## Excercises in Numerical Methods



Fredrik Berntsson (fredrik.berntsson@liu.se)
Linköpings universitet

## Contents

1 Basic Concepts and Floating Point systems ..... 3
2 Error Analysis ..... 5
3 Non-linear equations ..... 7
4 Polynomial and Spline Interpolation ..... 10
5 Integration, Differentiation and Extrapolation ..... 15
6 Differential Equations ..... 18
7 Basic matrix operations and linear systems of equations ..... 19
8 Least Squares Problems and Orthogonal Decompositions ..... 23
Answers ..... 26
1 Basic Concepts and Floating Point systems ..... 26
2 Error Analysis ..... 28
3 Non-linear equations ..... 31
4 Polynomial and Spline Interpolation ..... 35
5 Integration, Differentiation and Extrapolation ..... 39
6 Differential Equations ..... 42
7 Basic matrix operations and linear systems of equations ..... 44
8 Least Squares Problems and Orthogonal Decompositions ..... 49

## 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 $\pi$ 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 $\pi$ 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 $)$ | $\mathrm{f}(23$ 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 $m$, $\beta$ 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 $\mathrm{f}[a \cdot b]$ and $\mathrm{f}[a / b]$. 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 $f(x)=\mathrm{e}^{b x}$, 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 $\pi$ 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 $x \rightarrow 0$. 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 $\mu$.

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 $\mu$ 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

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

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 $\mu$. 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

$$
\lim _{x \rightarrow 0} f(x)=\frac{1}{6}
$$

but when we compute the expression for $x=10^{-7}$ we get the difference $\left|\bar{f}\left(10^{-7}\right)-\frac{1}{6}\right| \approx 5.4 \cdot 10^{-3}$. 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 $\mu$.

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

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

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 $\mu$. 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 $\Delta f$. The general error propagation formula states that

$$
|\Delta f| \lesssim\left|\frac{\partial f}{\partial x}\right||\Delta x|
$$

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

$$
|\Delta f| \lesssim\left|f^{\prime \prime}(\bar{x})\right| \frac{|\Delta x|^{2}}{2}
$$

## 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 $1.7 \leq f^{\prime}(x) \leq 2.2$ near the root $x^{*}$. Estimate the error in $\bar{x}$ for this case.

Exercise 3.5 Determine the positive root of the equation $x=5\left(1-\mathrm{e}^{-x}\right)$ 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 $f(x)=2 \cos (x)-4 x$ which has a positive root $x^{*} \approx 0.45$. Show theoretically that the fixed point iteration $x_{k+1}=\phi\left(x_{n}\right)=\frac{1}{2} \cos \left(x_{n}\right)$ 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 $f(x)=x-3 e^{-x}=0$ 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

$$
\begin{gathered}
\text { (i) } x_{n+1}=3 e^{-x_{n}}, \quad \text { (ii) } x_{n+1}=\left(2 x_{n}+3 e^{-x_{n}}\right) / 3, \\
\text { (iii) } x_{n+1}=1.05 x_{n}+3 e^{-x_{n}}, \quad \text { (iv) } x_{n+1}=\left(x_{n}+3 e^{-x_{n}}\right) / 2 .
\end{gathered}
$$

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

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

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

| $k$ | $x_{k}$ | $f\left(x_{k}\right)$ |
| :---: | :---: | :---: |
| 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 $f(x)=$ $x^{3}-9 x^{2}+24 x-20$. We obtain the following iteration sequence

| $k$ | $x_{k}$ | $\left\|x_{k}-x^{*}\right\|$ |
| :---: | :---: | :---: |
| 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 $x_{k}=\varphi\left(x_{k-1}\right)$ 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,

$$
\left|x_{k}-\sqrt{a}\right| \lesssim \frac{1}{2}\left|\varphi^{\prime \prime}(\sqrt{a})\right|\left|x_{k-1}-\sqrt{a}\right|^{2} .
$$

Take advantage of this and determine the number of iterations needed to achive an error bound $\left|x_{k}-\sqrt{a}\right| \leq \mu=1.1 \cdot 10^{-16}$ 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 $a_{j}=1+3 j / n, j=0, \ldots, n-1$, and compute $\sqrt{a_{j}}$ exactly. Clearly demonstrate how the values $\left(\sqrt{a_{j}}, a_{j}\right)$ 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 $p_{n}\left(x_{i}\right)=f_{i}$ 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 $\left.\left\{x_{i}, y_{i}\right)\right\}_{i=1}^{n}$, where $x_{1}=1, x_{n}=4$ and $h=x_{i+1}-x_{i}$ is the stepsize. In the table we store approximate values $y_{i} \approx \log \left(x_{i}\right)$, 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 $\ell_{i}\left(x_{j}\right)=1$ 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 $p(x)=c_{0}+c_{1} x+c_{2} x^{2}+c_{3} x^{3}$ be a cubic polynomial. We want to find values for the coefficients so that $p(0)=p(1)=0$ and $p^{\prime}(0)=p^{\prime}(1)=1$. Show how to derive a linear system of equations such that the solution $c=\left(c_{0}, c_{1}, c_{2}, c_{3}\right)^{T}$ 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

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

Is $s(x)$ a cubic spline?
Exercise 4.12 Let

$$
s(x)= \begin{cases}a x+1 & 0 \leq x<1 \\ b x^{3}+c x^{2} & 1 \leq x<2\end{cases}
$$

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. $s^{\prime}(0)=f^{\prime}(0)$ and $s^{\prime}(1)=f^{\prime}(1)$. Give the expression for $s(x)$.

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

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

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 |
| :---: | :---: | :---: | :---: |
| $\mathrm{~s}(\mathrm{x})$ | 0.9 | 2.0 | 6.7 |

with the end point conditions $s^{\prime}(0)=0.1$ and $s^{\prime}(2)=7.3$.

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

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

Is $s(x)$ a cubic spline? Present the calculations. Also determine if $s(x)$ is a natural cubic spline?
Exercise 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 $\left[x_{k}, x_{k+1}\right]$ we let $s(x)$ be given by a cubic polynomial

$$
s_{k}(x)=a_{k}+b_{k}\left(x-x_{k}\right)+c_{k}\left(x-x_{k}\right)^{2}+d_{k}\left(x-x_{k}\right)^{3}, \quad x_{k} \leq x<x_{k+1} .
$$

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 $\left\{x_{k}\right\}_{k=0}^{N}$.
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. $s^{\prime}(a)=f^{\prime}(a)$ and $s^{\prime}(b)=f^{\prime}(b)$. 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 $h=\max \left|x_{k+1}-x_{k}\right|$ as $C h^{p}$, 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

$$
p(t)=(1-t)^{2} P_{1}+2(1-t) t P_{2}+t^{2} P_{3}, \quad 0<t<1,
$$

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

$$
p(t)=(1-t)^{3} P_{1}+3(1-t)^{2} t P_{2}+3(1-t) t^{2} P_{3}+t^{3} P_{4}, \quad 0<t<1,
$$

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 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 $y=-2 x+b$ 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 $f^{\prime}(x)$ by the difference formula

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

Show that the truncation error can be written as $f^{\prime}(x)-D f(x) \approx C h$. Give an expression for $C$.
Exercise 5.2 The the function $f(x)$ be known at the points $\left\{x_{k}\right\}$ 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

$$
I=\int_{0}^{1} f(x) d x
$$

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 $\varepsilon$. Derive a bound for the resulting error when we want to compute an integral

$$
I=\int_{a}^{b} f(x) d x
$$

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 $f^{\prime}(2)$ we can use the formula

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

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 $F_{1}(h)=a+b h^{p_{1}}+c h^{p_{2}}$, for some constants $a, b$ and $c$ and positive integers $p_{1}$ and $p_{2}$ such that $p_{1}<p_{2}$. Show that

$$
F_{1}(h)+\frac{F_{1}(h)-F_{1}(q h)}{q^{p_{1}}-1}=a+\mathcal{O}\left(h^{p_{2}}\right) .
$$

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

| h | 0.1 | 0.05 |
| :---: | :---: | :---: |
| $\mathrm{~T}(\mathrm{~h})$ | 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 $R_{T}=C h^{p}$, 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

$$
\begin{array}{c|cccc}
\mathrm{h} & 0.4 & 0.2 & 0.1 & 0.05 \\
\hline \mathrm{~T}(\mathrm{~h}) & 3.100 & 2.701 & 2.604 & 2.578
\end{array}
$$

Use the table to determine $p$.
Exercise 5.8 Compute an approximation of $\int_{0}^{1} f(x) d x$ 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 $R_{T} \approx C h^{p}$. Use the table to compute both $C$ and $p$. Also determine the largest possible step size that can be used if we require that $\left|R_{T}\right| \leq 10^{-8}$.

Exercise 5.10 We want to compute an approximation of

$$
\int_{1}^{4} f(x) d x, \quad \text { where } f(x)=\cos \left(x^{2}\right) \sqrt{4-x}
$$

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

$$
I=\int_{0}^{1} \mathrm{e}^{1-x^{2}} d x
$$

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

| h | $1 / 2$ | $1 / 4$ | $1 / 8$ |
| :---: | :---: | :---: | :---: |
| $\mathrm{~S}(\mathrm{~h})$ | 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 $\left|R_{T}\right|<10^{-6}$. What step size would be required? Motivate your answer.

