Solutions to exercise set 11

Norwegian University of Science and Technology
Department of Mathematics

1

a) Euler's method for this particular initial value problem takes the form

$$x_{k+1} = x_k + h\sin(t_k^2 + x_k),$$

with initial values

$$x_0 = x(t_0) = 0 \qquad \text{and} \qquad t_0 = 0.$$

Here

$$t_k = t_0 + kh, \quad k = 0, 1, 2, \dots$$

With a step size of $h = 1/4$, we therefore obtain the (rounded) iterates

$$\begin{aligned}
x_1 &= 0 + \frac{1}{4}\sin(0) = 0,\\
x_2 &= 0 + \frac{1}{4}\sin(1/16) = 0.015615,\\
x_3 &= 0.015615 + \frac{1}{4}\sin(1/4 + 0.015615) = 0.081240,\\
x_4 &= 0.081240 + \frac{1}{4}\sin(9/16 + 0.081240) = 0.231288,\\
x_5 &= 0.231288 + \frac{1}{4}\sin(1 + 0.231288) = 0.467018,\\
x_6 &= 0.467018 + \frac{1}{4}\sin(25/16 + 0.467018) = 0.691173,\\
x_7 &= 0.691173 + \frac{1}{4}\sin(9/4 + 0.691173) = 0.740943,\\
x_8 &= 0.740943 + \frac{1}{4}\sin(49/16 + 0.740943) = 0.587299,
\end{aligned}$$

The value $x_8 = 0.587299$ is the desired approximation of $x(2)$.

b) The improved Euler method for this problem takes the form

$$\begin{aligned}
\hat{x}_{k+1} &= x_k + h\sin(t_k^2 + x_k),\\
x_{k+1} &= x_k + \frac{h}{2}\big(\sin(t_k^2 + x_k) + \sin((t_k + h)^2 + \hat{x}_{k+1})\big).
\end{aligned}$$

With a step size of $h = 1/2$, we obtain the (rounded) iterates

$$\begin{aligned}
\hat{x}_1 &= 0 + \frac{1}{2}\sin(0+0) = 0,\\
x_1 &= 0 + \frac{1}{4}\big(\sin(0+0) + \sin(1/4+0)\big) = 0.061851,\\
\hat{x}_2 &= 0.061851 + \frac{1}{2}\sin(1/4 + 0.061851) = 0.215261,\\
x_2 &= 0.061851 + \frac{1}{4}\big(\sin(1/4+0.061851) + \sin(1+0.215261)\big) = 0.372921,\\
\hat{x}_3 &= 0.372921 + \frac{1}{2}\sin(1 + 0.372921) = 0.863165,\\
x_3 &= 0.372921 + \frac{1}{4}\big(\sin(1+0.372921) + \sin(9/4+0.863165)\big) = 0.625149,\\
\hat{x}_4 &= 0.625149 + \frac{1}{2}\sin(9/4 + 0.625149) = 0.756800,\\
x_4 &= 0.756800 + \frac{1}{4}\big(\sin(9/4+0.625149) + \sin(4+0.756800)\big) = 0.441221.
\end{aligned}$$

The value $x_4 = 0.441221$ is the desired approximation of $x(2)$.

c) The classical Runge–Kutta method uses the iteration:

$$\begin{aligned}
K_1 &= hf(t_k, x_k),\\
K_2 &= hf(t_k + \frac{h}{2}, x_k + \frac{1}{2}K_1),\\
K_3 &= hf(t_k + \frac{h}{2}, x_k + \frac{1}{2}K_2),\\
K_4 &= hf(t_k + h, x_k + K_3),
\end{aligned}$$

and

$$x_{k+1} = x_k + \frac{1}{6}\Big(K_1 + 2K_2 + 2K_3 + K_4\Big).$$

In this particular example, we obtain with a step size of $h = 1$ the following iterates:

- First step:

$$\begin{aligned}
K_1 &= 0,\\
K_2 &= 0.247404,\\
K_3 &= 0.365064,\\
K_4 &= 0.978912,
\end{aligned}$$

  and

$$x_1 = 0.367308.$$

- Second step:

$$\begin{aligned}
K_1 &= 0.979368,\\
K_2 &= 0.034594,\\
K_3 &= 0.485546,\\
K_4 &= -0.990151,
\end{aligned}$$

  and

$$x_2 = 0.538891,$$

  which is also the desired approximation of $x(2)$.

