/2 = -0.5$ and $m_{31} = 1/2 = 0.5$. Therefore the Gauss transformation is

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.5 & 0 & 1 \end{pmatrix}.$$

Exercise 4.18 Set $m = (0\,,\, m_{21}\,,\, m_{31})^T$ and find that

$$M_1 = I - me_1^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ m_{21} \\ m_{31} \end{pmatrix} (1\,,\, 0\,,\, 0)$$

has the desired structure. To show that $M_1^{-1} = I + me_1^T$ we do

$$(I - me_1^T)(I + me_1^T) = I - me_1^T + me_1^T - me_1^T me_1^T = I$$

where $e_1^T m = 0$ since m has a zero in the first position. The elements of m are chosen so that $0 = a_{i1} - m_{i1}a_{11}$ and thus $m_{i1} = a_{i1}/a_{11}$, $i = 2, 3$.

Exercise 4.19 Set $m = (0\,,\,0\,,\,m_{32}\,,\,m_{42})^T$ and find that

$$M_2 = I - me_2^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ m_{32} \\ m_{42} \end{pmatrix} (0\,,\,1\,,\,0\,,\,0)$$

has the desired structure. To show that $M_2^{-1} = I + me_2^T$ we do

$$(I - me_2^T)(I + me_2^T) = I - me_2^T + me_2^T - me_2^T me_2^T = I$$

where $e_2^T m = 0$ since m has zeros in the first two positions.

Exercise 4.20 Pivoting is required as $|a_{21}| > |a_{11}$. The permutation

$$P_{12} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

should be used to switch rows one and two.

Exercise 4.21 We need to switch rows two and three to obtain

$$\begin{pmatrix} 2 & 1 & -2 \\ 0 & 1.7 & 0.3 \\ 0 & -0.5 & 1.6 \end{pmatrix}.$$

and then use the multiplier $m_{32} = -0.5/1.7 = -0.2941$. The permutation matrix and Gausstransformation are

$$P_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \text{ and } M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -0.2941 & 1 \end{pmatrix}.$$

The new element on position $(3,3)$ will be $1.6 - (-0.2941) \cdot 0.3 = 1.5118$. The upper triangular matrix is

$$U = \begin{pmatrix} 2 & 1 & -2 \\ 0 & 1.7 & 0.3 \\ 0 & 0 & 1.5118 \end{pmatrix}.$$

Exercise 4.22 In order to take advantage of the decomposition we rewrite $Ax = b$ as $PAx = Pb$ and $L(Ux) = Pb$. By introducing the intermediate variable $y = Ux$ we obtain two triangular systems $Ly = Pb$ and $Ux = y$. First solve

$$Ly = Pb \text{ or } \begin{pmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.5 & 0.8 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} -0.5 \\ 1.2 \\ 1.27 \end{pmatrix} = \begin{pmatrix} 1.2 \\ 1.27 \\ -0.5 \end{pmatrix}$$

to obtain $y = (1.2\,,\,1.87\,,\,-2.60)^T$. Next we solve $Ux = y$ by backwards substitution and find that $x = (1.02\,,\,0.34\,,\,0.84)^T$.

Exercise 4.23 Since the multiplier $m_{32} = 1.8 > 1$ pivoting cannot have been used correctly.

Exercise 4.24 Since the elements of $\bar{b}$ are correctly rounded to two correct decimals we have $\|\delta b\|_\infty \leq 0.5 \cdot 10^{-2}$. Thus the relative error in the solution is bounded by

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \|A\|_\infty \|A^{-1}\|_\infty \frac{\|\delta b\|_\infty}{\|b\|_\infty} \leq 3 \cdot 1 \cdot \frac{0.5 \cdot 10^{-2}}{1.34} < 0.012.$$

Exercise 4.25 The error in the right hand side is $\delta b = (0.003\,,\ 0.004\,,\ -0.003)^T$. We find that the error in the result is bounded by

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \|A\|_\infty \|A^{-1}\|_\infty \frac{\|\delta b\|_\infty}{\|b\|_\infty} \leq 6.9 \cdot 17.9052 \cdot \frac{0.004}{1.039} < 0.48.$$

The absolute error is $\|\delta x\|_\infty \leq 0.48 \cdot 3.5113 < 1.7$.

Exercise 4.26 Use the LU decomposition of A to obtain

$$A = P^T LU, \text{ so } \det(A) = \det(P^T)\det(L)\det(U).$$

Here both L and U are triangular so the determinant is the product of the diagonal elements. Also P is a permutation matrix. If we exchange two rows in a matrix then the determinant changes sign. So k is the number of row exchanges that actually occured during the Guassian elimination when computing the LU decomposition.