## 6 Differential Equations

Exercise 6.1 Consider the problem

$$
\left\{\begin{array}{l}
y^{\prime}=y^{2}-t . \\
y(0)=1
\end{array}\right.
$$

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 $y^{\prime}=t^{2}-y$, 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 $y^{\prime}=-100 y+\sin ^{2}(t), y(0)=0$. 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 $y^{\prime}=-25 y+t+0.04, y(0)=1$, 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

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

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 $y_{k} \approx y\left(t_{k}\right)$ by the steps

$$
\begin{aligned}
& k_{1}=h f\left(t_{k}, y_{k}\right), \\
& k_{2}=h f\left(t_{k}+h, y_{k}+k_{1}\right), \\
& y_{k+1}=y_{k}+\frac{1}{2}\left(k_{1}+k_{2}\right)
\end{aligned}
$$

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

Exercise 6.7 Rewrite the initial value problem $y^{\prime \prime}+3 y^{\prime}-t y+1=0, y(0)=1, y^{\prime}(0)=0$, as a system of first order.

Exercise 6.8 We have the second order equation

$$
y^{\prime \prime}=2 y\left(1+y^{2}\right), \quad y(0)=1, y^{\prime}(0)=2 .
$$

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

$$
y^{\prime \prime}+e\left(y^{2}-a\right) y^{\prime}+y=0, \quad y(0)=a, y^{\prime}(0)=b .
$$

Rewrite the equation as a system of first order equations.

## 7 Basic matrix operations and linear systems of equations

Exercise 7.1 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 $z=(A+I) B x+y$, where $x$ and $y$ are vectors.

Exercise 7.2 Suppose $A, B$, and $C$ are matrices and $b$ is a vector. How would you implement the formula

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

without computing any matrix inverse? Aim for as few arithmetic operations as possible.
Exercise 7.3 Suppose we want to solve the upper triangular system $R x=y$ by backwards substitution. Clearly show how many floating point operations are needed.

Exercise 7.4 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 $L U$-decomposition followed by two triangular systems is used. Is the algorithms computational complexity as expected?

Exercise 7.5 The unit circle can be defined as all points $x=\left(x_{1}, x_{2}\right)$ such that $\|x\|=1$. Draw the unit circle when $\|\cdot\|_{1},\|\cdot\|_{2}$ and $\|\cdot\|_{\infty}$ is used.

Exercise 7.6 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 $\|A B\| \leq\|A\|\|B\|$.

Exercise 7.7 Prove the inequality $\|x\|_{\infty} \leq\|x\|_{2} \leq \sqrt{n}\|x\|_{\infty}$.
Exercise 7.8 Let $u \in \mathbb{R}^{m}$ and $v \in \mathbb{R}^{n}$. Show that

$$
\left\|u v^{T}\right\|_{2}=\|u\|_{2}\|v\|_{2} .
$$

Exercise 7.9 Prove that $\|I\|=1$ and $\|A\|\left\|A^{-1}\right\| \geq 1$ for all matrix norms induced by a vector norm.

Exercise 7.10 Let $x=(1,-3,7)^{T}$. Compute $\|x\|_{1},\|x\|_{2}$, and $\|x\|_{\infty}$.
Exercise 7.11 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 7.12 Let

$$
A=\left(\begin{array}{ccc}
1.2 & 0.3 & -2.7 \\
3.1 & -0.5 & 3.2 \\
1.6 & -0.8 & -2.3
\end{array}\right) .
$$

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

Exercise 7.13 Formulate the following system of equations on matrix form $A x=b$ :

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

Exercise 7.14 Find a permutation matrix $P$ such that

$$
P\left(\begin{array}{l}
x_{1} \\
x_{2} \\
x_{3} \\
x_{4}
\end{array}\right)=\left(\begin{array}{l}
x_{1} \\
x_{4} \\
x_{2} \\
x_{3}
\end{array}\right)
$$

Exercise 7.15 Find a permutation matrix $P$ such that

$$
P\left(\begin{array}{cccc}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
9 & 0 & 1 & 2 \\
3 & 4 & 5 & 6
\end{array}\right)=\left(\begin{array}{cccc}
3 & 4 & 5 & 6 \\
9 & 0 & 1 & 2 \\
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8
\end{array}\right)
$$

Exercise 7.16 Find a Gauss transformation $M$ such that

$$
M\left(\begin{array}{c}
2 \\
3 \\
0.6 \\
-1.8
\end{array}\right)=\left(\begin{array}{l}
2 \\
3 \\
0 \\
0
\end{array}\right)
$$

Exercise 7.17 Suppose we have a linear system $A x=b$ where

$$
A=\left(\begin{array}{ccc}
2 & 1 & -2 \\
-1 & 0 & 3 \\
1 & 2 & -1
\end{array}\right) \text { and } b=\left(\begin{array}{c}
6 \\
1 \\
-3
\end{array}\right) .
$$

During the first step of Gaussian elimination we multiply the system with a matrix $M_{1}$ such that the new system $M_{1} A x=M_{1} b$ is

$$
\left(\begin{array}{ccc}
2 & 1 & -2 \\
0 & 0.5 & 2 \\
0 & 1.5 & 0
\end{array}\right) x=\left(\begin{array}{c}
6 \\
4 \\
-6
\end{array}\right) .
$$

Give the Gausstransformation $M_{1}$.
Exercise 7.18 A Gauss transformation $M_{1}$ that eliminates the non-zeros from the first column of a matrix has the form

$$
M_{1}=\left(\begin{array}{ccc}
1 & 0 & 0 \\
-m_{21} & 1 & 0 \\
-m_{31} & 0 & 1
\end{array}\right) .
$$

Show that the matrix can be written as $M_{1}=I-m e_{1}^{T}$. Also give the elements of the vector $m$ in terms of the elements of the matrix $A=\left(a_{i j}\right)$ and show that $M_{1}^{-1}=I+m e_{1}^{T}$.

Exercise 7.19 The Gauss transformation $M_{2}$ that is used during the second step of Gaussian elimination has the structure

$$
M_{2}=\left(\begin{array}{cccc}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & -m_{32} & 1 & 0 \\
0 & -m_{42} & 0 & 1
\end{array}\right) .
$$

Show that $M_{2}$ can be written as $M_{2}=I-m e_{2}^{T}$ where $m=\left(0,0, m_{32}, m_{42}\right)$. Also show that $M_{2}^{-1}=I+m e_{2}^{T}$.

Exercise 7.20 Let

$$
A=\left(\begin{array}{ccc}
2 & 1 & -2 \\
-3 & 0.5 & -2 \\
1 & 2.5 & 0
\end{array}\right)
$$

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

Exercise 7.21 After one step of Guassian elimination we have

$$
\left(\begin{array}{ccc}
2 & 1 & -2 \\
0 & -0.5 & 1.6 \\
0 & 1.7 & 0.3
\end{array}\right) .
$$

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 7.22 We need to solve a system of equations $A x=b$, where

$$
A=\left(\begin{array}{ccc}
1 & 1.2 & -2.3 \\
2 & 0 & -1 \\
-1 & 1.5 & 2.1
\end{array}\right) \text { and } b=\left(\begin{array}{c}
-0.5 \\
1.2 \\
1.27
\end{array}\right) .
$$

We compute the decomposition $P A=L U$ where

$$
L=\left(\begin{array}{ccc}
1 & 0 & 0 \\
-0.5 & 1 & 0 \\
0.5 & 0.8 & 1
\end{array}\right) \quad U=\left(\begin{array}{ccc}
2 & 0 & -1 \\
0 & 1.5 & 1.6 \\
0 & 0 & -3.08
\end{array}\right) \quad P=\left(\begin{array}{lll}
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{array}\right) .
$$

Take advantage of the decomposition to compute the solution $x$.
Exercise 7.23 A computer program has computed the decomposition $P A=L U$ and the output is

$$
L=\left(\begin{array}{ccc}
1 & 0 & 0 \\
-0.7 & 1 & 0 \\
0.3 & 1.8 & 1
\end{array}\right) \quad U=\left(\begin{array}{ccc}
1.7 & -2.3 & -1.4 \\
0 & 1.2 & -0.5 \\
0 & 0 & 3.1
\end{array}\right) \quad P=\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0
\end{array}\right) .
$$

Determine if pivoting was used correctly during the computations.

Exercise 7.24 Let a matrix $A$ and a vector $\bar{b}$ be given as

$$
A=\left(\begin{array}{ccc}
1 & -2 & 0 \\
2 & 0 & 1 \\
-1 & 0 & 1
\end{array}\right) \quad \bar{b}=\left(\begin{array}{c}
-1.03 \\
1.34 \\
0.78
\end{array}\right)
$$

where the elements of $\bar{b}$ are correctly rounded. The system of equations $A x=b$ 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 $\left\|A^{-1}\right\|_{\infty}=1$.
Exercise 7.25 Let

$$
A=\left(\begin{array}{ccc}
3.1 & -1.2 & 2.6 \\
1.5 & -0.7 & 3.6 \\
-4.1 & 1.1 & 0.7
\end{array}\right)
$$

If we solve the system $A x=b_{1}$, 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 $\left\|A^{-1}\right\|_{\infty}=17.9052$.
Exercise 7.26 Let $P A=L U$ be the $L U$ decomposition. Prove the formula

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

What is $k$ here?
Exercise 7.27 Let $r=b-A \widehat{x}$ be the residual for an approximate solution to the linear system $A x=b$. Prove the formula:

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

Exercise 7.28 Let

$$
A=\left(\begin{array}{ccc}
2 & 1 & -1 \\
1 & 3 & -2 \\
2 & 0 & 1
\end{array}\right) \quad \text { och }=\left(\begin{array}{c}
1.12 \\
-1.07 \\
2.34
\end{array}\right) .
$$

and the decomposition $P A=L U$, where

$$
L=\left(\begin{array}{ccc}
1 & 0 & 0 \\
0.5 & 1 & 0 \\
1 & -0.4 & 1
\end{array}\right) \quad U=\left(\begin{array}{ccc}
2 & 1 & -1 \\
0 & 2.5 & -1.5 \\
0 & 0 & 1.4
\end{array}\right) \quad P=\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right) .
$$

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

## 8 Least Squares Problems and Orthogonal Decompositions

Exercise 8.1 Suppose we have a series of measurements $\left(x_{i}, y_{i}\right), i=1,2, \ldots, m$, 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 $A x=b$.

