Numerical Methods for Ordinary Differential Equations

Answers of the exercises

C. Vuik, S. van Veldhuizen and S. van Loenhout

2020

Delft University of Technology
Faculty Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Contents

Chapter 1

Introduction

1.1 Solutions

1. We notice that the n^{th} order Taylor polynomial about the point $x = c$ from an $(n+1)$ times differentiable function is given by

$$P_n(x) = f(c) + (x-c)f'(c) + \frac{(x-c)^2}{2!}f''(c) + \cdots + \frac{(x-c)^n}{n!}f^{(n)}(c).$$

The upper bound for the error in $x = d$ is $\mid P_n(d) - f(d) \mid = \mid R_n(d) \mid$, in which

$$R_n(x) = \frac{(x-c)^{n+1}}{(n+1)!}f^{(n+1)}(\xi),$$

for a ξ between x en c.

The second order Taylor polynomial of $f(x) = x^3$ about the point $x = 1$ is then given by

$$\begin{aligned} P_2(x) &= f(1) + (x-1)f'(1) + \frac{(x-1)^2}{2!}f''(1) = \\ &= 1 + 3(x-1) + 3(x-1)^2. \end{aligned}$$

It now follows that

$$P_2(0.5) = 0.25.$$

The upperbound for the error in $x = 0.5$ is given by

$$\mid R_2(0.5) \mid = \mid \frac{(0.5-1)^3}{3!}f'''(\xi) \mid = \mid \frac{(0.5-1)^3}{3!}6 \mid = 0.125$$

The 'actual' error easily can be calculated with $\mid P_2(0.5) - f(0.5) \mid$. Note that for this problem the 'actual' error is equal to the upper bound of the error.

2. First, note that $f^{(i)}(x) = e^x$ for all $i \in \mathbb{N}$. So the n^{th} order Taylor polynomial of $f(x) = e^x$ about the point $x = 0$ is given by

$$P_n(x) = 1 + x + \frac{1}{2}x^2 + \cdots + \frac{1}{n!}x^n.$$

The error $R_n(x)$ for $x \in [0, 0.5]$ must be smaller than 10^{-6}. So we need an n such that $R_n(x) \leq 10^{-6}$ for $x \in [0, 0.5]$, in which

$$R_n(x) = \frac{(x)^{n+1}}{(n+1)!}e^{\xi},$$

for a ξ between 0 and x.

To make sure that $R_n(x)$ is smaller than 10^{-6} on the entire interval, we need to find a n such that

$$\frac{(0.5)^{n+1}}{(n+1)!}e^{0.5} \leq 10^{-6}. \tag{1.1}$$

Conclude this yourself. 'Trying' some values of n, for example $n = 5, 6, 7, \ldots$ gives that $n = 7$ is the smallest value which satisfies condition (1.1).

3. We use the polynomial $P_2(x) = 1 - \frac{1}{2}x^2$ to approximate $f(x) = \cos x$. It appears that if we take the third order Taylor polynomial about the point $x = 0$, this is equal to $P_2(x)$. The first, second and third derivative of $f(x)$ are given by

$$\begin{aligned} f(x) &= \cos x \\ f'(x) &= -\sin x \\ f''(x) &= -\cos x \\ f'''(x) &= \sin x. \end{aligned}$$

The third order Taylor polynomial of $f(x) = \cos x$ about the point $x = 0$ is given by

$$\begin{aligned} P_3(x) &= \cos 0 - x \sin 0 + \frac{x^2}{2}\cos 0 - \frac{x^3}{3!}\sin 0 \\ &= 1 - \frac{1}{2}x^2 = P_2(x). \end{aligned}$$

An upperbound for the error is

$$R_3(x) = \frac{1}{4!}x^4 f^{(iv)}(\xi) = \frac{1}{4!}x^4 \cos\xi,$$

for a ξ between 0 and x. We need to find an *upperbound* of $R_3(x)$. Notice that $|\cos\xi| \leq 1$ and $x^4 \leq \left(\frac{1}{2}\right)^4$ for $x \in [-\frac{1}{2}, \frac{1}{2}]$. So an upperbound for $|R_3(x)|$ is

$$\frac{1}{4!}\left(\frac{1}{2}\right)^4 \cdot 1 = 0.0026.$$

4. It is given that $x = \frac{1}{3}$ so we know that $fl(x) = 0.333 \cdot 10^0$, because we need to calculate with a precision of three digits. In the same way it follows from $y = \frac{5}{7}$ that $fl(y) = 0.714 \cdot 10^0$.

   To give an example we will calculate $fl(fl(x) + fl(y))$ and $x + y$ and also the rounding error.

   $$fl(x) + fl(y) = (0.333 + 0.714) \cdot 10^0 = 1.047 \cdot 10^0.$$

   So $fl(fl(x) + fl(y))$ is equal to $1.05 \cdot 10^0$; because we need to calculate with a precision of three digits. The rounding error, also called the absolute error, is given by

   $$| (x + y) - fl(fl(x) + fl(y)) | = | \frac{22}{21} - 1.05 \cdot 10^0 | = 2.38 \cdot 10^{-3}.$$

   The other calculations with the floating point numbers can be done in the same way. The answers are given in the table below.

$\circ$	$fl(x \circ y)$	$x \circ y$	absolute error
$+$	$0.105 \cdot 10$	$\frac{22}{21}$	$2.38 \cdot 10^{-3}$
-	$-0.381 \cdot 10^0$	$-\frac{8}{21}$	$0.48 \cdot 10^{-4}$
$*$	$0.238 \cdot 10^0$	$\frac{5}{21}$	$0.95 \cdot 10^{-4}$
$/$	$0.466 \cdot 10^0$	$\frac{7}{15}$	$0.66 \cdot 10^{-3}$

5. An example program is:

```
mu=1;
while 1+mu>1
    mu=mu/2;
end
```

   The end value of μ is then the machine precision of your system.

Chapter 2

Interpolation

2.1 Solutions

1. **(a)** Theorem 2.2.1 gives the following equation

$$f(x) - L_1(x) = \frac{1}{2}(x - x_0)(x - x_1)f''(\xi) \tag{2.1}$$

and we need to prove that

$$|f(x) - L_1(x)| \leq \frac{1}{8}(x_1 - x_0)^2 \max_{\xi \in (a,b)} |f''(\xi)| \tag{2.2}$$

We use the fact that $\frac{1}{2}|(x - x_0)(x - x_1)|$ has a maximum in $x = \frac{x_0+x_1}{2}$ and we do following steps.

$$|f(x)-L_1(x)| = |\frac{1}{2}(x-x_0)(x-x_1)f''(\xi)| \leq \frac{1}{2}|(\frac{x_0+x_1}{2}-x_0)(\frac{x_0+x_1}{2}-x_1)|\max_{\xi\in(a,b)}|f''(\xi)|$$

$$= \frac{1}{8}(x_1 - x_0)^2 \max_{\xi \in (a,b)} |f''(\xi)|$$

(b) The difference between the exact polynomial L_1 and the perturbed linear interpolation polynomial $\hat{L_1}$ is bounded by

$$|L_1(x) - \hat{L_1}(x)| \leq \left(\frac{|x_1 - x| + |x - x_0|}{x_1 - x_0}\right)\varepsilon$$

using the proof of Theorem 2.3.1. For $x \in [x_0, x_1]$ we can easily see that

$$|L_1(x) - \hat{L_1}(x)| \leq \left(\frac{|x_1 - x| + |x - x_0|}{x_1 - x_0}\right)\varepsilon = \left(\frac{x_1 - x + x - x_0}{x_1 - x_0}\right)\varepsilon = \varepsilon$$

because $x_1 - x > 0$ and $x - x_0 > 0$.
For $x \geq x_1$ we obtain the following

$$|L_1(x) - \hat{L_1}(x)| \leq \left(\frac{|x_1 - x| + |x - x_0|}{x_1 - x_0}\right)\varepsilon = \left(\frac{x - x_1 + x - x_0}{x_1 - x_0}\right)\varepsilon$$

$$= \left(\frac{(x_1 - x_0) + 2(x - x_1)}{x_1 - x_0} \right) \varepsilon = \left(1 + 2\frac{x - x_1}{x_1 - x_0} \right) \varepsilon$$

because $x_1 - x < 0$ and $x - x_0 > 0$.

2. In $x = 3$, our function $f(x) = \frac{1}{x}$ has value $f(3) = \frac{1}{3} = 0.333333\ldots$. We are going to approximate $f(3)$ with the second degree Lagrange polynomial, using nodes $x_0 = 2, x_1 = 2.5$ en $x_2 = 4$. The second degree Lagrange polynomial of $f(x)$ is given by

$$\begin{aligned} L_2(x) &= \sum_{k=0}^{2} f(x_k) L_{k2}(x) = \\ &= \frac{1}{2} L_{02}(x) + \frac{2}{5} L_{12}(x) + \frac{1}{4} L_{22}(x). \end{aligned}$$

The approximation of $f(3)$ is equal to $L_2(3)$, which is given by

$$L_2(3) = \frac{1}{2} L_{02}(3) + \frac{2}{5} L_{12}(3) + \frac{1}{4} L_{22}(3).$$

Following the theorem of Lagrange, we can calcute $L_{i2}(x)$, $i = 0, 1, 2$ in $x = 3$:

$$\begin{aligned} L_{02}(3) &= \frac{(3 - x_1)(3 - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(3 - 2.5)(3 - 4)}{(2 - 2.5)(2 - 4)} = -0.5, \\ L_{12}(3) &= \frac{(3 - x_0)(3 - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(3 - 2)(3 - 4)}{(2.5 - 2)(2.5 - 4)} = \frac{4}{3}, \text{ and} \\ L_{22}(3) &= \frac{(3 - x_0)(3 - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(3 - 2)(3 - 2.5)}{(4 - 2)(4 - 2.5)} = \frac{1}{6}. \end{aligned}$$

So $L_2(3)$ is equal to

$$L_2(3) = \frac{1}{2}(-0.5) + \frac{2}{5} \cdot \frac{4}{3} + \frac{1}{4} \cdot \frac{1}{6} = 0.325.$$

3. **(i)** The cubic spline s of the function $f(x) = x$ is an interpolation spline of degree 3 with nodes $x_0 = 0, x_1 = 1$ en $x_2 = 2$. With the definition of the natural spline of degree 3 it follows that s is a third degree polynomial on each subinterval $[0, 1]$ and $[1, 2]$, in which there is a special condition in $x = 1$ for the connection. We would like to determine $s(\frac{1}{2})$. Note that it is sufficient to determine the spline $s_0(x)$.
In Section 2.5 is shown that we need to solve the following equation for b_1

$$2(h_0 + h_1)b_1 = 3\left(\frac{f_2 - f_1}{h_1} - \frac{f_1 - f_0}{h_0} \right) \tag{2.3}$$

You can also see in this section that $b_0 = b_2 = 0$. With the knowledge that $h_i = x_{i+1} - x_i$ and $f_i = f(x_i)$ we obtain

$$2(1 + 1)b_1 = 3\left(\frac{2 - 1}{1} - \frac{1 - 0}{1} \right),$$

and so we know that $b_1 = 0$.
With formula

$$a_0 = \frac{1}{3h_0}(b_1 - b_0)$$

we obtain $a_0 = 0$. With the next formula we got that $c_0 = 1$

$$c_0 = \frac{f_1 - f_0}{h_0} - h_0\frac{2b_0 + b_1}{3}$$

The value of d_0 is given by $d_j = f_j$ which is 0 for d_0. The spline $s_0(x)$ on the interval $[0, 1]$ is then given by

$$s_0(x) = a_0(x - x_0)^3 + b_0(x - x_0)^2 + c_0(x - x_0) + d_0 = x.$$

We conclude that de spline s in $x = \frac{1}{2}$ is exact, namely $s(\frac{1}{2}) = \frac{1}{2}$.