2 In the k-th step of the implicit Euler method, we have to solve the equation

$$x_{k+1} = x_k + hf(t_k + h, x_{k+1}).$$

Defining

$$g_k(z) := x_k + hf(t_k + h, z) - z,$$

this means that x_{k+1} solves the equation

$$g_k(x_{k+1}) = 0.$$

By assumption, the function g_k is continuous and strictly decreasing (because $f(t_k + h, \cdot)$ is decreasing and $-z$ is strictly decreasing). Moreover, for $z > 0$ we have

$$g_k(z) = x_k + hf(t_k + h, z) - z \le x_k + hf(t_k + h, 0) - z = g_k(0) - z,$$

showing that $g_k(z) \to -\infty$ for $z \to +\infty$. Similarly, for $z < 0$ we have

$$g_k(z) \ge g_k(0) - z \to_{z \to -\infty} +\infty.$$

Since g_k is continuous and strictly decreasing, it follows that the equation $g_k(z) = 0$ has a unique solution, which is what we wanted to show.

3 **a)** Here the iteration (with $h = 1$) is

$$x_{k+1} = x_k + (-3x_k - e^{x_k}) = -2x_k - e^{x_k}.$$

Thus we obtain

$$\begin{aligned} x_0 &= 1, \\ x_1 &= -2 - e \approx -4.718282, \\ x_2 &\approx 9.427633, \\ x_3 &\approx -12446. \end{aligned}$$

b) Here the iteration is

$$\begin{aligned} \hat{x}_{k+1} &= x_k + (-3x_k - e^{x_k}), \\ x_{k+1} &= x_k + \frac{1}{2}\big(-3x_k - e^{x_k} - 3\hat{x}_{k+1} - e^{\hat{x}_{k+1}}\big). \end{aligned}$$

We obtain

$$\begin{aligned} \hat{x}_1 &\approx -4.718282, & x_1 &\approx 5.213817, \\ \hat{x}_2 &\approx -194.222, & x_2 &\approx 196.829, \\ \hat{x}_3 &\approx -3 \cdot 10^{85}, & x_3 &\approx 3 \cdot 10^{85}. \end{aligned}$$

c) The implicit Euler method requires the solution of the equation

$$x_{k+1} = x_k - 3x_{k+1} - e^{x_{k+1}}$$

with respect to x_{k+1} in each step. Setting

$$g(z) := 4z + e^z,$$

this means that x_{k+1} solves the equation

$$g(x_{k+1}) - x_k = 0.$$

We solve this equation using two steps of Newton's method with a suitable initial guess, for instance the result with the explicit Euler method. We then have (for each k) the iteration

$$x_{k+1,0} := x_k - 3x_k - e^{x_k}$$

(this is the explicit Euler step) and

$$x_{k+1,\ell+1} := x_{k+1,\ell} - \frac{g(x_{k+1,\ell}) - x_k}{g'(x_{k+1,\ell})}$$

(this is a Newton step), or, using the definition of g,

$$x_{k+1,\ell+1} := x_{k+1,\ell} - \frac{4x_{k+1,\ell} + e^{x_{k+1,\ell}} - x_k}{4 + e^{x_{k+1,\ell}}}.$$

In this exercise, we stop the Newton iteration after two steps, that is, we set

$$x_{k+1} := x_{k+1,2}.$$

We therefore obtain:

- First Euler step:

  $$x_{1,0} := -2 - e \approx -4.718282,$$

  and then

  $$\begin{aligned} x_{1,1} &= x_{1,0} - \frac{4x_{1,0} + e^{x_{1,0}} - 1}{4 + e^{x_{1,0}}} \approx 0.23670, \\ x_{1,2} &= x_{1,1} - \frac{4x_{1,1} + e^{x_{1,1}} - 1}{4 + e^{x_{1,1}}} \approx 0.00624. \end{aligned}$$

  Thus

  $$x_1 = x_{1,2} \approx 0.00624.$$

- Second Euler step:

  $$x_{2,0} := -2x_1 - e^{x_1} \approx -1.01873,$$

  and then

  $$\begin{aligned} x_{2,1} &= x_{2,0} - \frac{4x_{2,0} + e^{x_{2,0}} - 0.00624}{4 + e^{x_{2,0}}} \approx -0.16570, \\ x_{2,2} &= x_{2,1} - \frac{4x_{2,1} + e^{x_{2,1}} - 0.00624}{4 + e^{x_{2,1}}} \approx -0.20248. \end{aligned}$$