Exercise 8.2 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 $\left(t_{i}, u\left(t_{i}\right)\right), i=1,2, \ldots, m$, and want to find $A$ and $\phi$ 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 8.3 The time needed for a certain algorithm to complete can be described by the formula $t \approx C n^{p}$, 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 $A x=b$.

Exercise 8.4 Suppose $Q$ is an orthogonal matrix. Show that $\|Q x\|_{2}=\|x\|_{2}$, for all vectors $x$, and thus $\|Q\|_{2}=1$.

Exercise 8.5 If $A$ is both an ortgohonal matrix and an orthogonal projection. What can you conclude about $A$ ?

Exercise 8.6 Suppose $A \in \mathbb{R}^{m \times n}, m>n$, and that we have the reduced $Q R$ decomposition $A=Q_{1} R$. Show how the decomposition can be used to find the vector $x$ that minimize $\|A x-b\|_{2}$.

Exercise 8.7 Suppose $A \in \mathbb{R}^{m \times n}, m>n$, and that we have the $Q R$ decomposition

$$
A=Q\binom{R}{0}=Q_{1} R
$$

Show that the linear system $A x=b$ has an exact solution if $b=Q_{1} Q_{1}^{T} b$.
Exercise 8.8 Consider the vector $a$ as an $n \times 1$ matrix. Write out its reduced $Q R$ decomposition explicitly. Also write down a formula for the solution of the least squares problem $a x \approx b$, where $b$ is a given $n \times 1$ vector.

Exercise 8.9 We are intressted in the least squares problem $\min \|A x-b\|_{2}$. Suppose $A=Q_{1} R$ is the reduced $Q R$ decomposition. Use $Q_{1}$ to give a formula for a orthogonal projection $P$, such that $P b=r=b-A x$, where $x$ is the least squares solution.

Exercise 8.10 Compute the reduced $Q R$ factorization of the matrix

$$
A=\left(\begin{array}{rr}
0 & \sqrt{2} \\
-1 & 1 \\
1 & 1
\end{array}\right)
$$

Exercise 8.11 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,

$$
\min _{x}\|A x-b\|_{W} .
$$

Hint Use the Cholesky factorization $W=R^{T} R$.
Exercise 8.12 Show that $\|A\|_{2}=\sigma_{1}$ and if $A^{-1}$ exists then $\left\|A^{-1}\right\|_{2}=1 / \sigma_{n}$.
Exercise 8.13 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 $\operatorname{null}\left(A^{T}\right)$.

Exercise 8.14 Show that if $A \in \mathbb{R}^{m \times n}$ has rank $n$, then $\left\|A\left(A^{T} A\right)^{-1} A^{T}\right\|_{2}=1$.
Exercise 8.15 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

$$
\min _{x}\|B x\|_{2}, \text { subject to }\|x\|_{2}=1 .
$$

Exercise 8.16 Suppose $A \in \mathbb{R}^{m \times n}, m>n, \operatorname{rank}(A)=n$, 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 Range $(A)$ and null $(A)$. What are the dimension of the range and null space respectively.

Exercise 8.17 Let $A \in \mathbb{R}^{m \times n}, m>n$, and $\operatorname{rank}(A)=n$. Demonstrate how the decomposition $A=U \Sigma V^{T}$ can be used for solving the least squares problem

$$
\min _{x \in \mathbb{R}^{n}}\|A x-b\|_{2}
$$

Give formulas for both the solution $x$ and the redisual $r=b-A x$.
Exercise 8.18 Let $A \in \mathbb{R}^{m \times n}, m<n$, and $\operatorname{rank}(A)=m$. 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 $A x=b$. Is the solution unique?

Exercise 8.19 Suppose we want to find the solution to a linear system $A x=b$, where $\operatorname{rank}(A)=$ $k<n$ so that the solution $x$ is not unique. Demonstrate how the solution $x$ can be split into two parts,

$$
x=x_{1}+x_{2}, \quad x_{1} \in \operatorname{null}(A)^{\perp}, \quad \text { and, } \quad x_{2} \in \operatorname{null}(A),
$$

and how the SVD of $A$ can be used to write expressions for the solution components $x_{1}$ and $x_{2}$.
Exercise 8.20 Consider the Least Squares problem with linear constraints,

$$
\min \|A x-b\|_{2}, \quad \text { for all } x \in \mathbb{R}^{n} \text { such that } B x=0
$$

where $A$ is $m \times n, m>n$, and $B$ is $n \times n$.
a) Suppose $\operatorname{rank}(B)=n$. What is the solution of the least squares problem?
b) Suppose $\operatorname{rank}(B)=k<n$. Show how the SVD can be used to derive a formula for the solution of the least squares problem.

Exercise 8.21 Let $A \in \mathbb{R}^{m \times n}$, where $m \gg n$, have full column rank. Use the decomposition $A=U \Sigma V^{T}$ to develop a criteria that ensures that the linear system $A x=b$ has a solution. Try and make the criteria as inexpensive as possible to check.

Exercise 8.22 Tikhonov regularization means replacing an ill-conditioned linear system $A x=b$ by the more stable problem,

$$
\min _{x}\|A x-b\|_{2}^{2}+\lambda^{2}\|x\|_{2}^{2}
$$

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

$$
\left(A^{T} A+\lambda^{2} I\right) x=A^{T} b .
$$

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

## 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 $\pi$ 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

$$
\begin{array}{|l|l|l|}
\hline 0 & 10000010 & 00101000 \ldots 00 \\
\text { and }
\end{array} \begin{array}{|l|l|l|}
\hline 1 & 1000000 & 01010100 \ldots 00 \\
\hline
\end{array}
$$

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, $\beta$ 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 $m= \pm d_{0} \cdot d_{1} d_{2} \ldots d_{t}$. 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

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

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

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

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 $\bar{f}=1.8\left(\left|R_{B}\right|=0\right)$. 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

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

Thus $f=1.80 \pm 0.5 \cdot 10^{-2}$.
Exercise 2.2 The error propagation formula gives

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

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

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

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

The total error is $\left|R_{T O T}\right| \leq 3.3+0.3<4$. Thus $A=1774 \pm 4$.
Exercise 2.4 The approximate value is $\bar{y}=1.54 \mathrm{e}^{3.17}=36.7,\left|R_{B}\right| \leq 0.5 \cdot 10^{-1}$. The error propagation formula gives

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

The total error is $\left|R_{T O T}\right| \leq 2.55+0.5 \cdot 10^{-1}<2.6$. Thus $y=36.7 \pm 2.6$.
In hindsight it would probably have been better to round to 37 and use $\left|R_{B}\right|=|37-36.6635 \ldots|<$ 0.34 to obtain the answer $y=37 \pm 3$. Either works fine.

Exercise 2.5 Rewrite the expression to read

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

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

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

The total error is $\left|R_{T O T}\right| \leq 0.05+0.57<0.7$. 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 $\mu$. The error propagation formula gives

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

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

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

for small $x$. We obtain

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

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

$$
\begin{gathered}
|\Delta f| \lesssim\left|\frac{\partial f}{\partial c}\right||\Delta c|+\left|\frac{\partial f}{\partial s}\right||\Delta s|+\left|\frac{\partial f}{\partial d}\right||\Delta d|+\left|\frac{\partial f}{\partial e}\right||\Delta e|=\left|\frac{1}{s}\right||\Delta c|+\left|\frac{1-c}{s^{2}}\right||\Delta s|+\left|\frac{1}{s}\right||\Delta d|+|\Delta e| \leq \\
\mu\left(\left|\frac{c}{s}\right|+\left|\frac{1-c}{s}\right|+\left|\frac{d}{s}\right|+|e|\right) \approx \frac{\mu}{x}
\end{gathered}
$$

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
$|\Delta f| \lesssim\left|\frac{s}{(1+c)^{2}}\right||\Delta c|+\left|\frac{1}{1+c}\right||\Delta s|+\left|\frac{s}{d^{2}}\right||\Delta d|+|\Delta e| \leq \mu\left(\left|\frac{c s}{(1+c)^{2}}\right|+\left|\frac{s}{1+c}\right|+\left|\frac{s}{d}\right|+|e|\right) \approx 1.75 x \mu$.
These are the absolute errors. Insert $x=1.111 \cdot 10^{-8}$ in and $\left|\Delta f_{1}\right| \leq 10^{-8}$ 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

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

The error propagation formula gives us

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

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

where we have used $\mathrm{e}^{x} \approx 1, f(x)=c \approx 1$ 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 $2 \mu$ (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 $\mu$. The error propagation formula gives

$$
\begin{aligned}
& |\Delta f| \lesssim\left|\frac{\partial f}{\partial a}\right||\Delta a|+\left|\frac{\partial f}{\partial b}\right||\Delta b|+\left|\frac{\partial f}{\partial c}\right||\Delta c|+\left|\frac{\partial f}{\partial d}\right||\Delta d|= \\
& \left|-\frac{1}{b}\right||\Delta a|+\left|-\frac{c}{b^{2}}\right||\Delta b|+\left|\frac{1}{b}\right||\Delta c|+|1||\Delta d| \leq \mu\left(\left|\frac{a}{b}\right|+\left|\frac{c}{b}\right|\left|\frac{c}{b}\right|+|d|\right) \approx \mu\left(\left|\frac{1}{x^{2}}\right|+\frac{3}{6}\right) .
\end{aligned}
$$