Exercise 4.27 Let $r = b - A\widehat{x}$. Then $A^{-1}r = A^{-1}b - A^{-1}A\widehat{x} = x - \widehat{x}$ and we obtain $\|x - \widehat{x}\| \leq \|A^{-1}\|\|r\|$.

Exercise 4.28 Since $P = I$ we have $\det(P) = 1$ and we obatin

$$\det(A) = \det(P^T)\det(L)\det(U) = 1 \cdot 1 \cdot (2 \cdot 2.5 \cdot 1.4) = 7.$$

5 Least Squares Problems and Orthogonal Decompositions

Exercise 5.1 The model is $y_i = c_0 + c_1 x_i + c_2 \sin(x_i)$ and each measurement results in a row of a linear system $Ax = b$. We obtain

$$\begin{pmatrix} 1 & x_1 & \sin(x_1) \\ 1 & x_2 & \sin(x_2) \\ \vdots & \vdots & \vdots \\ 1 & x_m & \sin(x_m) \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

Exercise 5.2 The least squares method can not be used since the model is not linear. We rewrite the model as $u(t) = A\cos(\phi)\sin(t) + A\sin(\phi)\cos(t)$. The resulting over determined linear system is

$$\begin{pmatrix} \sin(t_1) & \cos(t_1) \\ \sin(t_2) & \cos(t_2) \\ \vdots & \vdots \\ \sin(t_m) & \cos(t_m) \end{pmatrix} \begin{pmatrix} A\cos(\phi) \\ A\sin(\phi) \end{pmatrix} = \begin{pmatrix} u(t_1) \\ u(t_2) \\ \vdots \\ u(t_m) \end{pmatrix}.$$

where the first have to find a least squares solution x and then find the parameters A and ϕ by, e.g., $\phi = \text{atan}(x_2/x_1)$.

Exercise 5.3 We rewrite the model as $\log(t_i) = \log(C) + p\log(n_i)$, and obtain the linear system

$$\begin{pmatrix} 1 & \log(100) \\ 1 & \log(200) \\ 1 & \log(300) \\ 1 & \log(400) \end{pmatrix} \begin{pmatrix} log(C) \\ p \end{pmatrix} = \begin{pmatrix} \log(1.213) \\ \log(2.370) \\ \log(3.619) \\ \log(4.875) \end{pmatrix}.$$

Exercise 5.4 Since Q is orthogonal $Q^TQ = I$. So

$$\|Qx\|_2 = (Qx)^T(Qx) = x^TQ^TQx = x^Tx = \|x\|_2.$$

This means that $\|Q_2\|_2 = \max \|Qx\|_2/\|x\|_2 = \max 1 = 1$.

Exercise 5.5 First Range$(A) = \mathbb{R}^n$ since A is orthogonal and thus has linearly independent columns. So A is an orthogonal projection on the whole of $\mathbb{R}^n$. So $Ax = x$ for every $x \in \mathbb{R}^n$ so $A = I$ is the identity matrix.

Exercise 5.6 Let

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1 R$$

where $Q = (Q_1, Q_2)$. Since Q is orthogonal we find that

$$\|Ax - b\|_2^2 = \|Q^T(Ax - b)\|_2^2 = \| \begin{pmatrix} R \\ 0 \end{pmatrix} x - \begin{pmatrix} Q_1^Tb \\ Q_2^Tb \end{pmatrix} \|_2^2 = \|Rx - Q_1^Tb\|_2^2 + \|Q_2^Tb\|_2^2.$$

The minimum is achived for $x = R^{-1}Q_1^Tb$. Thus only the reduced QR decomposition is needed.

Exercise 5.7 If $A = Q_1 R$ then range(A) = span(Q_1). So $P = Q_1 Q_1^T$ is an orthogonal projection onto the range space of A. Thus if $b = Q_1 Q_1^T b$ then b belongs to the space range(A) which by definition means that $Ax = b$ has a solution.

Exercise 5.8 The vector a can be seen as a matrix in $\mathbb{R}^{n\times 1}$. This means that

$$a = (a/\|a\|_2)\|a\|_2 = Q_1 R$$

where $Q_1 \in \mathbb{R}^{n\times 1}$ and $R \in \mathbb{R}^{1\times 1}$. The formula for the least squares solution can be written using the normal equations $a^T a x = a^T b$ or $x = (a^T b)/(a^T a)$. This is the same as $x = R^{-1} Q_1^T b$ with the decomposition above.