(ii) In this case the function f is given by $f(x) = x^2$. So we need to solve the following equation

$$2(h_0 + h_1)b_1 = 3\left(\frac{f_2 - f_1}{h_1} - \frac{f_1 - f_0}{h_0}\right)$$

After substituting the known values, we obtain

$$4b_1 = 6,$$

so $b_1 = 1.5$. We know that $b_0 = b_2 = 0$ and also that

$$a_0 = \frac{1}{3h_0}(b_1 - b_0) = \frac{1}{3}(1.5 - 0) = \frac{1}{2}.$$

The coefficient c_0 is equal to

$$c_0 = \frac{f_1 - f_0}{h_0} - h_0\frac{2b_0 + b_1}{3} = \frac{1}{2}.$$

We also know that $d_0 = f_0 = 0$. The spline $s_0(x)$ on the interval $[0, 1]$ is then given by

$$s_0(x) = \frac{1}{2}x^3 + \frac{1}{2}x.$$

Substituting the value of x gives

$$s_0(\frac{1}{2}) = \frac{5}{16}.$$

We conclude that the cubical spline is not exact for the function $f(x) = x^2$; $f(\frac{1}{2}) = \frac{1}{4}$.

4. (a) The Lagrange polynomial is given by

$$\begin{aligned} L_3(x) = 0 \cdot \frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} + 0.5 \cdot \frac{(x-0)(x-2)(x-3)}{(1-0)(1-2)(1-3)} \\ + 1 \cdot \frac{(x-0)(x-1)(x-3)}{(2-0)(2-1)(2-3)} + 2 \cdot \frac{(x-0)(x-1)(x-2)}{(3-0)(3-1)(3-2)} \quad . \end{aligned}$$

This can be simplified to

$$L_3(x) = \frac{1}{12}x^3 - \frac{1}{4}x^2 + \frac{2}{3}x \quad .$$

In order to find the speed we take the derivative

$$L_3'(x) = \frac{1}{4}x^2 - \frac{1}{2}x + \frac{2}{3} \; .$$

and to determine the accelaration we differentiate again

$$L_3''(x) = \frac{1}{2}x - \frac{1}{2}$$

The acceleration is zero at $x = 1$, but this corresponds to a minimum acceleration. The maximum acceleration is achieved at $x = 3$ and is the corresponding speed is $\frac{17}{12}$ km/min.

(b) Following the procedure in the book, we find the following spline:

$$\begin{aligned}
s_1(x) &= \frac{8}{15}x - \frac{1}{30}x^3, && 0 \le x \le 1; \\
s_2(x) &= \frac{1}{2} + \frac{13}{30}(x-1) - \frac{1}{10}(x-1)^2 + \frac{1}{6}(x-1)^3, && 1 \le x \le 2; \\
s_3(x) &= 1 + \frac{11}{15}(x-2) + \frac{2}{5}(x-2)^2 - \frac{4}{30}(x-2)^3, && 2 \le x \le 3.
\end{aligned}$$

For the velocity we get

$$\begin{aligned}
v_1(x) &= \frac{8}{15} - \frac{1}{10}x^2, && 0 \le x \le 1; \\
v_2(x) &= \frac{13}{30} - \frac{1}{5}(x-1) + \frac{1}{2}(x-1)^2, && 1 \le x \le 2; \\
v_3(x) &= \frac{11}{15} + \frac{4}{5}(x-2) - \frac{2}{5}(x-2)^2, && 2 \le x \le 3.
\end{aligned}$$

The maximum speed is achieved at the end point and is $\frac{17}{15}$ km/min, which is 68 km/h.

Chapter 3

Numerical differentiation

3.1 Solutions

1. First we notice that $f \in C^3[x-h, x+h]$ means that f is three times continuous differentiable on $[x-h, x+h]$. So the first, second and third derivative of f are continuous on $[x-h, x+h]$.

   We need to prove: $\mid f'(x) - Q(h) \mid = \mathcal{O}(h^2)$, in which $Q(h)$ is given by

   $$Q(h) = \frac{f(x+h) - f(x-h)}{2h}.$$

   The Taylor expansion from $f(x+h)$ about x with truncation error $\mathcal{O}(h^3)$ is

   $$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3). \tag{3.1}$$

   We have the following expansion for $f(x-h)$ about x with truncation error $\mathcal{O}(h^3)$:

   $$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3). \tag{3.2}$$

   Subtracting (3.2) from (3.1) gives us

   $$f(x+h) - f(x-h) = 2hf'(x) + \mathcal{O}(h^3), \tag{3.3}$$

   which is the same as

   $$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \mathcal{O}(h^2).$$

   The order of the truncation error is now given by

   $$\begin{aligned} \mid f'(x) - Q(h) \mid &= \mid f'(x) - \tfrac{f(x+h)-f(x-h)}{2h} \mid = \\ &= \mid f'(x) - f'(x) - \mathcal{O}(h^2) \mid = \mathcal{O}(h^2) \end{aligned}$$

2. The position of the ship during start up is given by $S(t) = 0.5at^2$. The velocity is approximated by a backward difference with step size h, so

$$S_b'(t) \approx \frac{S(t) - S(t-h)}{h}.$$

To determine the truncation error we expand $S(t-h)$ about point t with a Taylor polynomial of order 1;

$$S(t-h) = S(t) - hS_b'(t) + \frac{h^2}{2}S''(\xi), \tag{3.4}$$

with $\xi \in (t-h, t)$. Given the fact that $S''(t) = a$, we can re-order (3.4) to

$$S_b'(t) - \frac{S(t) - S(t-h)}{h} = \frac{1}{2}ah.$$

So the truncation error is $\frac{1}{2}ah$.

We can determine the measurement error in the speed as follows; we already now that the measurement error is less than 10 meters. So the approximated speed with measurement errors is given by;

$$S_b'(x) \approx \frac{S(t) \pm 10 - S(t-h) \pm 10}{h}.$$

This means that the measurement error in the approximated speed is bounded by $\frac{20}{h}$

The total error in the approximated speed is the sum of the truncation error and the measurement error. With $a = 0.004$ the total error (te) is given by

$$te(h) = 0.002h + \frac{20}{h}.$$

The error in the calculated velocity is minimal in the minimum of the function te. For positive step size h, the minimum is equal to $h = 100$. This can be obtained by setting $te'(h) = 0.002 - \frac{20}{h^2}$ equal to zero. The total error in the speed is then equal to $te(100) = 0.4m/s$.

3. **(a)** An upper bound for the truncation error is

$$|\frac{h^2}{12}f^{(4)}(\xi)| \leq \frac{h^2}{12}$$

because $|f^{(4)}(\xi)| = |\sin\xi| \leq 1$. An upper bound for the rounding error is $\frac{4\epsilon}{h^2}$. So an upper bound for the total error E_{tot} is given by

$$E_{tot} \leq \frac{h^2}{12} + \frac{4\epsilon}{h^2}$$

To minimize this upper bound, we need to calculte when the derivative is equal to zero.

$$\frac{h}{6} - \frac{8\epsilon}{h^3} = 0 \quad \rightarrow \quad h_{opt} = (48\epsilon)^{\frac{1}{4}} \approx 2.6 \cdot 10^{-4} \quad .$$

h	$Q_1(h)$	$\lvert -\sin(1) - Q_1(h)\rvert$
1	-0.77364	0.067826
0.1	-0.84077	0.00070099
0.01	-0.84146	7.0122e-06
0.001	-0.84147	7.0083e-08
0.0001	-0.84147	3.0251e-09
1e-05	-0.84147	7.5798e-07
1e-06	-0.84155	7.8068e-05
1e-07	-0.83267	0.0088037
1e-08	-1.1102	0.26875

(b) The table below gives $Q_1(h)$ and the error for different values of h.
The experiments show that the optimal value of h is about 10^{-4}. This value is close to the value determined in part (a).

4. The central difference formula $Q(h)$, approaches the unknown value M, in this case $f'(1)$, in which the error has the form

$$M - Q(h) = c_p h^p.$$

Given that the error in the central difference approximation is of order two, it follows that $p = 2$.

$f(x) = \sin x$, $h = 0.1$ and $f'(x)$ is approximated by a central difference, so

$$f'(1) \approx Q(0.1) = \frac{\sin(1.1) - \sin(0.9)}{0.2} = 0.5394.$$

Conclude for yourself that the error, $|\cos 1 - \frac{\sin(1.1)-\sin(0.9)}{0.2}|$, is equal to $0.9 \cdot 10^{-3}$. For $2h = 0.2$ we obtain that $Q(0.2) = 0.5367$.

With the formula

$$c_p = \frac{Q(h) - Q(2h)}{(h)^2((2)^p - 1)},$$

we found that with $p = 2$ and $h = 0.1$ that $c_p = 0.09$. Now it follows that

$$M - Q(h) = c_p h^p = 0.9 \cdot 10^{-3}.$$

This is a good estimate of the error by Richardson's extreapolation.