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

Exercise 2.10 The computational order is

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

The error propagation formula gives us

$$
\begin{gathered}
|\Delta f| \lesssim\left|\frac{\partial f}{\partial a}\right||\Delta a|+\left|\frac{\partial f}{\partial b}\right||\Delta b|+\left|\frac{\partial f}{\partial c} \| \Delta c\right|=|2 x||\Delta a|+|1||\Delta b|+|1||\Delta c| \lesssim \\
\mu(|2 x a|+|b|+|c|) \approx \mu(|2 x|+|2 x|+1) \approx \mu
\end{gathered}
$$

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 $\mu$ (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

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

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

$$
|\Delta f| \lesssim\left|f^{\prime \prime}(\bar{x})\right| \frac{|\Delta x|^{2}}{2}
$$

## 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 $\phi^{\prime}(x)=\frac{1}{2} \mathrm{e}^{x}$ and $\left|\phi^{\prime}(0.3)\right| \approx 0.68<1$.
Exercise 3.3 With $\bar{x}=1.7$ we get $f(\bar{x})=-0.387$. Also $f^{\prime}(x)=3 x^{2}+1$ so $f^{\prime}(\xi) \approx f^{\prime}(1.7)=9.67$. Thus

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

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

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

Exercise 3.5 Let $f(x)=5\left(1-\mathrm{e}^{-x}\right)-x$ 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

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

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 $\left|\phi^{\prime}\left(x^{*}\right)\right|$. For the two methods, and $x^{*} \approx 1.9$, we get $\phi_{1}^{\prime}(1.9) \approx 0.647$ and $\phi_{2}^{\prime}(1.9) \approx 0.177$. Thus the second method has the fastest convergence. Also note that a fixed point satisfies $x^{*}=\frac{x^{*}}{2}+\sin \left(x^{*}\right)$ which can be written as $x^{*}=\sin \left(x^{*}\right)$. 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

$$
\left|x_{n}-x^{*}\right|=\left|\phi\left(x_{n-1}\right)-\phi\left(x^{*}\right)\right| \leq\left|\phi^{\prime}\left(\xi_{n}\right)\right|\left|x_{n-1}-x^{*}\right| \leq C\left|x_{n-1}-x^{*}\right|,
$$

where the constant $C$ satisfies $\left|\phi^{\prime}\left(\xi_{n}\right)\right| \leq C$. In this case $\phi(x)=\frac{1}{2} \cos (x)$ and $\phi^{\prime}(x)=-\frac{1}{2} \sin (x)$. Thus $\left|\phi^{\prime}(x)\right| \leq \frac{1}{2}$. Thus we can pick $C=\frac{1}{2}<1$ and we have proved that the error $\left|x_{n}-x^{*}\right|$ 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 $f^{\prime}(x)=$ $-2 \sin (x)-4$ and use the estimate $=|-4.87026 \ldots|>4.8=M$. Thus

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

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

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

where $f^{\prime}(x)=1+3 \mathrm{e}^{-x}$.
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 $3 x^{*}=2 x^{*}+3 \mathrm{e}^{-x^{*}}$ or $x^{*}=3 \mathrm{e}^{-x^{*}}$ or $f\left(x^{*}\right)=0$. 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

$$
\begin{gathered}
\phi_{1}(x)=3 \mathrm{e}^{-x}, \text { so } \phi_{1}^{\prime}(1.05)=1.05 \text { and divergence, } \\
\phi_{2}(x)=\left(2 x+3 \mathrm{e}^{-x}\right) / 3, \text { so } \phi_{2}^{\prime}(1.05)=-0.32 \text { and convergence, }
\end{gathered}
$$

and

$$
\phi_{4}(x)=\left(x+3 \mathrm{e}^{-x}\right) / 2, \text { so } \phi_{4}^{\prime}(1.05)=-0.025, \text { and again convergence },
$$

Thus the method (iv) should converge the fastest to $x^{*}$.
Finally the error in step $k$ is $\left|x_{k}-x^{*}\right| \approx\left|\phi_{4}^{\prime}(1.05)\right|^{k}\left|1.05-x^{*}\right| \approx(0.025)^{k} 10^{-4}$. 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

$$
x_{k+1}=x_{k}-\frac{f\left(x_{k}\right)}{f^{\prime}\left(x_{k}\right)}=x_{k}-\frac{x_{k}^{2}}{2 x_{k}}=\frac{1}{2} x_{k},
$$

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

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

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

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

$$
\phi\left(x_{k}\right)=\phi\left(x^{*}\right)+\phi^{\prime}\left(x^{*}\right)\left(x_{k}-x^{*}\right)+\frac{\phi^{\prime \prime}(\xi)}{2}\left(x_{k}-x^{*}\right)^{2}, \xi \in\left(x_{k}, x^{*}\right) .
$$

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

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

which shows that the convergence is quadratic.

Exercise 3.11 We use a Taylor series expansion

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

and since $x^{*}$ is a double root we have $f\left(x^{*}\right)=f^{\prime}\left(x^{*}\right)=0$ and thus

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

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 $f^{\prime \prime}(\bar{x})=-3.0127$. Thus

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

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

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

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

$$
\lim _{k \rightarrow \infty} \frac{\left|x_{k}-x^{*}\right|}{\left|x_{k-1}-x^{*}\right|^{p}}=C<\infty .
$$

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 $\left|x_{0}-x^{*}\right| \approx 4.8 \cdot 10^{-1},\left|x_{1}-x^{*}\right| \approx 1.4 \cdot 10^{-1},\left|x_{2}-x^{*}\right| \approx$ $1.4 \cdot 10^{-2}$, and $\left|x_{3}-x^{*}\right| \approx 1.5 \cdot 10^{-4}$. Since $\left(\left|x_{1}-x^{*}\right|\right)^{2} \approx\left(1.4 \cdot 10^{-1}\right)^{2} \approx 2 \cdot 10^{-2} \approx\left|x_{2}-x^{*}\right|$ and $\left(\left|x_{2}-x^{*}\right|\right)^{2} \approx\left(1.4 \cdot 10^{-2}\right)^{2} \approx\left|x_{3}-x^{*}\right|$ 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

$$
\lim _{k \rightarrow \infty} \frac{\left|x_{k}-x^{*}\right|}{\left|x_{k-1}-x^{*}\right|^{p}}=C<\infty
$$

This means that $\left|x_{k}-x^{*}\right| \approx C\left|x_{k-1}-x^{*}\right|^{p}$, 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

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

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 $\phi^{\prime \prime}(x)=-\frac{a}{x^{3}}$ and therefore $\phi^{\prime \prime}(\sqrt{a})=1 / \sqrt{a} \leq 1$. Thus

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

If $x_{0}=1.5$ the maximum error is $\left|x_{0}-\sqrt{a}\right| \leq 0.5$. We get $\left|x_{1}-\sqrt{a}\right| \leq 0.125,\left|x_{2}-\sqrt{a}\right| \leq 0.078$, $\left|x_{3}-\sqrt{a}\right| \leq 3.1 \cdot 10^{-5},\left|x_{4}-\sqrt{a}\right| \leq 4.7 \cdot 10^{-10}$, and $\left|x_{5}-\sqrt{a}\right| \leq 1.1 \cdot 10^{-19}$. 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 $\left|x_{0}-\sqrt{a}\right|<$ 0.125 . Then the new $x_{0}$ is the same as $x_{1}$ above. This means dividing the interval $[1,4]$ into $n$ intervals $\left[a_{0}, a_{1}\right),\left[a_{1}, a_{2}\right), \ldots,\left[a_{n-2}, a_{n-1}\right)$. 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 $\left(\sqrt{a_{k}}+\sqrt{a_{k+1}}\right) / 2$ and the initial error is $\left(\sqrt{a_{+1}}-\sqrt{a_{k}}\right) / 2$. 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 $\left|x_{4}-\sqrt{a}\right| \leq 6.5 \cdot 10^{-18}$ 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
$p(0.9)=c_{0}=f(0.9)=0.4710$, and $p(1.1)=c_{0}+c_{1}(1.1-0.9)=f(1.1)=0.2452$ 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 $f(1.03) \approx p(1.03)=$ $0.3242,\left|R_{B}\right| \leq 0.5 \cdot 10^{-4}$, and $\left|R_{T}(1.03)\right| \leq 0.033$. The errors in the table also gives an error $\left|R_{X F}\right| \leq 0.5 \cdot 10^{-4}$ 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 $R_{T O T} \leq 10^{-5}$. If the values in the table is stored with 6 correct digits then the resulting error is $R_{X F} \leq 0.5 \cdot 10^{-6}$. 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

$$
R_{T} \leq \frac{h^{2}}{8} \max _{x_{i} \leq \xi \leq x_{i+1}}\left|f^{\prime \prime}(\xi)\right|
$$

and since $f(x)=\log (x)$ we find that $\left|f^{\prime \prime}(x)\right|=\left|-x^{-2}\right| \leq 1$ 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

$$
f(x)-p(x)=\frac{1}{2} f^{\prime \prime}(\eta(x))\left(x-x_{1}\right)\left(x-x_{2}\right),
$$

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

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

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 $p_{1}(x)=1.8+2 x$. 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 $\left|R_{B}\right|=0$, and $\left|R_{T}(0.4)\right| \leq 0.024$. The errors in the table also gives an error $\left|R_{X F}\right| \leq 0.5 \cdot 10^{-2}$ 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 $\left|R_{B}\right| \leq 0.5 \cdot 10^{-3}$ (not asked for in the exercise), and with the truncation error $\left|R_{T}\right| \leq-0.014$.