Exercise 5.9 Since $x = R^{-1} Q_1^T b$ is the least squares solution we can compute the residual as $r = b - Ax = b - Q_1 R R^{-1} Q_1^T b = (I - Q_1 Q_1^T) b = Pb$. We can also recall that $Ax = Q_1 Q_1^T b$ from the geometrical interpretation of the least squares problem.

Exercise 5.10 The matrix has orthogonal columns, i.e. if $A = (a_1, a_1)$ then $(a_1, a_2) = 0$. Thus the QR decomposition is

$$A = (a_1/\|a_1\|_2\,,\, \|a_2/\|a_2\|_2) \begin{pmatrix} \|a_1\|_2 & 0 \\ 0 & \|a_2\|_2 \end{pmatrix} = Q_1 R.$$

The numbers are not very important.

Exercise 5.11 Let $W = R^T R$ be the Cholesky decomposition ans rewrite

$$\|x\|_W^2 = x^T W x = x^T R^T R x = (Rx)^T (Rx) = \|Rx\|_2^2.$$

Since R is non-singular $Rx = 0$ if and only if $x = 0$. In order to find the normal equations we use

$$\|Ax - b\|_W = \|R(Ax - b)\|_2.$$

The normal equations are now $(RA)^T (RA)x = (RA)^T (Rb)$ or $A^T W A x = A^T W b$.

Exercise 5.12 If $A = U\Sigma V^T$ where U, V are orthogonal and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ then $\|A\|_2 = \|U\Sigma V^T\|_2 = \|\Sigma\|_2$. The norm of a diagonal matrix can be computed by

$$\|\Sigma\|_2 = \max_{y\in\mathbb{R}^n} \frac{\|\Sigma y\|_2}{\|y\|_2} = \max_{y\in\mathbb{R}^n} \sqrt{\frac{\sum \sigma_i^2 y_i^2}{\sum y_i^2}} \leq \sigma_1 \max_{y\in\mathbb{R}^n} \sqrt{\frac{\sum y_i^2}{\sum y_i^2}} = \sigma_1,$$

with equality for $y = e_1$. Thus $\|A\|_2 = \sigma_1$. If A^{-1} exists then $A^{-1} = V\Sigma^{-1}U^T$ and $\|A^{-1}\|_2 = \|\Sigma^{-1}\|_2$. Since the diagonal elements of Σ^{-1} are $1/\sigma_i$ the largest diagonal element is $1/\sigma_n$ and $\|A^{-1}\|_2 = 1/\sigma_n$.

Exercise 5.13 Let $A^T = V\Sigma^T U^T$. If $y \in \text{span}(u_{k+1}, \ldots, u_m)$ then $u_i^T y = 0$ for $i = 1, \ldots, k$. This is the null space of A^T.

Exercise 5.14 First compute $(A^TA)^{-1} = (V\Sigma^TU^TU\Sigma V^T)^{-1} = V(\Sigma^T\Sigma)^{-1}V^T$. Here $\Sigma^T\Sigma = \text{diag}(\sigma_i^2) \in \mathbb{R}^{n\times n}$. Thus $A(A^TA)^{-1}A^T = U\Sigma V^TV(\Sigma^T\Sigma)^{-1}V^TV\Sigma^TU^T = U\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^TU^T$. Since U is orthogonal $\|A(A^TA)^{-1}A\|_2 = \|\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T\|_2$. Evaluate the product of the diagonal matrices to obtain

$$\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad I \in \mathbb{R}^{n\times n}.$$

The norm is the largest diagonal entry, i.e. 1.

Exercise 5.15 Let $B = U\Sigma V^T$. Since $V = (v_1, \ldots, v_n)$ provides a basis for $\mathbb{R}^n$ any x can be written

$$x = \sum_{i=1}^n c_iv_i \Longrightarrow By = \sum_{i=1}^n c_i\sigma_iu_i.$$

If $\|x\|_2 = 1$ then $\sum c_i^2 = 1$. So

$$\|Bx\|_2^2 = \sum_{i=1}^n \sigma_i^2c_i^2 \geq \sigma_n \sum_{i=1}^n c_i^2 = \sigma_n^2,$$

with equality for $c = e_n$. So the minimum is σ_n and it is obtained for $x = \pm v_n$.

Exercise 5.16 The decomposition $A = U\Sigma V^T$ can be written

$$A = \sum_{i=1}^n \sigma_iu_iv_i^T,$$

where $\sigma_n > 0$ as $\text{rank}(A) = n$. This means that $Av_i = \sigma_iu_i \neq 0$ for $i = 1, \ldots, n$. So the null space is only the trivial one $\text{null}(A) = \{0\}$ with dimension 0. Similarily, if y belongs to the range then there is an x such that $y = Ax$, or

$$y = Ax = \sum_{i=1}^n \sigma_i(v_i^Tx)u_i,$$

so the y is a linear combination of $\{u_1, \ldots, u_n\}$. Thus $\text{range}(A) = \text{span}(u_1, \ldots, u_n)$ and the dimension of the range is n.