5. We look for a $Q(h)$ such that

$$|\, f'(x) - Q(h) \,| = \mathcal{O}(h^k),$$

with k maximal, with $f(x)$, $f(x+h)$ and $f(x+2h)$ given. When the error is minimal, the k is maximal. Notice that $Q(h)$ is given by

$$Q(h) = \frac{\alpha_0}{h} f(x) + \frac{\alpha_1}{h} f(x+h) + \frac{\alpha_2}{h} f(x+2h). \tag{3.5}$$

Taylor expansion of $f(x)$, $f(x+h)$ and $f(x+2h)$ about x gives

$$\begin{aligned} f(x) &= f(x), && (3.6)\\ f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3), && (3.7)\\ f(x+2h) &= f(x) + 2hf'(x) + \frac{4h^2}{2}f''(x) + \mathcal{O}(h^3). && (3.8) \end{aligned}$$

The values of α_0, α_1 and α_2 can be obtained by substituting (3.6), (3.7) and (3.8) into (3.5). Note that the coefficient in front of $f(x)$ and $f''(x)$ must be equal to zero and that the coefficient in front of $f'(x)$ must be equal to one. This gives the following 3×3 system

$$\begin{array}{rccccccc} f(x): & \frac{\alpha_0}{h} & + & \frac{\alpha_1}{h} & + & \frac{\alpha_2}{h} & = & 0, \\ f'(x): & & & \alpha_1 & + & 2\alpha_2 & = & 1, \\ f''(x): & & & \frac{h}{2}\alpha_1 & + & 2h\alpha_2 & = & 0. \end{array} \qquad (3.9)$$

Solving (3.9) yields

$$\begin{cases} \alpha_0 &= -\frac{3}{2} \\ \alpha_1 &= 2 \\ \alpha_2 &= -\frac{1}{2} \end{cases}.$$

Such that

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

The truncation error to approximate $f'(x)$ is given by $\mathcal{O}(h^2)$, because

$$f'(x) - Q(h) = f'(x) - f'(x) + \mathcal{O}(h^2)$$

Chapter 4

Nonlinear equations

4.1 Solutions

1. We want to determine p_3 on the interval $[-2, 1.5]$. First we need to check if $f(a) = f(-2)$ and $f(b) = f(1.5)$ have opposite signs. With $f(x) = 3(x+1)(x-\frac{1}{2})(x-1)$ we find that

$$\begin{aligned} f(a) = &\quad f(-2) &&= -22\frac{1}{2} \\ f(b) = &\quad f(1.5) &&= 3\frac{3}{4}. \end{aligned}$$

So they have opposite signs.

To start, we take $a_0 = a$ and $b_0 = b$ and the new approximation for the zero is computed by

$$p_0 = \frac{1}{2}(a_0 + b_0) = -0.25.$$

The stopping criterion is not satisfied, because $f(p_0) = 2.10 \neq 0$. So we construct a new interval $[a_{n+1},b_{n+1}]$. $f(p_0)f(a_0) < 0$ so we choose $a_1 = a_0$ and $b_1 = p_0$. This leads to $p_1 = -1.125$ and $f(p_1) = -1.295$.

Now we have that $f(p_1)f(a_1) > 0$ so we choose $a_2 = p_1$ and $b_2 = b_1$. This gives the value $p_2 = -0.6875$ and $f(p_2) = 1.879$. So $f(p_2)f(a_2) < 0$, so we choose $a_3 = a_2$ and $b_3 = p_2$ such that $p_3 = -0.991$.

We can determine p_3 in the same way for interval $[-1.25, 2.5]$.

2. (i) We start with the following fixed-point method:

$$p_n = \frac{20p_{n-1} + \frac{21}{p_{n-1}^2}}{21}. \tag{4.1}$$

Define the function g as

$$g(x) = \frac{20x + \frac{21}{x^2}}{21}.$$

Recursion (4.1) can now be written as $p_n = g(p_{n-1})$. The fixed point is the value of p for which: $g(p) = p$. This leads othe following equation

$$p = \frac{20p + \frac{21}{p^2}}{21}$$

Which is equivalent to

$$21p = 20p + \frac{21}{p^2}.$$

Rewriting leads to $p^3 = 21$ and so

$$p = 21^{\frac{1}{3}}.$$

The fixed point is $p = 21^{\frac{1}{3}}$.
The convergence speed is given by $g'(p)$. Conclude yourself that

$$g'(x) = \frac{20}{21} - \frac{2}{x^3},$$

and so $g'(p) = \frac{18}{21} \approx 0.8571$.
If p_0 is equal to 1 then with $p_1 = g(p_0)$ we find

$$p_1 = \frac{20p_0 + \frac{21}{p_0^2}}{21} = \frac{20 + 21}{21} = \frac{41}{21} = 1.9524.$$

With $p_2 = g(p_1)$ we have

$$p_2 = \frac{20 \cdot 1.9524 + \frac{21}{1.9524^2}}{21} \approx 2.1218.$$

In the same way we have

$$p_3 \approx 2.2428.$$

(ii) For the next fixed-point method, we define $g(x)$ as

$$g(x) = x - \frac{x^3 - 21}{3x^2},$$

so $p_n = g(p_{n-1})$. Conclude yourself that the fixed point $p = g(p)$ is obtained by solving

$$p^3 - 21 = 0$$

where we assume that $3p^2 \neq 0$. Again we find $p = \sqrt[3]{21}$, note that our assumption is satisfied.
The convergence speed is given by $g'(p)$. The first derivative of $g(x)$ is

$$g'(x) = \frac{2}{3} - \frac{14}{x^3}.$$

So we found $g'(p) = \frac{2}{3} - \frac{14}{21} = 0$. With Theorem 4.4.2 we can see that the second method converges faster.
With $p_0 = 1$ we get the following values of p_i, $i = 1, 2, 3$

$$\begin{aligned} p_1 &= 7.6667, \\ p_2 &= 5.2302 \quad \text{and} \\ p_3 &= 3.7427. \end{aligned}$$

3. (a) The linear interpolation polynomial $l(x)$ is given by

$$l(x) = f(p_0) + \frac{x - p_0}{p_1 - p_0}\left(f(p_1) - f(p_0)\right).$$

See also Chapter 2.

(b) The new point p_2 is the point where $l(p_2) = 0$. With **(a)** we have to solve p_2 from

$$f(p_0) + \frac{p_2 - p_0}{p_1 - p_0}\left(f(p_1) - f(p_0)\right) = 0$$

Solving this gives

$$p_2 = p_0 + \frac{(p_1 - p_0)f(p_0)}{f(p_0) - f(p_1)}.$$

A general method for computing p_n from p_{n-2} and p_{n-1} can be obtained in the same way. Then we have the following equation

$$p_n = p_{n-1} + \frac{(p_{n-1} - p_{n-2})f(p_{n-2})}{f(p_{n-2}) - f(p_{n-1})}.$$

(c) With the formula from **(b)** it follows that

$$p_2 = 1 + \frac{1f(1)}{f(1) - f(2)} = \frac{4}{3}.$$

So we have the following expression for p_3:

$$p_3 = 2 + \frac{(\frac{4}{3} - 2)f(2)}{f(2) - f(\frac{4}{3})} = \frac{7}{5} = 1.4.$$

4. We determine an approximation of the zero of $f(x) = x - \cos x$ on the interval $\left[0, \frac{\pi}{2}\right]$ using the Newton-Raphson method.[1] With start approximation $p_0 = 0$ we can generate the sequence $\{p_n\}_{n=0}^{N}$, with N such that the error is smaller than 10^{-4}, using

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}.$$

We take the following stop criteria: Stop if the fourth decimal does not change in two following iterations. Newton-Raphson has a quadratic convergence so this means that the error is smaller than 0.0001.

Every iteration p_n should be determined with four decimals. Starting with $p_0 = 0$ we obtain that

$$p_1 = 0 - \frac{0 - \cos 0}{1 + \sin 0} = \cos 0 = 1.$$

p_2 is given by

$$p_2 = 1 - \frac{1 - \cos 1}{1 + \sin 1} \approx 0.7504.$$

[1] Notice that we approximate the solution of the equation $x = \cos x$

In the same way we obtained

$$\begin{aligned} p_3 &= 0.7391 \\ p_4 &= 0.7391. \end{aligned}$$

p_3 and p_4 are identical in four decimals which means that the error is smaller than 0.0001.

5. The system that needs to be solved is

$$\left\{ \begin{array}{rcl} x^2 - y - 3 &=& 0 \\ -x + y^2 + 1 &=& 0 \end{array} \right. .$$

In vector notation, this is equal to $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ with $\mathbf{x} = [x, y]^T$. The Jacobian matrix of $\mathbf{F}(\mathbf{x})$ is

$$J(\mathbf{x}) = \begin{pmatrix} 2x & -1 \\ -1 & 2y \end{pmatrix}.$$

With starting vector $\mathbf{x}^{(0)} = [1, 1]^T$ the first iteration is given by

$$\begin{aligned} \mathbf{x}^{(1)} &= \mathbf{x}^{(0)} - J(\mathbf{x}^{(0)})^{-1}\mathbf{F}(\mathbf{x}^{(0)}) = \\ &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} -3 \\ 1 \end{pmatrix} = \\ &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} -3 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{8}{3} \\ \frac{4}{3} \end{pmatrix} \approx \begin{pmatrix} 2.6 \\ 1.3 \end{pmatrix}. \end{aligned}$$

In the same way we found $\mathbf{x}^{(2)} = \begin{pmatrix} 2.0980 \\ 1.0784 \end{pmatrix}$.

We used the following formula for calculating the inverse of a 2×2 matrix. If

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

and $ad - bc \neq 0$, then $\mathbf{A}$ is invertible and $\mathbf{A}^{-1}$ is given by

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

This is also known as Kramer's rule for 2×2 matrices.

Chapter 5

Numerical integration

5.1 Solutions

1. We want to compute the following integral

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

with $a = -1$, $b = 1$ and $f(x) = (10x)^3 + 0.001$.

We assume that the relative rounding error in the function values is less than ε:

$$|f(x) - \hat{f}(x)| \leq |f(x)|\varepsilon,$$

Now we can determine the following upperbound for the relative rounding:

$$\frac{\left|\int_a^b f(x)dx - \int_a^b \hat{f}(x)dx\right|}{\left|\int_a^b f(x)dx\right|} \leq K\varepsilon,$$

in which K is called the condition number of the integral. The condition number is defined by

$$K = \frac{\int_a^b |f(x)|\,dx}{\left|\int_a^b f(x)dx\right|}.$$

See page 59 of the textbook.