Exercise 4.7 The basis function satisfies $\ell_{2}\left(x_{2}\right)=1$ and $\ell_{2}\left(x_{i}\right)=0, i \neq 2$. Thus

$$
\ell_{2}(x)=\frac{\left(x-x_{1}\right)\left(x-x_{3}\right)\left(x-x_{4}\right)}{\left(x_{2}-x_{1}\right)\left(x_{2}-x_{3}\right)\left(x_{2}-x_{4}\right)}
$$

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 $p^{\prime}(x)=c_{1}+2 c_{2} x+3 c_{3} x^{2}$ so we also obtain $p^{\prime}(0)=c_{1}=1$ and $p^{\prime}(1)=c_{1}+2 c_{2}+3 c_{3}=1$. Thus the system of equations is

$$
\left(\begin{array}{llll}
1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 \\
0 & 1 & 2 & 3
\end{array}\right)\left(\begin{array}{l}
c_{0} \\
c_{1} \\
c_{2} \\
c_{3}
\end{array}\right)=\left(\begin{array}{l}
0 \\
0 \\
1 \\
1
\end{array}\right)
$$

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 $p(x)=x-3 x^{2}+2 x^{3}$.

Exercise 4.10 The conditions for $s(x)$ to be a cubic spline are $(i)$ on each sub interval $\left[x_{i}, x_{i+1}\right]$ the spline $s(x)$ should be given by a cubic polynomial, and (ii) $s(x), s^{\prime}(x)$ and $s^{\prime \prime}(x)$ should be continuous on the whole interval $\left[x_{1}, x_{n}\right]$. Also the (iii) the interpolation conditions $s\left(x_{i}\right)=f\left(x_{i}\right)$ 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 $s(x), s^{\prime}(x)$ and $s^{\prime \prime}(x)$ are continuous at $x=1$.

Exercise 4.12 The conditions that has to be satisfied are $s(1)=a+1=b+c, s^{\prime}(1)=a=3 b+2 c$ and $s^{\prime \prime}(1)=0=6 b+2 c$. 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 $s_{1}^{\prime}(1)=2.5$ and use $s_{2}^{\prime}(1)=b=2.5$. 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 $s_{1}(1)=0.6=s_{2}(1), s_{1}^{\prime}(1)=-0.6=s_{2}^{\prime}(1)$ and finally $s_{1}^{\prime \prime}(1)=-0.8=s_{2}^{\prime \prime}(1)$. Thus $s(x)$ is a cubic spline. We also compute $s_{1}^{\prime \prime}(0)=-0.8$ 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 $s_{k-1}\left(x_{k}\right)=s_{k}\left(x_{k}\right), s_{k-1}^{\prime}\left(x_{k}\right)=s_{k}^{\prime}\left(x_{k}\right)$ and $s_{k-1}^{\prime \prime}\left(x_{k}\right)=s_{k}^{\prime \prime}\left(x_{k}\right)$, for $k=1,2, \ldots, N-1$. In addition we need the interpolation conditions $s_{k}\left(x_{k}\right)=f\left(x_{k}\right)$ and $s_{k}\left(x_{k+1}\right)=f\left(x_{k+1}\right)$, for $k=0,1, \ldots, N-1$. A sketch of the case $N=3$ is given below


Exercise 4.17 Since the error satisfies $E(N) \approx C h^{p}=C^{\prime}(1 / N)^{p}$, where $C^{\prime}=C(b-a)^{p}$ is a constant. Thus $E(N) / E(2 N)=2^{p}$ 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

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

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

$$
p^{\prime}(t)=-2(1-t) P_{1}+2(1-2 t) P_{2}+2 t P_{3}, \text { and } p^{\prime}(0)=2\left(P_{2}-P_{1}\right) .
$$

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
$p^{\prime}(t)=-3(1-t)^{2} P_{1}+3\left(-2(1-t) t+(1-t)^{2}\right) P_{2}+3\left(2(1-t) t-t^{2}\right) P_{3}+3 t^{2} P_{4}$, and $p^{\prime}(0)=3\left(P_{2}-P_{1}\right)$.
For $\mathbf{b}$ ) the definition of the convex hull is the set of all convex linear combinations $\alpha_{1} P_{1}+\alpha_{2} P_{2}+$ $\alpha_{3} P_{3}+\alpha_{4} P_{4}$, 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 $\mathbf{c}$ ) we compute $p(1 / 2)$ by inserting $t=1 / 2$ into the expression (note that $t=0.5$ means also $1-t=0.5$ )

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

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

$$
P_{3}=\binom{2}{3}-\alpha\binom{1}{-2} \text { and } P_{5}=\binom{2}{3}+\alpha\binom{1}{-2},
$$

where $\alpha$ is a positive number. There is no unique solution to this problem.

## 5 Integration, Differentiation and Extrapolation

Exercise 5.1 Taylors formula gives

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

Thus

$$
\frac{1}{h}(f(x+h)-f(x))=f^{\prime}(x)+\frac{1}{2} f^{\prime \prime}(x) h+\frac{1}{6} f^{(3)}(x) h^{2}+\ldots=f^{\prime}(x)+C h+\mathcal{O}\left(h^{2}\right), \quad C=\frac{1}{2} f^{\prime \prime}(x) .
$$

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 $R_{T}=c h^{2}$ 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 $\left|R_{X F}\right| \leq(1-0) 0.5 \cdot 10^{-3}=0.5 \cdot 10^{-3}$. If we round the result to 3 correct digits we get $I=0.934 \pm\left(0.5 \cdot 10^{-3}+4.7 \cdot 10^{-3}+0.5 \cdot 10^{-3}\right)=0.934 \pm 6 \cdot 10^{-3}$.

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

$$
\left|R_{X F}\right|=\left|\int_{a}^{b} f(x) d x-\int_{a}^{b} \bar{f}(x) d x\right| \leq \int_{a}^{b}|f(x)-\bar{f}(x)| d x \leq \int_{a}^{b} \varepsilon|f(x)| d x \leq\left(\int_{a}^{b}|f(x)| d x\right) \varepsilon .
$$

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

$$
\left|\int_{a}^{b} f(x) d x\right| \ll \int_{a}^{b}|f(x)| d x,
$$

and the result may have a large relative error.
Exercise 5.4 We denote the error by $\epsilon_{h} \approx C h^{p}$. Then

$$
\frac{\epsilon_{h_{1}}}{\epsilon_{h_{2}}} \approx \frac{C h_{1}^{p}}{C h_{2}^{p}}=\left(\frac{h_{1}}{h_{2}}\right)^{p} .
$$

Insert numbers from the table we obtain

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

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

Exercise 5.5 We see that

$$
\begin{aligned}
& F_{1}(h)+\frac{F_{1}(h)-F_{1}(q h)}{q^{p_{1}}-1}=a+b h^{p_{1}}+c h^{p_{2}}+\frac{a+b h^{p_{1}}+c h^{p_{2}}-\left(a+b q^{p_{1}} h^{p_{1}}+c q^{p_{2}} h^{p_{2}}\right)}{q^{p_{1}}-1}= \\
& =a+b h^{p_{1}}+c h^{p_{2}}+\frac{b\left(1-q^{p_{1}}\right) h^{p_{1}}+c\left(1-q^{p_{2}}\right) h^{p_{2}}}{q^{p_{1}}-1}=a+c\left(1-\frac{1-q^{p_{2}}}{1-q^{p_{1}}}\right) h^{p_{2}}=a+\mathcal{O}\left(h^{p_{2}}\right)
\end{aligned}
$$

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

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

where $I$ is the exact value of the integral and $R_{T}$ is the truncation error for $h=0.05$. Thus

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

Exercise 5.7 With $T(h)=T_{0}+C h^{p}$ we form the quotient

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

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

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 $R_{T} \approx-0.4 h^{1}$ 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 $f^{\prime \prime}(x)$ on the intervall $[1,4]$ to have a truncation error $R_{T} \approx c h^{2}$. In this case the first derivative is

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

which tends to $-\infty$ as $x \rightarrow 4$. 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

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

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

$$
S(1 / 4)-S(1 / 8)=2.03208-2.03021 \approx c\left(2^{4}-1\right)(1 / 8)^{4}
$$

to obtain $c=0.51$. For $\mathbf{b}$ ) we note that $\left|R_{T}\right| \approx 0.51 h^{4}$ is smaller than $10^{-6}$ if $h \leq\left(10^{-6} / 0.51\right)^{1 / 4} \approx$ 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

$$
y(0.5) \approx y_{1}=y_{0}+h f\left(t_{0}, y_{0}\right)=y_{0}+h\left(y_{0}^{2}-t_{0}\right)=1+0.5\left(1^{2}-0\right)=1.5 .
$$

Also, since the truncation error for Euler's method is $R_{T} \approx c h$ we find $y(0.5 ; 5 h)-y(0.5 ; h) \approx y(0.5)+$ $c 5 h-y(0.5)-c h=4 c h$. Thus, with $h=0.1$ we get $R_{T} \approx c 5 h \approx 5(y(0.5 ; 5 h)-y(0.5 ; h)) / 4 \approx 0.021$.