  Thus

  $$x_2 = x_{2,2} \approx -0.20248.$$

- Third Euler step:

  $$x_{3,0} := -2x_2 - e^{x_2} \approx -0.41176,$$

  and then

  $$\begin{aligned} x_{3,1} &= x_{3,0} - \frac{4x_{3,0} + e^{x_{3,0}} + 0.20248}{4 + e^{x_{3,0}}} \approx -0.24402, \\ x_{3,2} &= x_{3,1} - \frac{4x_{3,1} + e^{x_{3,1}} + 0.20248}{4 + e^{x_{3,1}}} \approx -0.24608. \end{aligned}$$

  Thus

  $$x_3 = x_{3,2} \approx -0.24608.$$

Some additional remarks are here in order: Writing the differential equation in this exercise as $x' = f(x)$ with $f(x) = -3x - e^x$, we see that the function f is strictly decreasing with a unique zero $\hat{x} \in \mathbb{R}$ (with $\hat{x} \approx -0.25763$). In this case it is easy to show that every solution of the differential equation (regardless of the initial value) has the property that $x(t) \to \hat{x}$ for $t \to \infty$. Moreover this convergence is monotoneous: if $x(0) > \hat{x}$, then $x(t)$ is strictly decreasing to $\hat{x}$; if $x(0) < \hat{x}$, then $x(t)$ is strictly increasing to $\hat{x}$. If we compare this expected theoretical behaviour with the actual behaviour of the numerical solution, we see that:

- The solution with the Euler method is not monotoneous but instead jumps in each step over $\hat{x}$. Moreover, after three iterations it is quite far from the actual solution; after five iterations it is numerically equal to $-\infty$.
- The solution with the improved Euler method is increasing instead of decreasing, and after three iterations the result is (obviously) way off.
- The result of the implicit Euler method (using only two Newton steps) matches exactly the expected behaviour. The iterates are monotoneous and appear to be converging to the correct value. The added work of having to solve a non-linear system at each step, since the method is implicit, appears to be well worth it in this case.

$\boxed{4}$ The Newton form of the interpolation polynomial through the points (t_k, f_k), (t_{k-1}, f_{k-1}), and (t_{k-2}, f_{k-2}) (which can be computed using divided differences) is

$$p(\tau) = f_k + (\tau - t_k)\frac{f_k - f_{k-1}}{h} + (\tau - t_k)(\tau - t_{k-1})\frac{f_k - 2f_{k-1} + f_{k-2}}{2h^2}.$$

Thus

$$\begin{aligned}
&\int_{t_k}^{t_k+h} p(\tau)\,d\tau \\
&= f_k h + \frac{f_k - f_{k-1}}{h}\int_{t_k}^{t_k+h}(\tau - t_k)\,d\tau + \frac{f_k - 2f_{k-1} + f_{k-2}}{2h^2}\int_{t_k}^{t_k+h}(\tau - t_k)(\tau - t_{k-1})\,d\tau \\
&\qquad = f_k h + \frac{f_k - f_{k-1}}{h}\frac{h^2}{2} + \frac{f_k - 2f_{k-1} + f_{k-2}}{2h^2}\frac{5h^3}{6}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
x_{k+1} &:= x_k + \int_{t_k}^{t_k+h} p(\tau)\,d\tau \\
&= x_k + f_k h + \frac{f_k - f_{k-1}}{2}h + \frac{5f_k - 10f_{k-1} + 5f_{k-2}}{12}h \\
&= x_k + \frac{23}{12}f_k h - \frac{16}{12}f_{k-1}h + \frac{5}{12}f_{k-2}h.
\end{aligned}$$

$\boxed{5}$ We first apply the improved Euler method in order to computing the first iterate x_1. The general formula for this method is (for an autonomous ODE)

$$x_{k+1} = x_k + \frac{h}{2}\big(f(x_k) + f(x_k + hf(x_k))\big).$$

With $h = 1$ and $x_0 = 1$ we obtain first

$$f(x_0 + f(x_0)) = f(1 + (1 - 1/2)) = f(3/2) = \frac{3}{2} - \frac{9}{8} = \frac{3}{8}$$

and then

$$x_1 = 1 + \frac{1}{2}\Big(\frac{1}{2} + \frac{3}{8}\Big) = \frac{23}{16} = 1.4375.$$