(a) The relative rounding error is $K\varepsilon$. To calculate the condition number we need to determine the following two integrals :

$$\left|\int_{-1}^{1} \left((10x)^3 + 0.001\right) dx\right| \qquad \text{and}$$
$$\int_{-1}^{1} \left|(10x)^3 + 0.001\right| dx.$$

The first integral gives us

$$\left|\int_{-1}^{1} \left((10x)^3 + 0.001\right) dx\right| = 0.002.$$

We need to split the second integral into two parts. So we obtain

$$\begin{aligned}\int_{-1}^{1} \left|(10x)^3 + 0.001\right| dx &= \int_{-1}^{-0.01} -(10x)^3 - 0.001 dx \\ &+ \int_{-0.01}^{1} (10x)^3 + 0.001 dx = 500.\end{aligned}$$

The relative rounding error is equal to $\frac{500}{0.002}\varepsilon = 2.5 \cdot 10^5 \varepsilon$.

(b) The relative rounding error in the function values is less than (or equal to) ε, so the absolute rounding error is equal to

$$|f(x) - \hat{f}(x)| \leq |f(x)|\varepsilon.$$

With $f(x) = (10x)^3 + 0.001$ it follows that the absolute rounding error of the integral when we choose $\varepsilon = 4 \cdot 10^{-8}$ is given by

$$\int_{-1}^{1} |\varepsilon(10x)^3| dx = 2\int_{0}^{1} \varepsilon(10x)^3 dx = 2 \cdot 10^{-5},$$

The truncation error of the Midpoint rule is given by

$$-\frac{b-a}{24} h^2 M,$$

in which M is the maximum of $f''(x)$, $x \in [-1, 1]$. The second derivative of f is $f''(x) = 6000x$. So the maximal truncation error is equal to

$$500h^2.$$

We have a good assumption for the intergral if the rounding error and the truncation error are in balance. We have to notice that both errors can't be avoided (in our case) and so we have a good estimation if the errors are equal. So the step size h must fulfil

$$500h^2 = 2 \cdot 10^{-5},$$

so $h \approx 10^{-4}$.

2. In this exercise we determine or estimate the integral $\int_{0.5}^{1} x^4 dx$ with the following methods:

   (i) Exact

   (ii) Trapezoidal

   (iii) Composite Trapezoidal

   When $\int_{0.5}^{1} x^4 dx$ is calculated exact, we obtain

$$\int_{0.5}^{1} x^4 dx = \left(\frac{1}{5}x^5\right)\Big|_{0.5}^{1} = \frac{1}{5}(1 - (0.5)^5) = 0.19375.$$

With the Trapezoidal rule[1] we receive

$$\int_{0.5}^{1} x^4 dx \approx \frac{1-0.5}{2}\left((0.5)^4 + 1^4\right) = 0.265625.$$

An estimate of the error can be determined with Theorem 5.3.3 of the book.[2] The estimate of the error with $f''(x) = 12x^2$ is given by

$$\frac{(1-0.5)^3}{12} \cdot \left(\max_{x\in[x_R,x_L]} |f''(x)|\right) = (0.5)^3 = 0.125$$

The real error is |0.19375-0.265625| = 0.0719.

The same integral is calculated with the composite Trapezoidal rule.[3] We have $h = 0.25$, so $x_0 = 0.5$, $x_1 = 0.75$ en $x_2 = 1$, the integral is approximated by

$$\int_{0.5}^{1} x^4 dx \approx 0.25\left(\frac{1}{2}(0.5)^4 + (0.75)^4 + \frac{1}{2}1^4\right) = 0.2119.$$

The error is estimated by Richardson's method. So we must also approximate the integral with the composite Trapezoidal rule with $\frac{h}{2} = 0.125$. This approximation is 0.1983. From the error estimation follows that $\alpha = 2$, so

$$c_p = \frac{Q(h) - Q(2h)}{h^2(2^2-1)} = -0.2901.$$

See Chapter 3. It follows that the estimate of the error is equal to

$$|M - N(h)| = |c_p(h)^p| = 0.018.$$

The 'real' error is 0.0182, so the estimate with Richardson is very good.

3. For the Trapezoidal rule we use intervals with the same lengths, so $h = 0.25$ and $x_0 = 1$, $x_1 = 1.25$ and $x_2 = 1.5$. Using the composite Trapezoidal rule, we obtain

$$\int_{1}^{1.5} x^7 dx \approx 0.25\left(\frac{1}{2}(1)^7 + (1.25)^7 + \frac{1}{2}(1.5)^7\right) = 3.45$$

[1] **Trapezoidal rule**:

$$\int_{x_L}^{x_R} f(x)dx \approx \frac{x_R - x_L}{2}\left(f(x_L) + f(x_R)\right)$$

[2] **Theorem 5.3.3** For $m_2 = \max_{x\in[x_L,x_R]}|f''(x)|$ holds

$$\left|\int_{x_L}^{x_R} f(x)dx - \frac{x_R - x_L}{2}\left(f(x_L) + f(x_R)\right)\right| \leq \frac{1}{12}m_2(x_R - x_L)^3.$$

[3] **Composite Trapezoidal**:

$$I_T = \frac{h}{2}\sum_{k=1}^{n}\left(f(x_{k-1}) + f(x_k)\right) = h\left(\frac{1}{2}f(x_L) + f(x_L + h) + \cdots + f(x_R - h) + \frac{1}{2}f(x_R)\right),$$

in which $h = \frac{x_R - x_L}{n}$ and $x_k = x_L + kh$. For the error follows:

$$\left|\int_{x_L}^{x_R} f(x)dx - I_T\right| \leq \frac{1}{12}M_2(x_R - x_L)h^2 \quad \text{where, again, } M_2 = \max_{x\in[x_L,x_R]}|f''(x)|.$$

The error for this rule is calculated with $f''(x) = 42x^5$,

$$\frac{(1.5-1)^3}{12} \cdot \left(\max_{x \in [x_R, x_L]} |f''(x)| \right) = (0.5)^3 \frac{318.94}{12} = 3.32$$

Using Simpson's rule [4]

gives us

$$\int_1^{1.5} x^7 dx \approx 0.25 \left(\frac{1}{6} 1^7 + \frac{2}{3}(1.125)^7 + \frac{1}{3}(1.25)^7 + \frac{2}{3}(1.375)^7 + \frac{1}{6}(1.5)^7 \right) = 3.0798$$

The error is calculated with Theorem 5.3.4 of the book.[5]
The estimate of the error with $f^{(4)} = 840x^3$ is given by

$$\frac{(1.5-1)^5}{2880} \cdot \left(\max_{x \in [x_R, x_L]} |f^{(4)}(x)| \right) = (0.5)^5 \frac{2835}{2880} = 0.031$$

If we compare our results with Table 5.3 of the book, we can see that the Simpson's rule is better than the 1-point Gauss quadrature rule.

[4] **Simpson's rule**

$$I_S = \frac{h}{6} \sum_{k=1}^{n} \left(f(x_{k-1}) + 4f\left(\frac{x_{k-1} + x_k}{2} \right) + f(x_k) \right)$$

$$= h \left(\frac{1}{6} f(a) + \frac{2}{3} f(a + \frac{1}{2}h) + \frac{1}{3} f(a+h) + \frac{2}{3} f(a + \frac{3}{2}h) + \cdots + \frac{2}{3} f(b - \frac{1}{2}h) + \frac{1}{6} f(b) \right)$$

[5] **Theorem 5.3.4** For $m_4 = \max_{x \in [x_L, x_R]} |f^{(4)}(x)|$ holds

$$\left| \int_{x_L}^{x_R} f(x)dx - \frac{x_R - x_L}{6} \left(f(x_L) + 4f(x_M) + f(x_R) \right) \right| \leq \frac{1}{2880} m_4 (x_R - x_L)^5.$$

Chapter 6

Numerical time integration of initial-value problems

6.1 Solutions

1. The *Modified Euler method* approximates the solution to an initial-value problem $y' = f(t, y)$ in the following way:

   Suppose that the numerical approximation w_n of $y(t_n)$ is known. Then we can compute the following predictor and corrector:

   $$\begin{aligned} \text{predictor}: \overline{w}_{n+1} &= w_n + \Delta t f(t_n, w_n), \\ \text{corrector}: w_{n+1} &= w_n + \frac{\Delta t}{2}\left(f(t_n, w_n) + f(t_{n+1}, \overline{w}_{n+1})\right). \end{aligned}$$

   The initial-value problem in this exercise is given by

   $$\begin{cases} y' = 1 + (t - y)^2, & 2 \leq t \leq 3 \\ y(2) = 1 \end{cases}$$

   The solution is given by

   $$y(t) = t + \frac{1}{1 - t}.$$

   With a time step $\Delta t = 0.5$ we need to determine the following values of w for $2 \leq t \leq 3$:

   $$\begin{aligned} w_0 &\cong y(2), \\ w_1 &\cong y(2.5) \quad \text{and} \\ w_2 &\cong y(3). \end{aligned}$$

   From the initial condition we obtain that $w_0 = y(2) = 1$. So the error in this point is equal to zero.

The approximation of $y(t)$ at the time $t = 2.5$ follows with the Modified Euler Method,

$$\begin{aligned}
\overline{w}_1 &= w_0 + \Delta t f(t_0, w_0) = \\
&= 1 + 0.5\left(1 + (2-1)^2\right) = \\
&= 1 + 0.5 \cdot 2 = 2, \\
w_1 &= w_0 + \frac{\Delta t}{2}\left(f(t_0, w_0) + f(t_1, \overline{w}_1)\right) = \\
&= 1 + 0.25\left(2 + (1 + (2.5-2)^2)\right) = \\
&= 1 + 0.25 \cdot 3.25 = 1.8125.
\end{aligned}$$

The solution $y(t)$ gives the value 1.8333 in $t = 2.5$. So the error in the numerical approximation is 0.0208.