Exercise 6.2 Using Euler's method we find that, using $h=0.2$, we have $y(0.2 ; h=0.2)=y_{1}=$ $y_{0}+h f\left(t_{0}, y_{0}\right)=1.6$. If we instead use $h=0.1$ we get

$$
y(0.1) \approx y_{1}=y_{0}+h f\left(t_{0}, y_{0}\right)=2-0.1 \cdot\left(0^{2}-2\right)=1.8 \text { and } y(0.2) \approx y_{2}=y_{1}+h f\left(t_{1}, y_{1}\right) \approx 1.621
$$

Considering that the truncation error for Euler is $R_{T}=C h$ we also find $y(0.2 ; 2 h)-y(0.2 ; h) \approx$ $y(0.2)+c 2 h-y(0.2)-c h=c h$. Thus, with $h=0.1$ we get $R_{T} \approx c h \approx(y(0.2 ; 2 h)-y(0.2 ; h)) \approx 0.021$. 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 $\left|R_{B}\right| \leq 0.5 \cdot 10^{-4}$ and $y(0.2)=1.621 \pm 0.022$.

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

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

Exercise 6.5 The method is implicit since the right hand side has the term $f\left(t_{k+1}, y_{k+1}\right)$ 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 $y^{\prime}=\lambda y, y(0)=1$, and obtain $y_{k+1}=y_{k}+\frac{h \lambda}{2}\left(y_{k}+y_{k+1}\right)$ 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 $\left|1+\frac{h \lambda}{2}\right| \leq\left|1-\frac{h \lambda}{2}\right|$. 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
$k_{1}=h f\left(t_{k}, y_{k}\right)=h \lambda y_{k}$, and $k_{2}=h f\left(t_{k}, y_{k}+k 1\right)=h \lambda\left(y_{k}+k_{1}\right)=h \lambda\left(y_{k}+h \lambda y_{k}\right)=\left(h \lambda+(h \lambda)^{2}\right) y_{k}$.
The next iterate $y_{k+1}$ is thus given by

$$
y_{k+1}=y_{k}+\frac{1}{2}\left(k_{1}+k_{2}\right)=y_{k}+\frac{1}{2}\left(h \lambda+h \lambda+(h \lambda)^{2}\right) y_{k}=\left(1+h \lambda+(h \lambda)^{2} / 2\right) y_{k} .
$$

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

Exercise 6.7 We define $v=y^{\prime}$. Then $v^{\prime}=y^{\prime \prime}$ and

$$
\binom{y}{v}^{\prime}=\binom{v}{t y-3 v-1}, \quad\binom{y(0)}{v(0)}=\binom{1}{0}
$$

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

$$
\binom{y}{v}^{\prime}=\binom{v}{2 y\left(1+y^{2}\right)}, \quad\binom{y(0)}{v(0)}=\binom{1}{2}
$$

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

$$
\binom{y_{1}}{v_{1}}=\binom{y_{0}}{v_{0}}+h\binom{v_{0}}{2 y_{0}\left(1+y_{0}^{2}\right)}=\binom{1.4}{2.8}
$$

Similarily $\left(y_{2}, v_{2}\right)^{T}=(1.96,4.576)^{T}$. Thus we obtain the approximation $y(0.4) \approx y_{2}=1.96$.
Exercise 6.9 Introduce $v=y^{\prime}$ to obtain

$$
v^{\prime}+e\left(y^{2}-a\right) v+y=0, \quad y(0)=a, \quad v(0)=b
$$

The system is thus

$$
\binom{y}{v}^{\prime}=\binom{v}{-e\left(y^{2}-a\right) v-y}, \quad\binom{y(0)}{v(0)}=\binom{a}{b}
$$

## 7 Basic matrix operations and linear systems of equations

Exercise 7.1 We evaluate the expression using the following operations

$$
z=(A+I) B x+y=(A+I) x_{1}+y=A x_{1}+x_{1}+y=x_{2}+x_{1}+y=x_{3}+y=x_{4}
$$

Computing the matrix vector product $x_{1}=B x$ requires $m n$ multiplications and additions each, i.e. a total of $2 m n$ operations. The product $x_{2}=A x_{1}$ requires $2 m^{2}$ operations. The remaining two vector additons require $m$ additions (as $y, x_{1} \in \mathbb{R}^{m}$ ). So the operation count is $m(2 m+2 n+2$ ).

Exercise 7.2 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 $A z=x$. The order of computation is

$$
z_{1}=A b, \quad C z_{2}=b, \quad z_{3}=z_{1}+z_{2}, \quad z_{4}=A z_{3}, \quad z_{5}=2 z_{4}+z_{3}, \quad \text { and finally } B x=z_{5} .
$$

All intermediate results are vectors.
Exercise 7.3 An upper triangular system $R x=y$ can be solved using,

$$
x_{i}=\left(b_{i}-\sum_{j=i+1}^{n} r_{i j} x_{j}\right) / r_{i i}, \quad i=n, n-1, \ldots, 1 .
$$

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 7.4 The $L U$-decomposition and two triangular systems require $2 n^{3} / 3+2 n^{2}$ 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

$$
\frac{t(2 n)}{t(n)} \approx \frac{c 2^{3} n^{3}}{c n^{3}}=8
$$

If we compute the same quotients using the table we obtain

$$
\begin{array}{c|lll}
n & 1024 & 2048 & 4096 \\
\hline t(2 n) / t(n) & 7.4166 & 7.1835 & 7.8295
\end{array}
$$

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

Exercise 7.5 The unit circles are


Exercise 7.6 A matrix norm is induced if its definition is based on a vector norm, i.e.

$$
\|A\|=\max _{x \neq 0} \frac{\|A x\|}{\|x\|}
$$

For such norms we have

$$
\|A B\|=\max _{x \neq 0} \frac{\|A B x\|}{\|x\|}=\max _{x \neq 0} \frac{\|A B x\|}{\|B x\|} \frac{\|B x\|}{\|x\|} \leq\left(\max _{x \neq 0} \frac{\|A B x\|}{\|B x\|}\right)\left(\max _{x \neq 0} \frac{\|B x\|}{\|x\|}\right) \leq \max _{y \neq 0} \frac{\|A y\|}{\|y\|}\|B\| \leq\|A\|\|B\| .
$$

Exercise 7.7 Demonstrate the first inequality by

$$
\|x\|_{\infty}^{2}=\max _{1 \leq i \leq n}\left|x_{i}\right|^{2} \leq \sum_{i=1}^{n}\left|x_{i}\right|^{2}=\|x\|_{2}^{2}
$$

Also, since $\left|x_{i}\right| \leq\|x\|_{\infty}$, we have

$$
\|x\|_{2}^{2}=\sum_{i=1}^{n}\left|x_{i}\right|^{2} \leq \sum_{i=1}^{n}\|x\|_{\infty}^{2}=n\|x\|_{\infty}
$$

Exercise 7.8 Recall the definition

$$
\left\|u v^{T}\right\|_{2}=\max _{x \in \mathbb{R}^{n}} \frac{\left\|u v^{T} x\right\|_{2}}{\|x\|_{2}}=\max _{x \in \mathbb{R}^{n}} \frac{\left|v^{T} x\right|\|u\|_{2}}{\|x\|_{2}}
$$

The Cauchy-Schwarz inequality is $\left|v^{T} x\right| \leq\|v\|_{2}\|x\|_{2}$ with equality for $x=v$. So

$$
\left\|u v^{T}\right\|_{2}=\frac{\left|v^{T} v\right|\|u\|_{2}}{\|v\|_{2}}=\|v\|_{2}\|u\|_{2} .
$$

Exercise 7.9 First from the definition of the matrix norm, and since $I x=x$ we have

$$
\|I\|=\max _{x \neq 0} \frac{\|I x\|}{\|x\|}=\max _{x \neq 0} \frac{\|x\|}{\|x\|}=1, \text { so } 1=\|I\|=\left\|A A^{-1}\right\| \leq\|A\|\left\|A^{-1}\right\| .
$$

Exercise 7.10 If $x=(1,-3,7)^{T}$ then $\|x\|_{\infty}=7,\|x\|_{2}=\sqrt{59}$ and $\|x\|_{1}=11$.
Exercise 7.11 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 7.12 From the second row we get the largest sum and $\|A\|_{\infty}=3.1+0.5+3.2=6.8$.

Exercise 7.13 The linear system is $A x=b$ where

$$
A=\left(\begin{array}{ccc}
2 & -1 & 2 \\
3 & 1 & -1 \\
-3 & 1 & 2
\end{array}\right) \text { and } b=\left(\begin{array}{c}
1 \\
0 \\
-3
\end{array}\right)
$$

Exercise 7.14 The rows should appear in the order $1-4-2-3$ and thus

$$
P=\left(\begin{array}{llll}
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{array}\right)
$$

Exercise 7.15 The rows should appear in the order $4-3-1-2$ which gives

$$
P=\left(\begin{array}{llll}
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{array}\right)
$$

Exercise 7.16 The multipliers are $m_{3}=0.6 / 3=0.2$ and $m_{4}=-1.8 / 3=-0.6$. Therefore the Gauss transformation is

$$
M=\left(\begin{array}{cccc}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0.2 & 1 & 0 \\
0 & -0.6 & 0 & 1
\end{array}\right)
$$

Exercise 7.17 The multipliers are $m_{21}=-1 / 2=-0.5$ and $m_{31}=1 / 2=0.5$. Therefore the Gauss transformation is

$$
M_{1}=\left(\begin{array}{ccc}
1 & 0 & 0 \\
0.5 & 1 & 0 \\
-0.5 & 0 & 1
\end{array}\right)
$$

Exercise 7.18 Set $m=\left(0, m_{21}, m_{31}\right)^{T}$ and find that

$$
M_{1}=I-m e_{1}^{T}=\left(\begin{array}{ccc}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{array}\right)-\left(\begin{array}{c}
0 \\
m_{21} \\
m_{31}
\end{array}\right)(1,0,0)
$$

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

$$
\left(I-m e_{1}^{T}\right)\left(I+m e_{1}^{T}\right)=I-m e_{1}^{T}+m e_{1}^{T}-m e_{1}^{T} m e_{1}^{T}=I
$$

where $e_{1}^{T} m=0$ since $m$ has a zero in the first position. The elements of $m$ are chosen so that $0=a_{i 1}-m_{i 1} a_{11}$ and thus $m_{i 1}=a_{i 1} / a_{11}, i=2,3$.