Exercise 5.17 Let $A = U\Sigma V^T$. Since $U = (u_1, \ldots, u_m)$ is a basis for $\mathbb{R}^m$ we can write

$$b = \sum_{i=1}^m (u_i^Tb)u_i,$$

Similarily, $V = (v_1, \ldots, v_n)$ is a basis for $\mathbb{R}^n$ so

$$Ax = A(\sum_{i=1}^n (v_i^Tx)v_i) = \sum_{i=1}^n \sigma_i(v_i^Tx)u_i.$$

We obtain

$$\|Ax - b\|_2^2 = \|\sum_{i=1}^n (\sigma_i(v_i^Tx) - (u_i^Tb))u_i - \sum_{i=n+1}^m (u_i^Tb))u_i\|_2^2 = \sum_{i=1}^n |\sigma_i(v_i^Tx) - (u_i^Tb)|^2 + \sum_{i=n+1}^m |u_i^Tb|^2.$$

The minimum is obtained for $\sigma_i(v_i^T x) - u_i^T b$ for $i = 1, \ldots, n$, so

$$x = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i.$$

For this particular x we get

$$r = b - Ax = \sum_{i=n+1}^{m} (u_i^T b) u_i.$$

Exercise 5.18 Compute Ax to obtain

$$Ax = A(\sum_{i=1}^{m} \frac{u_i^T b}{\sigma_i} v_i) = \sum_{i=1}^{m} \frac{u_i^T b}{\sigma_i} A v_i = \sum_{i=1}^{m} \frac{u_i^T b}{\sigma_i} \sigma_i u_i = \sum_{i=1}^{m} (u_i^T b) u_i = b,$$

where the last equality holds since $U = (u_1, \ldots, u_m)$ provides an orthogonal basis for $\mathbb{R}^m$ which is the space b belongs to.

Since $m < n$ the matrix has a null space $\text{null}(A) = \text{span}(v_{m+1}, \ldots, v_n)$. If x_2 belongs to the nullspace then $A(x + x_2) = Ax = b$ so the solution is not unique. Since the above formula for x does not include a component from the null space it can be characterized as

$$\min \|x\|_2 \text{ such that } Ax = b,$$

that is the *minimum norm* solution of the linear system $Ax = b$.

Exercise 5.19 Since $\text{rank}(A) = k$ we note that $\{v_{k+1}, \ldots, v_n\}$ is a basis for $\text{null}(A)$ and $\{v_1, \ldots, v_k\}$ is a basis for its orthogonal complement $(\text{null}(A))^{\perp}$. Thus for evey x we can write

$$x = x_1 + x_2 = (\sum_{i=1}^{k} c_i v_i) + (\sum_{i=k+1}^{n} c_i v_i).$$

In order to determine x_1 we compute

$$Ax = A(x_1 + x_2) = Ax_1 + 0 = \sum_{i=1}^{k} c_i \sigma_i u_i = b = \sum_{i=1}^{m} (u_i^T b) u_i.$$

Where $(u_i^T b) = 0$, for $i = k + 1, \ldots, m$, or a solution doesn't exist. Thus

$$x_1 = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i \text{ and } x_2 = \sum_{i=k+1}^{n} c_i v_i,$$

where $c_i, i = k + 1, \ldots, n$, are undetermined parameters.

Exercise 5.20 a) If B has full rank then $Bx = 0$ if and only if $x = 0$ so the unique, and only feasible, solution is precisely $x = 0$.

b) If $\text{rank}(B) = k < n$ then B has a non-trivial null space and write $V = (V_k, V_{n-k})$ so that the null space is given by V_{n-k} then the feasible solutions are $x = V_{n-k}c$, $c \in \mathbb{R}^{n-k}$. So in fact we have a regular least squares problem

$$\min_{c \in \mathbb{R}^{n-k}} \|(AV_{n-k})c - b\|_2 \text{ and } x = V_{n-k}c.$$

The above qualifies as a formula. Otherwise continmue and write the normal equations for the above least squares problem.

Exercise 5.21 Let $A = U\Sigma V^T$ and $U = (u_1, \ldots, u_m)$. A solution exists if $b \in \text{range}(A) = \text{span}(u_1, \ldots, u_n)$. We can check this by, for instance, verifying that $u_i^T b = 0$, for $i = n+1, \ldots, m$. If $m >> n$ it is cheaper to instead check if

$$b - \sum_{i=1}^{n} (u_i^T b) u_i = 0.$$

If we split the matrix $U = (U_1, U_2)$ then the same criteria can be written as $U_2^T b = 0$ or $b - U_1 U_1^T b = 0$.