In the second step of the Modified Euler Method we obtain

$$\begin{aligned}
\overline{w}_2 &= 2.5488, \\
w_2 &= 2.4816.
\end{aligned}$$

The solution in $t = 3$ gives $y(3) = 2.5$. So the error in the numerical approximation is 0.0184. A summary of the results is shown in the table below

i	time t	$y(t)$	w_i	error
0	2	1	1	0
1	2.5	1.8333	1.8125	0.0208
2	3	2.5	2.4816	0.0184

2. The local truncation error τ_{n+1} of the Midpoint method is defined by

$$\begin{aligned}
\tau_{n+1} &= \frac{y_{n+1} - z_{n+1}}{\Delta t} \quad \text{in which} \\
z_{n+1} &= y_n + \Delta t \left(f\left(t_n + \frac{\Delta t}{2}, y_n + \frac{\Delta t}{2} f(t_n, y_n) \right) \right).
\end{aligned}$$

Taylor expansion of y_{n+1} about y_n gives

$$y_{n+1} = y_n + \Delta t y_n' + \frac{\Delta t^2}{2} y_n'' + \mathcal{O}(\Delta t^3).$$

Using this in τ_{n+1} gives

$$\tau_{n+1} = y_n' + \frac{\Delta t}{2} y_n'' - f(t_n + \frac{\Delta t}{2}, y_n + \frac{\Delta t}{2} f(t_n, y_n)) + \mathcal{O}(\Delta t^2). \tag{6.1}$$

The second step is to expand $f(t_n + \frac{\Delta t}{2}, y_n + \frac{\Delta t}{2} f(t_n, y_n))$ in a Taylor series about (t_n, y_n). This gives

$$f(t_n + \frac{\Delta t}{2}, y_n + \frac{\Delta t}{2} f(t_n, y_n)) = f(t_n, y_n) + \frac{\Delta t}{2} f_t(t_n, y_n) + \frac{\Delta t}{2} f_y(t_n, y_n) y_n' + \mathcal{O}(\Delta t^2).$$

We will use that $y' = f(t, y)$. The second derivative in time of y is then given by

$$y'' = f_t + f_y y'.$$

So

$$y'_n = f(t_n, y_n) \tag{6.2}$$

and

$$y''_n = (f_t + f_y y')_n = f_t(t_n, y_n) + f_y(t_n, y_n) y'_n. \tag{6.3}$$

Using the Taylor series of $f(t_n + \frac{\Delta t}{2}, y_n + \frac{\Delta t}{2} f(t_n, y_n))$ in (6.1) gives us

$$\tau_{n+1} = y'_n + \frac{\Delta t}{2} y''_n - f(t_n, y_n) - \frac{\Delta t}{2} f_t(t_n, y_n) - \frac{\Delta t}{2} f_y(t_n, y_n) y'_n + \mathcal{O}(\Delta t^2). \tag{6.4}$$

To show that the Midpoint rule is of order two, it is sufficient to see that

$$y'_n + \frac{\Delta t}{2} y''_n - f(t_n, y_n) - \frac{\Delta t}{2} f_t(t_n, y_n) - \frac{\Delta t}{2} f_y(t_n, y_n) y'_n = 0. \tag{6.5}$$

With (6.2) and (6.3) we see that (6.5) holds. The conclusion is that the Midpoint rule is of order two.

3. (i) The Trapezoidal method to solve an initial-value problem is given by

$$w_{n+1} = w_n + \frac{\Delta t}{2} \left(f(t_n, w_n) + f(t_{n+1}, w_{n+1}) \right). \tag{6.6}$$

Note that the Trapezoidal method is an implicit method to approximate the solution of the initial-value problem $y' = f(t, y)$.

The amplification factor $Q(\lambda \Delta t)$ is determined in the following way; consider the test equation $y' = \lambda y$. Using the Trapeziodal method on this test equation gives a $Q(\lambda \Delta t)$ such that

$$w_{n+1} = Q(\lambda \Delta t) w_n.$$

Using (6.6) on the test equation $y' = \lambda y$ gives:

$$w_{n+1} = w_n + \frac{1}{2} \Delta t \left(\lambda w_n + \lambda w_{n+1} \right) \tag{6.7}$$

Restructuring of w_{n+1} and w_n in (6.7) gives

$$\left(1 - \frac{1}{2} \lambda \Delta t \right) w_{n+1} = \left(1 + \frac{1}{2} \lambda \Delta t \right) w_n.$$

So

$$w_{n+1} = \frac{1 + \frac{1}{2} \lambda \Delta t}{1 - \frac{1}{2} \lambda \Delta t} w_n,$$

and thus

$$Q(\lambda \Delta t) = \frac{1 + \frac{1}{2} \lambda \Delta t}{1 - \frac{1}{2} \lambda \Delta t}.$$

(ii) In paragraph 6.4.2 in the book on page 74 we find that the truncation error is given by

$$\tau_{n+1} = \frac{y_{n+1} - Q(\lambda\Delta t)y_n}{\Delta t}.$$

The exact solution of the test equation is

$$y_{n+1} = y_n e^{\lambda\Delta t}.$$

When we combine these results we can obtain that the truncation error of the test equation is determined by the difference of the exponential function and the amplification factor $Q(\lambda\Delta t)$:

$$\tau_{n+1} = \frac{e^{\lambda\Delta t} - Q(\lambda\Delta t)}{\Delta t} y_n. \tag{6.8}$$

We determine the difference between the exponential function and the amplification factor in the following way:

(1) Taylor expansion of $e^{\lambda\Delta t}$

(2) Taylor expansion of $\frac{1}{1-\frac{1}{2}\lambda\Delta t}$ multiplied by $1 + \frac{1}{2}\lambda\Delta t$

(3) Subtract: (1) − (2)

The Taylor expansion of $e^{\lambda\Delta t}$ around 0 is:

$$e^{\lambda\Delta t} = 1 + \lambda\Delta t + \frac{(\lambda\Delta t)^2}{2} + \mathcal{O}(h^3). \tag{6.9}$$

The Taylor expansion of $\frac{1}{1-\frac{1}{2}\lambda\Delta t}$ around 0 is:

$$\frac{1}{1-\frac{1}{2}\lambda\Delta t} = 1 + \frac{1}{2}\lambda\Delta t + \frac{1}{4}\lambda^2\Delta t^2 + \mathcal{O}(\Delta t^3). \tag{6.10}$$

With (6.10) we obtain that $\frac{1+\frac{1}{2}\lambda\Delta t}{1-\frac{1}{2}\lambda\Delta t}$ is equal to

$$\frac{1+\frac{1}{2}\lambda\Delta t}{1-\frac{1}{2}\lambda\Delta t} = 1 + \lambda\Delta t + \frac{1}{2}(\lambda\Delta t)^2 + \mathcal{O}(\Delta t^3). \tag{6.11}$$

To determine $e^{\lambda\Delta t} - Q(\lambda\Delta t)$, we need to substract (6.11) of (6.9).

$$e^{\lambda\Delta t} - Q(\lambda\Delta t) = \mathcal{O}(\Delta t^3). \tag{6.12}$$

The truncation error is now given by using (6.12) in (6.8). We obtain that

$$\tau_{n+1} = \mathcal{O}(\Delta t^2).$$

(iii) In paragraph 6.4.1 on page 71 it is shown that a numerical method is stable if and only if

$$|Q(\lambda\Delta t)| \leq 1.$$

The amplification factor of the Trapezoidal method is

$$Q(\lambda\Delta t) = \frac{1 + \frac{1}{2}\lambda\Delta t}{1 - \frac{1}{2}\lambda\Delta t}.$$

We need to show that if $\lambda \leq 0$ then it follows that $|Q(\lambda\Delta t)| \leq 1$ for all step sizes $\Delta t > 0$.
From $|Q(\lambda\Delta t)| \leq 1$ we know that the following inequalty must hold

$$-1 \leq \frac{1 + \frac{1}{2}\lambda\Delta t}{1 - \frac{1}{2}\lambda\Delta t} \leq 1.$$

Multiplying this inequality with $1 - \frac{1}{2}\lambda\Delta t$ gives

$$-1 + \frac{1}{2}\lambda\Delta t \leq 1 + \frac{1}{2}\lambda\Delta t \leq 1 - \frac{1}{2}\lambda\Delta t. \tag{6.13}$$

Note that $(1 - \frac{1}{2}\lambda\Delta t) \geq 0$, so the sign in the inequality remains the same. To show stability for all Δt we have to check that (6.13) holds for all Δt.
We start with the left-hand inequality (6.13), $-1 + \frac{1}{2}\lambda\Delta t \leq 1 + \frac{1}{2}\lambda\Delta t$. This inequality holds for all $\Delta t > 0$ if $\lambda \leq 0$.
We receive the same result for the right-hand inequality (6.13), $1 + \frac{1}{2}\lambda\Delta t \leq 1 - \frac{1}{2}\lambda\Delta t$. This inequality holds for every $\Delta t > 0$.
The conclusion is that the method is stable for all $\Delta t > 0$ if $\lambda \leq 0$.

4. We consider the nonlinear initial-value problem $y' = f(t, y)$ with

$$f(t, y) = 1 + (t - y)^2.$$

To investigate the stability of Modified Euler at the point $(t = 2, y = 1)$ we need the following results:

- The amplification factor of Modified Euler
- Linearisation of $f(t, y)$.

The amplification factor of Modified Euler is

$$Q(\lambda\Delta t) = 1 + \lambda\Delta t + \frac{1}{2}\lambda^2(\Delta t)^2.$$

See paragraph 6.4.2 from the book.

The linearisation of $f(t, y)$ about the point $(t, y) = (2, 1)$ is

$$\begin{aligned} f(t, y) &\approx f(2, 1) + (t - 2)\frac{\partial f}{\partial t}(2, 1) + (y - 1)\frac{\partial f}{\partial y}(2, 1) = \\ &= 2 + 2(t - 2) - 2(y - 1) = -2y + 2t = -2y + g(t). \end{aligned}$$

The initial-value problem after linearisation is given by

$$y' = -2y + g(t).$$

For stability it is sufficient to consider $y' = -2y$. Note that this is the test equation with $\lambda = -2$. Modified Euler is stable if $|Q(\lambda\Delta t)| \leq 1$ with $\lambda = -2$. So the method is stable if for the step size Δt the following holds

$$-1 \leq 1 - 2\Delta t + 2\Delta t^2 \leq 1.$$

The left-hand inequality $-1 \leq 1 - 2\Delta t + 2\Delta t^2$ holds if

$$\Delta t \geq 0.$$

The right-hand inequality $1 - 2\Delta t + 2\Delta t^2 \leq 1$ if

$$\Delta t \leq 1.$$

So the stability condition is given by

$$0 \leq \Delta t \leq 1.$$

5. (a) The method in this exercise is defined by

$$\begin{aligned} \bar{w}_{n+1} &= w_n + \beta\Delta t f(t_n, w_n) && (6.14)\\ w_{n+1} &= \bar{w}_{n+1} + (1-\beta)\Delta t f(t_n + \beta\Delta t, \bar{w}_{n+1}) && (6.15) \end{aligned}$$

We follow the steps as in determining the local truncation error in the Modified Euler method in paragraph 6.4.2 in the book. We start with replacing the numerical approximation w_n by the exact value y_n, so

$$\begin{aligned} \bar{z}_{n+1} &= y_n + \beta\Delta t f(t_n, y_n) && (6.16)\\ z_{n+1} &= \bar{z}_{n+1} + (1-\beta)\Delta t f(t_n + \beta\Delta t, \bar{z}_{n+1}) && (6.17) \end{aligned}$$