Exercise 7.19 Set $m=\left(0,0, m_{32}, m_{42}\right)^{T}$ and find that

$$
M_{2}=I-m e_{2}^{T}=\left(\begin{array}{cccc}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{array}\right)-\left(\begin{array}{c}
0 \\
0 \\
m_{32} \\
m_{42}
\end{array}\right)(0,1,0,0)
$$

has the desired structure. To show that $M_{2}^{-1}=I+m e_{2}^{T}$ we do

$$
\left(I-m e_{2}^{T}\right)\left(I+m e_{2}^{T}\right)=I-m e_{2}^{T}+m e_{2}^{T}-m e_{2}^{T} m e_{2}^{T}=I
$$

where $e_{2}^{T} m=0$ since $m$ has zeros in the first two positions.
Exercise 7.20 Pivoting is required as $\left|a_{21}\right|>\mid a_{11}$. The permutation

$$
P_{12}\left(\begin{array}{lll}
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1
\end{array}\right)
$$

should be used to switch rows one and two.
Exercise 7.21 We need to switch rows two and three to obtain

$$
\left(\begin{array}{ccc}
2 & 1 & -2 \\
0 & 1.7 & 0.3 \\
0 & -0.5 & 1.6
\end{array}\right) .
$$

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

$$
P_{23}=\left(\begin{array}{lll}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0
\end{array}\right) \text { and } M_{2}=\left(\begin{array}{ccc}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0.2941 & 1
\end{array}\right) .
$$

The new element on position $(3,3)$ will be $1.6-(-0.2941) \cdot 0.3=1.6882$. The upper triangular matrix is

$$
U=\left(\begin{array}{ccc}
2 & 1 & -2 \\
0 & 1.7 & 0.3 \\
0 & 0 & 1.6882
\end{array}\right)
$$

Exercise 7.22 In order to take advantage of the decomposition we rewrite $A x=b$ as $P A x=P b$ and $L(U x)=P b$. By introducing the intermediate variable $y=U x$ we obtain two triangular systems $L y=P b$ and $U x=y$. First solve

$$
L y=P b \text { or }\left(\begin{array}{ccc}
1 & 0 & 0 \\
-0.5 & 1 & 0 \\
0.5 & 0.8 & 1
\end{array}\right)\left(\begin{array}{l}
y_{1} \\
y_{2} \\
y_{3}
\end{array}\right)=\left(\begin{array}{lll}
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{array}\right)\left(\begin{array}{c}
-0.5 \\
1.2 \\
1.27
\end{array}\right)=\left(\begin{array}{c}
1.2 \\
1.27 \\
-0.5
\end{array}\right)
$$

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

Exercise 7.23 Since the multiplier $m_{32}=1.8>1$ pivoting cannot have been used correctly.
Exercise 7.24 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

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

Exercise 7.25 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

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

The absolute error is $\|\delta x\|_{\infty} \leq 0.48 \cdot 3.5113<1.7$.
Exercise 7.26 Use the $L U$ decomposition of $A$ to obtain

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

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 $L U$ decomposition.

Exercise 7.27 Let $r=b-A \widehat{x}$. Then $A^{-1} r=A^{-1} b-A^{-1} A \widehat{x}=x-\widehat{x}$ and we obtain $\|x-\widehat{x}\| \leq$ $\left\|A^{-1}\right\|\|r\|$.

Exercise 7.28 Since $P=I$ we have $\operatorname{det}(P)=1$ and we obatin

$$
\operatorname{det}(A)=\operatorname{det}\left(P^{T}\right) \operatorname{det}(L) \operatorname{det}(U)=1 \cdot 1 \cdot(2 \cdot 2.5 \cdot 1.4)=7
$$

## 8 Least Squares Problems and Orthogonal Decompositions

Exercise 8.1 The model is $y_{i}=c_{0}+c_{1} x_{i}+c_{2} \sin \left(x_{i}\right)$ and each measurement results in a row of a linear system $A x=b$. We obtain

$$
\left(\begin{array}{ccc}
1 & x_{1} & \sin \left(x_{1}\right) \\
1 & x_{2} & \sin \left(x_{2}\right) \\
\vdots & \vdots & \vdots \\
1 & x_{m} & \sin \left(x_{m}\right)
\end{array}\right)\left(\begin{array}{l}
c_{0} \\
c_{1} \\
c_{2}
\end{array}\right)=\left(\begin{array}{c}
y_{1} \\
y_{2} \\
\vdots \\
y_{m}
\end{array}\right) .
$$

Exercise 8.2 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

$$
\left(\begin{array}{cc}
\sin \left(t_{1}\right) & \cos \left(t_{1}\right) \\
\sin \left(t_{2}\right) & \cos \left(t_{2}\right) \\
\vdots & \vdots \\
\sin \left(t_{m}\right) & \cos \left(t_{m}\right)
\end{array}\right)\binom{A \cos (\phi)}{A \sin (\phi)}=\left(\begin{array}{c}
u\left(t_{1}\right) \\
u\left(t_{2}\right) \\
\vdots \\
u\left(t_{m}\right)
\end{array}\right) .
$$

where the first have to find a least squares solution $x$ and then find the parameters $A$ and $\phi$ by, e.g., $\phi=\operatorname{atan}\left(x_{2} / x_{1}\right)$.

Exercise 8.3 We rewrite the model as $\log \left(t_{i}\right)=\log (C)+p \log \left(n_{i}\right)$, and obtain the linear system

$$
\left(\begin{array}{cc}
1 & \log (100) \\
1 & \log (200) \\
1 & \log (300) \\
1 & \log (400)
\end{array}\right)\binom{\log (C)}{p}=\left(\begin{array}{c}
\log (1.213) \\
\log (2.370) \\
\log (3.619) \\
\log (4.875)
\end{array}\right) .
$$

Exercise 8.4 Since $Q$ is orthogonal $Q^{T} Q=I$. So

$$
\|Q x\|_{2}^{2}=(Q x)^{T}(Q x)=x^{T} Q^{T} Q x=x^{T} x=\|x\|_{2}^{2}
$$

This means that $\left\|Q_{2}\right\|_{2}=\max \|Q x\|_{2} /\|x\|_{2}=\max 1=1$.
Exercise 8.5 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 $A x=x$ for every $x \in \mathbb{R}^{n}$ so $A=I$ is the identity matrix.

Exercise 8.6 Let

$$
A=Q\binom{R}{0}=Q_{1} R
$$

where $Q=\left(Q_{1}, Q_{2}\right)$. Since $Q$ is orthogonal we find that

$$
\|A x-b\|_{2}^{2}=\left\|Q^{T}(A x-b)\right\|_{2}^{2}=\left\|\binom{R}{0} x-\binom{Q_{1}^{T} b}{Q_{2}^{T} b}\right\|_{2}^{2}=\left\|R x-Q_{1}^{T} b\right\|_{2}^{2}+\left\|Q_{2}^{T} b\right\|_{2}^{2} .
$$

The minimum is achived for $x=R^{-1} Q_{1}^{T} b$. Thus only the reduced $Q R$ decomposition is needed.

Exercise 8.7 If $A=Q_{1} R$ then range $(A)=\operatorname{span}\left(Q_{1}\right)$. 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 $A x=b$ has a solution.

Exercise 8.8 The vector $a$ can be seen as a matrix in $\mathbb{R}^{n \times 1}$. This means that

$$
a=\left(a /\|a\|_{2}\right)\|a\|_{2}=Q_{1} R
$$

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 $x=\left(a^{T} b\right) /\left(a^{T} a\right)$. This is the same as $x=R^{-1} Q_{1}^{T} b$ with the decomposition above.

Exercise 8.9 Since $x=R^{-1} Q_{1}^{T} b$ is the least squares solution we can compute the residual as $r=b-A x=b-Q_{1} R R^{-1} Q_{1}^{T} b=\left(I-Q_{1} Q_{1}^{T}\right) b=P b$. We can also recall that $A x=Q_{1} Q_{1}^{T} b$ from the geometrical interpretation of the least squares problem.

Exercise 8.10 The matrix has orthogonal columns, i.e. if $A=\left(a_{1}, a_{2}\right)$ then $a_{1}^{T} a_{2}=0$. Thus the $Q R$ decomposition is

$$
A=\left(a_{1} /\left\|a_{1}\right\|_{2},\left\|a_{2} /\right\| a_{2} \|_{2}\right)\left(\begin{array}{cc}
\left\|a_{1}\right\|_{2} & 0 \\
0 & \left\|a_{2}\right\|_{2}
\end{array}\right)=Q_{1} R .
$$

The numbers are not very important.
Exercise 8.11 Let $W=R^{T} R$ be the Cholesky decomposition ans rewrite

$$
\|x\|_{W}^{2}=x^{T} W x=x^{T} R^{T} R x=(R x)^{T}(R x)=\|R x\|_{2}^{2}
$$

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

$$
\|A x-b\|_{W}=\|R(A x-b)\|_{2} .
$$

The normal equations are now $(R A)^{T}(R A) x=(R A)^{T}(R b)$ or $A^{T} W A x=A^{T} W b$.
Exercise 8.12 If $A=U \Sigma V^{T}$ where $U, V$ are orthogonal and $\Sigma=\operatorname{diag}\left(\sigma_{1}, \sigma_{2}, \ldots, \sigma_{n}\right)$ then $\|A\|_{2}=$ $\left\|U \Sigma V^{T}\right\|_{2}=\|\Sigma\|_{2}$. 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 $\left\|A^{-1}\right\|_{2}=$ $\left\|\Sigma^{-1}\right\|_{2}$. Since the diagonal elements of $\Sigma^{-1}$ are $1 / \sigma_{i}$ the largest diagonal element is $1 / \sigma_{n}$ and $\left\|A^{-1}\right\|_{2}=1 / \sigma_{n}$.