Exercise 5.22 The normal equations can be derived by the identity

$$\min_x \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2 = \min_x \left\| \begin{pmatrix} Ax - b \\ \lambda x \end{pmatrix} \right\|_2 = \min_x \left\| \begin{pmatrix} A \\ \lambda I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2.$$

The last is a regular least squares problem with an extended matrix. The normal equations are

$$\begin{pmatrix} A^T & \lambda I \end{pmatrix} \begin{pmatrix} A \\ \lambda I \end{pmatrix} x = \begin{pmatrix} A^T & I \end{pmatrix} \begin{pmatrix} b \\ 0 \end{pmatrix} \text{ or } (A^T A + \lambda^2 I)x = A^T b.$$

Now we can derive the solution formula using the decomposition $A = U\Sigma V^T$. Since $A^T A + \lambda I = V\Sigma^T \Sigma V^T + \lambda^2 V V^T = V(\Sigma^T \Sigma + \lambda^2 I)V^T$ and $A^T b = V\Sigma U^T b$ we obtain the solution

$$x_\lambda = V(\Sigma^T \Sigma + \lambda I)^{-1} \Sigma U^T b = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \lambda^2} (u_i^T b) v_i.$$

Too see that the normal equations are not ill-conditioned we look at $A^T A$ which has singular values $\sigma_i^2 + \lambda^2 \geq \lambda^2$. So the addition of the regularization parameter removes the small singular values and makes the condition number smaller.

6 Polynomial and Spline Interpolation

Exercise 6.1 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 6.2 A polynomial, $p_n(x)$, of degree n, can be written $p_n(x) = c_0 + c_1x + \ldots + c_nx^n$. Thus there are $n + 1$ parameters to determine and the polynomial can satisfy exactly $n + 1$ interpolation conditions.

Exercise 6.3 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 $|f''(x)| = |-x^{-2}| \geq 1$ 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 6.4 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 6.5 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 6.6 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 6.7 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 6.8 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 6.9 The function $s(x)$ is a cubic spline since $s(x)$, $s'(x)$ and $s''(x)$ are continuous at $x = 1$.

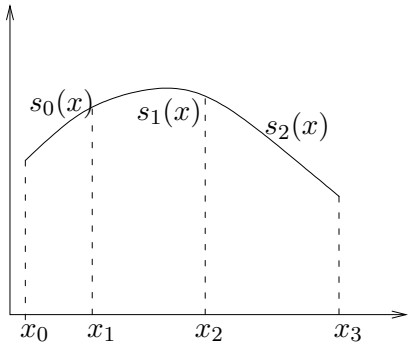
Exercise 6.10 The conditions that has to be satisfied are $s(1) = a + 1 = b + c$, $s'(1) = a = 3b + 2c$ and $s''(1) = 0 = 6b + 2c$. The solution is $a = -3$, $b = 1$ and $c = -3$.

Exercise 6.11 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 6.12 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 6.13 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 6.14 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 6.15 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

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

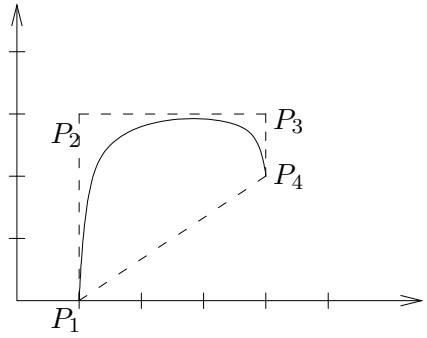
In both cases the quitient is sufficiently close to $2^4 = 16$ to conclude that $p = 4$.

Exercise 6.16 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 6.17 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 6.18 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

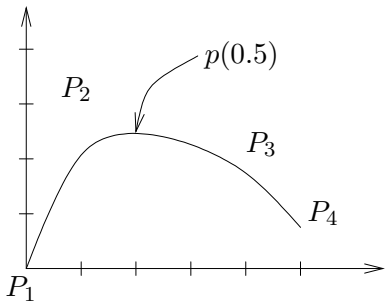
is located within the convex hull we note that the weights in the expression for $p(t)$ are calculated from the identity

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

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 6.19 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.

7 Taylor expansions and extrapolation

Exercise 7.1 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 7.2 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 7.3 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 7.4 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 7.5 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 7.6 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 $h < 10^-8/0.4 = 2.5 \cdot 10^{-8}$.