We start with the Taylor expansion of $f(t_n + \beta\Delta t, \bar{z}_{n+1})$ about the point (t_n, y_n):

$$\begin{aligned} f(t_n + \beta\Delta t, \bar{z}_{n+1}) &= f(t_n, y_n) + \beta\Delta t (f_t)_n + (\bar{z}_{n+1} - y(t_n))(f_y)_n + \frac{1}{2}\beta^2\Delta t^2 (f_{tt})_n + \\ &\quad + \frac{1}{2}(\bar{z}_{n+1} - y(t_n))^2 (f_{yy})_n + \beta\Delta t(\bar{z}_{n+1} - y(t_n))(f_{ty})_n. \qquad (6.18) \end{aligned}$$

With equation (6.16) we obtain that (6.18) is equal to

$$f(t_n + \beta\Delta t, \bar{z}_{n+1}) = f(t_n, y_n) + \beta\Delta t(f_t + f f_y) + \mathcal{O}(\Delta t^2). \qquad (6.19)$$

Subsituting (6.19) in (6.15) gives us

$$\begin{aligned} z_{n+1} &= y_n + \beta\Delta t f(t_n, y_n) + (1-\beta)\Delta t f(t_n, y_n) + (1-\beta)\beta\Delta t^2 (f_t + f f_y)_n + \mathcal{O}(\Delta t^3) = \\ &= y_n + \Delta t f(t_n, y_n) + (1-\beta)\beta\Delta t^2 (f_t + f f_y)_n + \mathcal{O}(\Delta t^3). \end{aligned}$$

The Taylor expansion of y_{n+1} is given by

$$y_{n+1} = y_n + \Delta t y'_n + \frac{1}{2}\Delta t^2 y''_n + \mathcal{O}(\Delta t^3).$$

Using the fact that $y' = f(t, y)$ we can obtain that $y_{n+1} - z_{n+1}$ is equal to

$$y_{n+1} - z_{n+1} = \left(\frac{1}{2}\Delta t^2 - \beta(1-\beta)\Delta t^2\right) y_n'' + \mathcal{O}(\Delta t^3) = \mathcal{O}(\Delta t^2).$$

The last equality follows from the fact that $\frac{1}{2} - \beta(1-\beta)$ is not equal to zero for every β. Now it's easy to see that the local truncation error τ_{n+1} is given by

$$\frac{y_{n+1} - z_{n+1}}{\Delta t} = \mathcal{O}(\Delta t).$$

(b) The amplification factor is determined by using the method on the test equation $y' = \lambda y$. With $f(t_n, w_n)$ equal to λw_n in (6.14) and (6.15) we can obtain:

$$\begin{aligned} \bar{w}_{n+1} &= w_n + \beta\lambda\Delta t w_n \\ w_{n+1} &= \bar{w}_{n+1} + (1-\beta)\lambda\Delta t \bar{w}_{n+1}. \end{aligned} \tag{6.20}$$

Substituting $\bar{w}_{n+1}$ in (6.20) gives us

$$\begin{aligned} w_{n+1} &= w_n + \beta\lambda\Delta t w_n + (1-\beta)\lambda\Delta t(w_n + \beta\lambda\Delta t w_n) = \\ &= \left(1 + \beta\lambda\Delta t + (1-\beta)\lambda\Delta t + \beta(1-\beta)(\lambda\Delta t)^2\right) w_n = \\ &= \left(1 + \lambda\Delta t + \beta(1-\beta)(\lambda\Delta t)^2\right) w_n = \\ &= Q(\lambda\Delta t) w_n. \end{aligned}$$

The amplification error is $Q(\lambda\Delta t) = 1 + \lambda\Delta t + \beta(1-\beta)(\lambda\Delta t)^2$.

(c) The nonlinear differential equation $y' = f(y)$ with $f(y) = 2y - 4y^2$ need to be linearized about $y = \frac{1}{2}$. We obtain that the linearisation of $f(y)$ about $y = \frac{1}{2}$ is given by

$$f(y) \approx f(\frac{1}{2}) + (y - \frac{1}{2}) f'(\frac{1}{2}) = -2y + 1.$$

So the linearized differential equation is equal to $y' = -2y + 1$. For stability it is sufficient to check the equation $y' = -2y$. This is the test equation with $\lambda = -2$. The test equation is stable if the following holds:

$$|Q(\lambda\Delta t)| \leq 1 \quad \text{voor } \lambda = -2 \text{ en } \beta = \frac{1}{2}.$$

Using the value of $Q(\lambda\Delta t)$ gives that for step size Δt the inequality must satisfy

$$-1 \leq 1 - 2\Delta t + \Delta t^2 \leq 1.$$

The right inequality holds if $\Delta t \leq 2$. The left inequality also satisfies this condition (and we can't find a more strict condition). The conclusion is that the method is stable in this point if $\Delta t \leq 2$.

6. The vector notation for the Forward Euler method is given by

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \Delta t \mathbf{f}(t_n, \mathbf{w}_n).$$

In our case we have the value $t = 2$,

$$\mathbf{w}_n = \begin{bmatrix} w_1^{(n)} \\ w_2^{(n)} \end{bmatrix}.$$

and

$$\mathbf{f}(\mathbf{w}_n) = \begin{bmatrix} -4w_1^{(n)} - 2w_2^{(n)} + e^{t_n} \\ 3w_1^{(n)} + w_2^{(n)} \end{bmatrix}.$$

With the initial values we know the vector $\mathbf{w}_0$:

$$\mathbf{w}_0 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

So the first step of Forward Euler is

$$\begin{aligned} \mathbf{w}_1 &= \mathbf{w}_0 + \Delta t \mathbf{f}(\mathbf{w}_0) = \\ &= \begin{bmatrix} 0 \\ -1 \end{bmatrix} + 0.1 \cdot \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \\ &= \begin{bmatrix} 0.3 \\ -1.1 \end{bmatrix}. \end{aligned}$$

7. We use Paragraph 6.8 of the book. In this exercise we have a second order initial value problem. We first consider a simplified initial-value problem for a linear differential equation. So we rewrite our differential equation to a system of two first order differential equations:

$$\begin{cases} y_1' &= y_2 \\ y_2' &= -y_1 + 2y_2 + te^t - t \end{cases}$$

with initial values

$$\begin{cases} y_1(0) &= 0 \\ y_2(0) &= 0 \end{cases}.$$

Using Euler Forward with $\Delta t = 0.1$ gives us

$$\begin{aligned} \begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \end{bmatrix} &= \begin{bmatrix} y_1^{(0)} \\ y_2^{(0)} \end{bmatrix} + 0.1 \cdot \begin{bmatrix} y_2^{(0)} \\ -y_1^{(0)} + 2y_2^{(0)} + 0 \cdot e^0 - 0 \end{bmatrix} = \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

The numerical solution is zero after one time step.

The solution of this problem is $\frac{1}{6}t^3e^t - te^t + 2e^t - t - 2$. At $t = 0.1$, the solution has the value $y(0.1) = 8.939 \cdot 10^{-6}$. The error is thus equal to the value of the solution.

8. Stable integration of a system of differential equations, written as

$$\mathbf{y}' = \mathbf{A}\mathbf{y},$$

is only possible if all eigenvalues of the matrix $\mathbf{A}$ have a real part smaller or equal to zero.

In this excersie we consider a mathematical pendulum. The angle as a function of t satisfies:

$$\phi'' + \frac{g}{L}\phi = 0. \tag{6.21}$$

This second order differential equation can also be written as a system of two first order differential equations. So (6.21) is equivalent to

$$\left\{ \begin{array}{rcl} \phi_1' & = & \phi_2 \\ \phi_2' & = & -\frac{g}{L}\phi_1 \end{array} \right. . \tag{6.22}$$

If we define vector ψ as

$$\psi = \left[\begin{array}{c} \phi_1 \\ \phi_2 \end{array} \right],$$

then (6.22) can be written as

$$\psi = \mathbf{S}\psi \quad \text{met} \quad \mathbf{S} = \left[\begin{array}{cc} 0 & 1 \\ -\frac{g}{L} & 0 \end{array} \right].$$

The eigenvalues of $\mathbf{S}$ are calculated by setting

$$|\mathbf{S} - \lambda \mathbf{I}| = 0$$

equal to zero. The eigenvalues are the roots of

$$\lambda^2 + \frac{g}{L} = 0.$$

It follows that $\lambda_{1,2} = \pm i\sqrt{\frac{g}{L}}$. The real part of both eigenvalues is zero, so the system is stable.

9. We consider the system

$$\left\{ \begin{array}{rcl} y_1' & = & 1195y_1 + 1995y_2 \\ y_2' & = & 1197y_1 - 1997y_2 \end{array} \right. ,$$

with initial values

$$\left\{ \begin{array}{rcl} y_1(0) & = & 2 \\ y_2(0) & = & -2 \end{array} \right. .$$

We define the following vector on time $t = n\Delta t$

$$\mathbf{w}_{n+1} = \left[\begin{array}{c} w_1^{(n)} \\ w_2^{(n)} \end{array} \right].$$

Euler Forward is given by

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \Delta t \mathbf{f}(\mathbf{w}_n),$$

and Euler Backward by

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \Delta t \mathbf{f}(\mathbf{w}_{n+1}),$$

with

$$\mathbf{f}(\mathbf{w}_n) = \left[\begin{array}{c} 1195w_1^{(n)} + 1995w_2^{(n)} \\ 1197w_1^{(n)} - 1997w_2^{(n)} \end{array} \right].$$

(a) Choose step size $\Delta t = 0.1$, with $\mathbf{w}_0 = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$ we obtain that for Euler Forward $\mathbf{w}_1$ is given by

$$\begin{aligned} \mathbf{w}_1 &= \mathbf{w}_0 + \Delta t \mathbf{f}(\mathbf{w}_0) \\ &= \begin{bmatrix} 2 \\ -2 \end{bmatrix} + 0.1 \cdot \begin{bmatrix} 1195 \cdot 2 + 1995 \cdot (-2) \\ 1197 \cdot 2 - 1997 \cdot (-2) \end{bmatrix} = \\ &= \begin{bmatrix} 640 \\ 636.8 \end{bmatrix}. \end{aligned}$$

Euler Backward gives us the following linear system that needs to be solved :

$$\begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} + 0.1 \cdot \begin{bmatrix} 1195 & -1995 \\ 1197 & -1997 \end{bmatrix} \begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} \tag{6.23}$$