Exercise 8.13 Let $A^{T}=V \Sigma^{T} U^{T}$. If $y \in \operatorname{span}\left(u_{k+1}, \ldots, u_{m}\right)$ then $u_{i}^{T} y=0$ for $i=1, \ldots, k$. This is the null space of $A^{T}$.

Exercise 8.14 First compute $\left(A^{T} A\right)^{-1}=\left(V \Sigma^{T} U^{T} U \Sigma V^{T}\right)^{-1}=V\left(\Sigma^{T} \Sigma\right)^{-1} V^{T}$. Here $\Sigma^{T} \Sigma=$ $\operatorname{diag}\left(\sigma_{i}^{2}\right) \in \mathbb{R}^{n \times n}$. Thus $A\left(A^{T} A\right)^{-1} A^{T}=U \Sigma V^{T} V\left(\Sigma^{T} \Sigma\right)^{-1} V^{T} V \Sigma^{T} U^{T}=U \Sigma\left(\Sigma^{T} \Sigma\right)^{-1} \Sigma^{T} U^{T}$. Since $U$ is orthogonal $\left\|A\left(A^{T} A\right)^{-1} A\right\|_{2}=\left\|\Sigma\left(\Sigma^{T} \Sigma\right)^{-1} \Sigma^{T}\right\|_{2}$. Evaluate the product of the diagonal matrices to obtain

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

The norm is the largest diagonal entry, i.e. 1 .
Exercise 8.15 Let $B=U \Sigma V^{T}$. Since $V=\left(v_{1}, \ldots, v_{n}\right)$ provides a basis for $\mathbb{R}^{n}$ any $x$ can be written

$$
x=\sum_{i=1}^{n} c_{i} v_{i} \Longrightarrow B y=\sum_{i=1}^{n} c_{i} \sigma_{i} u_{i} .
$$

If $\|x\|_{2}=1$ then $\sum c_{i}^{2}=1$. So

$$
\|B x\|_{2}^{2}=\sum_{i=1}^{n} \sigma_{i}^{2} c_{i}^{2} \geq \sigma_{n} \sum_{i=1}^{n} c_{i}^{2}=\sigma_{n}^{2}
$$

with equality for $c=e_{n}$. So the minimum is $\sigma_{n}$ and it is obtained for $x= \pm v_{n}$.
Exercise 8.16 The decomposition $A=U \Sigma V^{T}$ can be written

$$
A=\sum_{i=1}^{n} \sigma_{i} u_{i} v_{i}^{T}
$$

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

$$
y=A x=\sum_{i=1}^{n} \sigma_{i}\left(v_{i}^{T} x\right) u_{i}
$$

so the $y$ is a linear combination of $\left\{u_{1}, \ldots, u_{n}\right\}$. Thus range $(A)=\operatorname{span}\left(u_{1}, \ldots, u_{n}\right)$ and the dimension of the range is $n$.

Exercise 8.17 Let $A=U \Sigma V^{T}$. Since $U=\left(u_{1}, \ldots, u_{m}\right)$ is a basis for $\mathbb{R}^{m}$ we can write

$$
b=\sum_{i=1}^{m}\left(u_{i}^{T} b\right) u_{i},
$$

Similarily, $V=\left(v_{1}, \ldots, v_{n}\right)$ is a basis for $\mathbb{R}^{n}$ so

$$
A x=A\left(\sum_{i=1}^{n}\left(v_{i}^{T} x\right) v_{i}\right)=\sum_{i=1}^{n} \sigma_{i}\left(v_{i}^{T} x\right) u_{i} .
$$

We obtain

$$
\left.\|A x-b\|_{2}^{2}=\| \sum_{i=1}^{n}\left(\sigma_{i}\left(v_{i}^{T} x\right)-\left(u_{i}^{T} b\right)\right) u_{i}-\sum_{i=n+1}^{m}\left(u_{i}^{T} b\right)\right) u_{i} \|_{2}^{2}=\sum_{i=1}^{n}\left|\sigma_{i}\left(v_{i}^{T} x\right)-\left(u_{i}^{T} b\right)\right|^{2}+\sum_{i=n+1}^{m}\left|u_{i}^{T} b\right|^{2} .
$$

The minimum is obtained for $\sigma_{i}\left(v_{i}^{T} x\right)-u_{i}^{T} b$ 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

$$
r=b-A x=\sum_{i=n+1}^{m}\left(u_{i}^{T} b\right) u_{i} .
$$

Exercise 8.18 Compute $A x$ to obtain

$$
A x=A\left(\sum_{i=1}^{m} \frac{u_{i}^{T} b}{\sigma_{i}} v_{i}\right)=\sum_{i=1}^{m} \frac{u_{i}^{T} b}{\sigma_{i}} A v_{i}=\sum_{i=1}^{m} \frac{u_{i}^{T} b}{\sigma_{i}} \sigma_{i} u_{i}=\sum_{i=1}^{m}\left(u_{i}^{T} b\right) u_{i}=b,
$$

where the last equality holds since $U=\left(u_{1}, \ldots, u_{m}\right)$ provides an orthogonal basis for $\mathbb{R}^{m}$ which is the space $b$ belongs to.
Since $m<n$ the matrix has a null space $\operatorname{null}(A)=\operatorname{span}\left(v_{m+1}, \ldots, v_{n}\right)$. If $x_{2}$ belongs to the nullspace then $A\left(x+x_{2}\right)=A x=b$ 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

$$
\min \|x\|_{2} \text { such that } A x=b,
$$

that is the minimum norm solution of the linear system $A x=b$.
Exercise 8.19 Since $\operatorname{rank}(A)=k$ we note that $\left\{v_{k+1}, \ldots, v_{n}\right\}$ is a basis for null $(A)$ and $\left\{v_{1}, \ldots, v_{k}\right\}$ is a basis for its orthogonal complement $(\operatorname{null}(A))^{\perp}$. Thus for evey $x$ we can write

$$
x=x_{1}+x_{2}=\left(\sum_{i=1}^{k} c_{i} v_{i}\right)+\left(\sum_{i=k+1}^{n} c_{i} v_{i}\right) .
$$

In order to determine $x_{1}$ we compute

$$
A x=A\left(x_{1}+x_{2}\right)=A x_{1}+0=\sum_{i=1}^{k} c_{i} \sigma_{i} u_{i}=b=\sum_{i=1}^{m}\left(u_{i}^{T} b\right) u_{i} .
$$

Where $\left(u_{i}^{T} b\right)=0$, 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 8.20 a) If $B$ has full rank then $B x=0$ if and only if $x=0$ so the unique, and only feasible, solution is precisely $x=0$.
b) If $\operatorname{rank}(B)=k<n$ then $B$ has a non-trivial null space and write $V=\left(V_{k}, V_{n-k}\right)$ 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

$$
\min _{c \in \mathbb{R}^{n-k}}\left\|\left(A V_{n-k}\right) c-b\right\|_{2} \text { and } x=V_{n-k} c .
$$

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

Exercise 8.21 Let $A=U \Sigma V^{T}$ and $U=\left(u_{1}, \ldots, u_{m}\right)$. A solution exists if $b \in \operatorname{range}(A)=$ $\operatorname{span}\left(u_{1}, \ldots, u_{n}\right)$. We can check this by, for instance, verifying that $u_{i}^{T} b=0$, for $i=n+1, \ldots, m$. If $m \gg n$ it is cheaper to instead check if

$$
b-\sum_{i=1}^{n}\left(u_{i}^{T} b\right) u_{i}=0
$$

If we split the matrix $U=\left(U_{1}, U_{2}\right)$ then the same criteria can be written as $U_{2}^{T} b=0$ or $b-U_{1} U_{1}^{T} b=$ 0 .

Exercise 8.22 The normal equations can be derived by the identity

$$
\min _{x}\|A x-b\|_{2}^{2}+\lambda^{2}\|x\|_{2}^{2}=\min _{x}\left\|\binom{A x-b}{\lambda x}\right\|_{2}=\min _{x}\left\|\binom{A}{\lambda I} x-\binom{b}{0}\right\|_{2} .
$$

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

$$
\left(\begin{array}{ll}
A^{T} & \lambda I
\end{array}\right)\binom{A}{\lambda I} x=\left(\begin{array}{ll}
A^{T} & I
\end{array}\right)\binom{b}{0} \text { or }\left(A^{T} A+\lambda^{2} I\right) x=A^{T} b .
$$

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

$$
x_{\lambda}=V\left(\Sigma^{T} \Sigma+\lambda I\right)^{-1} \Sigma U^{T} b=\sum_{i=1}^{n} \frac{\sigma_{i}}{\sigma_{i}^{2}+\lambda^{2}}\left(u_{i}^{T} b\right) v_{i} .
$$

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.