As a consequence,

$$f_1 = f(x_1) \approx 0.4043.$$

a) For the Adams–Bashforth method of second order

$$x_2 = x_1 + \frac{3}{2}f_1 - \frac{1}{2}f_0 = 1.4375 + \frac{3}{2}0.4043 - \frac{1}{2}\frac{1}{2} = 1.7939$$

and

$$f_2 = f(1.7939) = 0.1848;$$

then

$$x_3 = x_2 + \frac{3}{2}f_2 - \frac{1}{2}f_1 = 1.7939 + \frac{3}{2}0.1848 - \frac{1}{2}0.4043 = 1.8690$$

and

$$f_2 = f(1.8690) = 0.1224;$$

finally,

$$x_4 = x_3 + \frac{3}{2}f_3 - \frac{1}{2}f_2 = 1.8690 + \frac{3}{2}0.1848 - \frac{1}{2}0.1224 = 1.9602$$

and

$$f_4 = f(1.9602) = 0.0390.$$

b) For the Adams–Moulton method of second order we have for this particular function with step size $h = 1$

$$x_{k+1} = x_k + \frac{1}{2}f_k + \frac{1}{2}\Big(x_{k+1} - \frac{x_{k+1}^2}{2}\Big)$$

or

$$\frac{1}{4}x_{k+1}^2 + \frac{1}{2}x_{k+1} - x_k - \frac{1}{2}f_k = 0,$$

which means that

$$x_{k+1} = -1 \pm \sqrt{1 + 4x_k + 2f_k}.$$

Because $x_0 = 1 > -1$, it makes sense to choose the positive root[1] in this expression and thus set

$$x_{k+1} = -1 + \sqrt{1 + 4x_k + 2f_k}.$$

[1] More precisely, it is easy to see that the differential equation has the two constant solutions $x = 0$ and $x = 2$. The solution of an initial value problem with $0 < x(0) < 2$ therefore satisfies $0 < x(t) < 2$ for all $t > 0$. In addition, because $f(x) > 0$ for $0 < x < 2$, it follows that in this case the solution is a strictly increasing function. Now note that the iteration can be explicitly written as $x_{k+1} = -1 + \sqrt{1 + 6x_k - x_k^2}$. One can show that this iteration has the same properties (that is, it is increasing and remains smaller than 2) for all starting values between 0 and 2.

We obtain

$$\begin{aligned} x_2 &\approx 1.7493, & f_2 &\approx 0.2193,\\ x_3 &\approx 1.9044, & f_3 &\approx 0.0910,\\ x_4 &\approx 1.9664, & f_4 &\approx 0.0330. \end{aligned}$$

(Note that here it would have been possible (and sensible) to compute also x_1 with the Adams–Moulton methd.)

c) The Adams–Bashforth–Moulton method of second order is defined by the iteration

$$\begin{aligned} \hat{x}_{k+1} &= x_k + h\Big(\frac{3}{2}f_k - \frac{1}{2}f_{k-1}\Big),\\ \hat{f}_{k+1} &= f(t_{k+1}, \hat{x}_{k+1}),\\ x_{k+1} &= x_k + h\Big(\frac{1}{2}\hat{f}_{k+1} + \frac{1}{2}f_k\Big). \end{aligned}$$

Here we obtain in the first step

$$\begin{aligned} \hat{x}_2 &= x_1 + \frac{3}{2}f_1 - \frac{1}{2}f_0 \approx 1.7939,\\ \hat{f}_2 &= f(\hat{x}_2) \approx 0.1848,\\ x_2 &= x_1 + \frac{1}{2}f_1 + \frac{1}{2}\hat{f}_2 \approx 1.7321,\\ f_2 &= f(x_2) \approx 0.2320, \end{aligned}$$