Reordering in (6.23) gives

$$\begin{aligned} \begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} &= \left[\mathbf{I} - 0.1 \cdot \begin{bmatrix} 1195 & -1995 \\ 1197 & -1997 \end{bmatrix} \right]^{-1} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \\ &= \frac{1}{97} \cdot \begin{bmatrix} 220.7 & -199.5 \\ 119.7 & -118.5 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \begin{bmatrix} 8.23 \\ 4.90 \end{bmatrix} \end{aligned}$$

Notice that the inverse matrix can be determined with Kramer's rule.
The solution in $t = 0.1$ is given by

$$\begin{bmatrix} y_1(0.1) \\ y_2(0.1) \end{bmatrix} = \begin{bmatrix} 8.187 \\ 4.912 \end{bmatrix}.$$

(b) Note that system

$$\begin{cases} y_1' &= 1195 y_1 + 1995 y_2 \\ y_2' &= 1197 y_1 - 1997 y_2 \end{cases},$$

can be written as

$$\mathbf{y}' = \begin{bmatrix} 1195 & -1995 \\ 1197 & -1997 \end{bmatrix} \mathbf{y} = \mathbf{A}\mathbf{y},$$

in which the vector $\mathbf{y}$ is definied as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

As we can see in Paragraph 6.8.3, Euler Forward is stable if the following condition holds for all eigenvalues :

$$|1 + \lambda_i \Delta t| < 1, \quad i = 1, 2, \tag{6.24}$$

with Δt the step size.
We start with determining the eigenvalues of $\mathbf{A}$. So the eigenvalues are determined by setting

$$\begin{vmatrix} 1195 - \lambda & -1995 \\ 1197 & -1997 - \lambda \end{vmatrix} = 0$$

So the following equation should be solved

$$\begin{array}{rcl} \lambda^2 + 802\lambda & + & 1600 = 0 \\ & \Updownarrow & \\ (\lambda + 2)(\lambda & + & 800) = 0 \\ & \Updownarrow & \\ \lambda = -2 \qquad & \text{of} & \qquad \lambda = -800. \end{array}$$

To satisfy (6.24) it is sufficient to sattisfy this for $\lambda = -800$. Note yourself that with $\lambda = -2$ we also automatically fullfill (6.24). So we look for all Δt such that

$$|1 - 800\Delta t| < 1.$$

We obtain that $\Delta t < 0.0025$.

(c) We need to do the same calculations as in (a). For Euler Forward we have

$$\begin{aligned} \mathbf{w}_1 &= \mathbf{w}_0 + \Delta t \mathbf{f}(\mathbf{w}_0) \\ &= \begin{bmatrix} 2 \\ -2 \end{bmatrix} + 0.0001 \cdot \begin{bmatrix} 1195 \cdot 2 + 1995 \cdot (-2) \\ 1197 \cdot 2 - 1997 \cdot (-2) \end{bmatrix} = \\ &= \begin{bmatrix} 2.64 \\ -1.36 \end{bmatrix}. \end{aligned}$$

For Euler Backwards we solve the system

$$\begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} + \begin{bmatrix} 0.1195 & -0.1995 \\ 0.1197 & -0.1997 \end{bmatrix} \begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} \tag{6.25}$$

Note that the solution can be determined by

$$\begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} = 0.9257 \cdot \begin{bmatrix} 1.1997 & -0.1995 \\ 0.1197 & 0.8805 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \begin{bmatrix} 2.59 \\ -1.41 \end{bmatrix}$$

The solution on $t = 0.0001$ is given by

$$\begin{bmatrix} y_1(0.0001) \\ y_2(0.0001) \end{bmatrix} = \begin{bmatrix} 2.61 \\ -1.39 \end{bmatrix}$$

With this step size, both methods give good results. However, Euler Backward is 'more expensive' in calculation time.

Conclusion : With a stiff system, Euler Backward gives a good answer for all step sizes Δt, while Euler Forward only gives good results with (very small) step sizes, that has to satisfy the stability condition.

10. The solution of the differential equation $y' = y - t^2 + 1$ in $t = 0.1$ is approximated in this exercise by Euler Forward and Runge Kutta 4 (RK4). We have already seen the Euler Forward in previous excersises. The RK4 method is explained in Paragraph 6.5 and is definied as follows

$$w_{n+1} = w_n + \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4], \tag{6.26}$$

in which the predictors $k_1, \ldots, k_4$ are given by

$$\begin{aligned} k_1 &= \Delta t f(t_n, w_n) && (6.27)\\ k_2 &= \Delta t f(t_n + \frac{1}{2}\Delta t, w_n + \frac{1}{2}k_1) && (6.28)\\ k_3 &= \Delta t f(t_n + \frac{1}{2}\Delta t, w_n + \frac{1}{2}k_2) && (6.29)\\ k_4 &= \Delta t f(t_n + \Delta t, w_n + k_3) && (6.30) \end{aligned}$$

(i) In the first part we will approximate $y(0.1)$ with the Euler Forward method with step size $h = 0.025$. This means that we need to use four steps of the system

$$w_{n+1} = w_n + \Delta t(w_n - t_n^2 + 1).$$

With initial value $y(0) = w_0 = \frac{1}{2}$, we obtain the following w_i's:

$$\begin{aligned} w_1 &= w_0 + \Delta t(w_0 - 0^2 + 1) = \\ &= \frac{1}{2} + 0.025 \cdot 1\frac{1}{2} = 0.5375, \\ w_2 &= w_1 + 0.025(w_1 - 0.025^2 + 1) = 0.5759219 \\ w_3 &= w_2 + 0.025(w_2 - 0.050^2 + 1) = 0.6152574 \\ w_4 &= w_3 + 0.025(w_3 - 0.075^2 + 1) = 0.6554982 \end{aligned}$$

The approximation of $y(0.1)$ is given by $w_4 = 0.6554982$.

(ii)In this part we are going to approximate $y(0.1)$ with the RK4 method, explained by (6.26) t/m (6.30), with step size $\Delta t = 0.1$. Note that in (6.27) t/m (6.30) $f(t_n, w_n) = w_n - t_n^2 + 1$.
We start with calculating the four predictors, in which $y(0) = w_0 = \frac{1}{2}$. So

$$\begin{aligned} k_1 &= \Delta t f(t_0, w_0) = \\ &= \Delta t(w_0 - t_0^2 + 1) = \\ &= 0.1(\frac{1}{2} - 0^2 + 1) = 0.15 \\ k_2 &= 0.1(w_0 + \frac{1}{2}k_1 - 0.05^2 + 1) = 0.15725 \\ k_3 &= 0.1(w_0 + \frac{1}{2}k_2 - 0.05^2 + 1) = 0.1576125 \\ k_4 &= 0.1(w_0 + k_3 - 0.1^2 + 1) = 0.1647613. \end{aligned}$$

So for w_1 we obtain:

$$\begin{aligned} w_1 &= w_0 + \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4] = \\ &= \frac{1}{2} + \frac{1}{6} \cdot 0.9444863 = 0.6574144. \end{aligned}$$

With RK4, the approximaton of $y(0.1)$ is given by $w_1 = 0.6574144$.

The solution of the differential equation is $y(t) = -\frac{1}{2}e^t + t^2 + 2t + 1$. At $t = 0.1$ is the solution thus given by $y(0.1) = 0.6574145$.

Conclusion: We obtain that the approximations of both methods are good and that the work we need to do is almost the same. While the error in the approximation of RK4 is equal to 10^{-7} and of Euler Forward is equal to 2×10^{-3}, we prefer RK4.

11. To determine the order of the error, we use the formula given in Paragraph 6.6.2 of the book: Error estimate if p is unknown:

$$\frac{w_{N/2}^{2\Delta t} - w_{N/4}^{4\Delta t}}{w_N^{\Delta t} - w_{N/2}^{2\Delta t}} = 2^p.$$

In which $w_{N/2}^{2\Delta t}$ and $w_N^{\Delta t}$ denote the approximation for $y(t) = y(N\delta t)$ using $N/2$ time steps of length $2\Delta t$ and using N time steps of length Δt, respectively. In the first column we obtain

$$\frac{0.750686 - 0.752790}{0.750180 - 0.750686} = 4.1581 \approx 2^2,$$

so the order of Method 1 is 2. For the second column we obtain

$$\frac{0.730912 - 0.710791}{0.740587 - 0.730912} = 2.0797 \approx 2^1,$$

so the order of Method 2 is 1.

Chapter 7

The finite-difference method for boundary-value problems

7.1 Solutions

1. The condition number $\kappa(\mathbf{A})$ of a symmetric matrix $\mathbf{A}$ is the quotient of the largest and smallest absolute eigenvalue of $\mathbf{A}$ as in Paragraph 7.3 of the book.

   The matrix $\mathbf{A}_1$ is de symmetric matrix

   $$\mathbf{A}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

   The eigenvalues of $\mathbf{A}_1$ can be easily calculated by

   $$\begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1). \tag{7.1}$$

   Setting (7.1) equal to zero gives us the eigenvalues $\lambda_{\max} = 3$ and $\lambda_{\min} = 1$. It follows that $k(\mathbf{A}_1) = 3$.

   The matrix $\mathbf{A}_2$ is the symmetric matrix

   $$\mathbf{A}_2 = \begin{bmatrix} 100 & 99 \\ 99 & 100 \end{bmatrix}.$$

   The eigenvalues of $\mathbf{A}_2$ are $\lambda_{\max} = 199$ and $\lambda_{\min} = 1$. Thus the condition number of $\mathbf{A}_2$ is equal to $k(\mathbf{A}_2) = 199$.

   The solution of the system $\mathbf{A}_2\mathbf{w} = \mathbf{f}$ with $\mathbf{f} = [199, 199]^T$ is easy to see, namely $\mathbf{x} = [1, 1]^T$.

   We take $\Delta\mathbf{f} = [1, 0]^T$ as an error in the right-hand side. In Paragraph 7.3 we obtain the following approximation for the relative error:

   $$\|\Delta\mathbf{w}\| \leq \kappa(\mathbf{A}_2)\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}\|\mathbf{w}\|. \tag{7.2}$$

   The norm of the vector is $\mathbf{w} = [1, 1]^T$ is

   $$\|\mathbf{w}\| = \sqrt{\frac{1}{2}1^2 + \frac{1}{2}1^2} = 1.$$

In the same way we obtain

$$\|\mathbf{f}\| = 199 \quad \text{en} \quad \|\Delta\mathbf{f}\| = \frac{1}{\sqrt{2}}.$$

Substituting in (7.2) gives us

$$\|\Delta\mathbf{w}\| \leq \frac{1}{\sqrt{2}} \approx 0.7071.$$

We will now determine $\Delta\mathbf{w}$ and then we will compare $\|\Delta\mathbf{w}\|$ with the estimated upper bound. If the right-hand side $\mathbf{f}$ has error $\Delta\mathbf{f} = [1, 0]^T$, then we can determine $\Delta\mathbf{w}$ with

$$\mathbf{A}_2(\mathbf{w} + \Delta\mathbf{w}) = (\mathbf{f} + \Delta\mathbf{f}), \tag{7.3}$$

because $\mathbf{w}$ is known. Reorder known and unknown in (7.3) gives

$$\mathbf{A}_2\Delta\mathbf{w} = (\mathbf{f} + \Delta\mathbf{f}) - \mathbf{A}_2\mathbf{w}.$$

Wit Gauss' elimination it follows that $\Delta\mathbf{w}$ is equal to

$$\Delta\mathbf{w} = \begin{bmatrix} 0.5025 \\ -0.4975 \end{bmatrix}.$$

The norm of $\Delta\mathbf{w}$ is

$$\|\Delta\mathbf{w}\| = \sqrt{\frac{1}{2}(0.5025)^2 + \frac{1}{2}(-0.4975)^2} \approx 0.5000.$$

So $\|\Delta\mathbf{x}\|$ is indeed smaller than 0.7071, as predicted.

2. The boundary-value probem can be rewritten as

$$\begin{cases} -y''(x) + 2y(x) = 2x, x \in [0, 1], \\ y(0) = y(1) = 0. \end{cases}$$

   (a) The interval $[0, 1]$ can be divided into $n + 1$ equidistant subintervals with length $\Delta x = \frac{1}{n+1}$. The nodes are given by $x_j = j\Delta x$ for $j = 0, 1, \ldots, n + 1$. The numerical approximation of $y_j = y(x_j)$ is denoted by w_j, which is computed using the differential equation in x_j :

$$-y_j'' + 2y_j = 2x_j \quad \text{voor} \quad 1 \leq j \leq n. \tag{7.4}$$

   The finite-difference method approximates the second derivative in (7.4) by a central-difference formula to obtain

$$-\frac{w_{j-1} - 2w_j + w_{j+1}}{\Delta x^2} + 2w_j = 2j\Delta x. \tag{7.5}$$

   When $j = 1$, equation (7.5) is given by

$$-\frac{w_0 - 2w_1 + w_2}{\Delta x^2} + 2w_1 = 2\Delta x, \tag{7.6}$$

in which $w_0 = 0$ because of the boundary condition $y(0) = 0$. For $j = 1$ we have the following equation

$$\frac{2w_1 - w_2}{\Delta x^2} + 2w_1 = 2\Delta x. \tag{7.7}$$

(7.5) stays the same for all $1 < j < n$. For $j = n$ we obtain

$$-\frac{w_{n-1} - 2w_n}{\Delta x^2} + 2w_n = 2n\Delta x. \tag{7.8}$$

We can solve this system of equations with m equations and m unknown values. The system consists of linear equations, so we can rewrite it in the form $\mathbf{Aw} = \mathbf{f}$. With

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix},$$

we obtain

$$\mathbf{A} = \begin{bmatrix} \frac{2}{\Delta x^2} + 2 & -\frac{1}{\Delta x^2} & & & \\ -\frac{1}{\Delta x^2} & \frac{2}{\Delta x^2} + 2 & -\frac{1}{\Delta x^2} & & \\ \ddots & \ddots & \ddots & & \\ & & & & \\ & & & & \\ & & & -\frac{1}{\Delta x^2} & \frac{2}{\Delta x^2} + 2 \end{bmatrix}$$

and

$$\mathbf{f} = \begin{bmatrix} 2\Delta x \\ 4\Delta x \\ \vdots \\ 2(n-1)\Delta x \\ 2n\Delta x \end{bmatrix}.$$

(b) Note that $\mathbf{A}$ is symmetric. This means that all eigenvalues of $\mathbf{A}$ are real values. To determine the largest and smallest eigenvalues of $\mathbf{A}$ we use Gershgorin's theorem.

Theorem 7.1.1 (Gershgorin circle theorem) *The eigenvalues of a general $n \times n$ matrix $\mathbf{A}$ are located in the complex plane in the union of circles*

$$|z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|,$$

where $z \in \mathbb{C}$.

Using Gershgorin on row k of $\mathbf{A}$, for $1 < k < n$, gives that the eigenvalues are in the union of circles

$$|z - \frac{2}{\Delta x^2} - 2| \leq \frac{2}{\Delta x^2}.$$

For rows 1 and n of $\mathbf{A}$

$$|z - \frac{2}{\Delta x^2} - 2| \leq \frac{1}{\Delta x^2}.$$

So all eigenvalues of $\mathbf{A}$ lie in the union of circles

$$|z - \frac{2}{\Delta x^2} - 2| \leq \frac{2}{\Delta x^2} \quad \text{and} \tag{7.9}$$

$$|z - \frac{2}{\Delta x^2} - 2| \leq \frac{1}{\Delta x^2}, \tag{7.10}$$

in the complex plane.

We note that the smallest (real) eigenvalue is equal to $\lambda_{\min} = 2$ and the biggest $\lambda_{\max} = 2 + \frac{4}{\Delta x^2}$.

(c) Since $\mathbf{A}$ is symmetric, the condition number of $\mathbf{A}$ is definied as

$$\kappa(\mathbf{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

In Paragraph 7.3 it is shown that

$$\frac{\|\Delta \mathbf{w}\|}{\|\mathbf{w}\|} \leq \kappa(\mathbf{A}) \frac{\|\Delta \mathbf{f}\|}{\|\mathbf{f}\|}.$$

Together with

$$\frac{\|\Delta \mathbf{f}\|}{\|\mathbf{f}\|} \leq 10^{-4},$$

we obtain

$$\frac{\|\Delta \mathbf{w}\|}{\|\mathbf{w}\|} \leq \kappa(\mathbf{A}) \frac{\|\Delta \mathbf{f}\|}{\|\mathbf{f}\|} \leq \frac{2 + \frac{4}{\Delta x^2}}{2} \cdot 10^{-4} = \left(1 + \frac{2}{\Delta x^2}\right) \cdot 10^{-4}.$$

3. The differential equation

$$-y''(x) = \sin x \quad x \in [0, \pi],$$

integrated two times gives the general solution

$$y(x) = \sin x + c_1 x + c_2. \tag{7.11}$$

Using the boundary conditions $y(0) = y(\pi) = 0$ gives

$$\begin{aligned} y(0) &= c_2 = 0 \\ y(\pi) &= sin\pi + c_1 \pi + c_2 = 0, \end{aligned}$$

which gives us $c_1 = c_2 = 0$. The solution of the boundary-value problem is thus $y(x) = \sin x$.

The numerical solution is determined by taking $N = 2$ and approximating the second derrivative with central differences. With $N = 2$, we divide the area in 3 equivalent intervals. The numerical approximation consists of the following points

$$\begin{aligned} y(0) &\approx w_0 = 0 \quad \text{because of the boundary value,} \\ y(\frac{\pi}{3}) &\approx w_1 \\ y(\frac{2\pi}{3}) &\approx w_2 \\ y(\pi) &\approx w_3 = 0 \quad \text{because of the boundary value.} \end{aligned}$$

So we need to solve the system

$$\begin{cases} 2w_1 - w_2 &= \sin(\frac{\pi}{3})\Delta x^2 \\ -w_1 + 2w_2 &= \sin(\frac{2\pi}{3})\Delta x^2 \end{cases},$$

in which $\Delta x = \frac{\pi}{3}$. With Gauss elimination the solution is given by

$$\begin{aligned} u_1 &= 0.9497 \\ u_2 &= 0.9497. \end{aligned}$$

The global error, $y_j - w_j$ is equal to -0.0837 for $j = 1, 2$.

4. We obtain the boundary-value problem

$$\begin{cases} -y''(x) + y(x) = 0, x \in [0, 1] \\ y(0) = 0 \\ y'(1) = 0 \end{cases},$$

The grid points are given by $x_j = jh$ with $\Delta x = \frac{2}{7}$ for $j = 0, 1, \ldots, 4$. The differential equation in the point x_j is given by

$$-y''(x_j) + y(x_j) = 0.$$

Discretization of the equation above gives

$$\frac{-w_{j-1} + 2w_j - w_{j+1}}{\Delta x^2} + w_j = 0, \quad 1 \leq j \leq 3 \tag{7.12}$$

in which $w_j \approx y(x_j)$.

For $j = 1$ in (7.12) we obtain with the boundary value $w_0 = 1$

$$\frac{2w_1 - w_2}{\Delta x^2} + w_1 = \frac{1}{\Delta x^2}.$$

The boundary condition $y'(1) = 0$ is discretizated by :

$$y'(1) = 0 \rightsquigarrow \frac{w_4 - w_3}{\Delta x} = 0.$$

This results for $j = 3$ in (7.12) in

$$\frac{-w_2 + w_3}{\Delta x^2} + w_3 = 0.$$

If we define the vector of unknowns by

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix},$$

and taking the right-hand equal to the zero vector, the disretization of this boundary problem is given by $\mathbf{A}\mathbf{w} = \mathbf{f}$, with

$$\mathbf{A} = \frac{1}{\Delta x^2} \begin{bmatrix} 2+\Delta x^2 & -1 & 0 \\ -1 & 2+\Delta x^2 & -1 \\ 0 & -1 & 1+\Delta x^2 \end{bmatrix}.$$

The matrix $\mathbf{A}$ is thus symmetric, which means that all eigenvalues are real.

Using Gershgorin's circle theorem gives us the fact that the eigenvalues are in the union of circles $C = \bigcup_{i=1}^{3} C_i$, with

$$\begin{aligned} C_1 &: \left| z - \left(\frac{2}{\Delta x^2} + 1 \right) \right| \leq \frac{1}{\Delta x^2} \\ C_2 &: \left| z - \left(\frac{2}{\Delta x^2} + 1 \right) \right| \leq \frac{2}{\Delta x^2} \quad \text{en} \\ C_3 &: \left| z - \left(\frac{1}{\Delta x^2} + 1 \right) \right| \leq \frac{1}{\Delta x^2}. \end{aligned}$$

All eigenvalues lie in the union of C_2 and C_3. All (real) eigenvalues in the circles are greater than 1, so positive. We conclude that $\mathbf{A}$ has only positive real eigenvalues.

Chapter 8

The instationary heat equation*

8.1 Solutions