then

$$\begin{aligned} \hat{x}_3 &= x_2 + \frac{3}{2}f_2 - \frac{1}{2}f_1 \approx 1.8780,\\ \hat{f}_3 &= f(\hat{x}_3) \approx 0.1146,\\ x_3 &= x_2 + \frac{1}{2}f_2 + \frac{1}{2}\hat{f}_3 \approx 1.9054,\\ f_3 &= f(x_3) \approx 0.0902, \end{aligned}$$

and finally

$$\begin{aligned} \hat{x}_4 &= x_3 + \frac{3}{2}f_3 - \frac{1}{2}f_2 \approx 1.9246,\\ \hat{f}_4 &= f(\hat{x}_3) \approx 0.0726,\\ x_4 &= x_3 + \frac{1}{2}f_3 + \frac{1}{2}\hat{f}_4 \approx 1.9867,\\ f_4 &= f(x_4) \approx 0.0132. \end{aligned}$$

$\boxed{6}$ **a)** The equation can be rewritten as the system of first order equations

$$\begin{aligned} x_1' &= x_2, & x_1(0) &= 0,\\ x_2' &= -\sin(x_1) + x_2, & x_2(0) &= 1. \end{aligned}$$

b) To avoid double subscripts we here denote by $x_i^{(k)}$ the approximation from the method to the component function $x_i(t)$ at time $t_k = 0 + kh$ and let $\mathbf{X}_k = [x_1^{(k)}, x_2^{(k)}]^T$. The classical Runge–Kutta method with step size $h = 1/2$ yields:

- First step:

$$\mathbf{K}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$
$$\mathbf{K}_2 = \begin{bmatrix} 1.25 \\ 1.0026 \end{bmatrix},$$
$$\mathbf{K}_3 = \begin{bmatrix} 1.2506 \\ 0.9432 \end{bmatrix},$$
$$\mathbf{K}_4 = \begin{bmatrix} 1.4716 \\ 0.8862 \end{bmatrix},$$

and

$$\mathbf{X}_1 = \begin{bmatrix} 0.6227 \\ 14815 \end{bmatrix}.$$

- Second step:

$$\mathbf{K}_1 = \begin{bmatrix} 1.4815 \\ 0.8982 \end{bmatrix},$$
$$\mathbf{K}_2 = \begin{bmatrix} 1.7060 \\ 0.8683 \end{bmatrix},$$
$$\mathbf{K}_3 = \begin{bmatrix} 1.6986 \\ 0.8315 \end{bmatrix},$$
$$\mathbf{K}_4 = \begin{bmatrix} 1.8972 \\ 0.9021 \end{bmatrix},$$

and

$$\mathbf{X}_2 = \begin{bmatrix} 1.4717 \\ 1.9148 \end{bmatrix}.$$

The first entry of $\mathbf{X}_2$ is an approximation of $x(1)$, that is, $x(1) \approx 1.4717$.

$\boxed{7}$ **a)** We follow the normal procedure for this. Defining a vector $\mathbf{X} = [x_0, x_1, x_2, x_3, x_4, x_5]^T$ where $x_0 = t$, $x_1 = x$, $x_2 = x'$, $x_3 = y$, $x_4 = y'$ and $x_5 = y''$ the system gets the required vector form if we define

$$\mathbf{S} = \begin{bmatrix} 2 \\ x(2) \\ x'(2) \\ y(2) \\ y'(2) \\ y''(2) \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 1 \\ -1 \\ 0 \\ 3 \end{bmatrix}.$$

and

$$\mathbf{F}(\mathbf{X}) = \begin{bmatrix} 1 \\ x_2 \\ \sqrt{\frac{6}{1+x_1^2}} - \frac{1}{2+x_3^2} + \sin x_0 - 2\cos(x_2 x_5) \\ x_4 \\ x_5 \\ -\sqrt{\frac{4}{1+x_1^2}} - \frac{1}{1+x_3^4} + \cos x_0 + \sin(x_1^2 x_4) \end{bmatrix}.$$

b) As in the task 6, to avoid double subscripts, we denote by $x_i^{(k)}$ the approximation from the method to the component function $x_i(t)$ at time $t_k = a + kh$. With the notational convention $\mathbf{X}_k = [x_0^{(k)}, x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)}, x_5^{(k)}]^T$ and $\mathbf{F}(\mathbf{X})$ as defined, the Explicit Euler method becomes

$$\mathbf{X}_{k+1} = \mathbf{X}_k + h\mathbf{F}(\mathbf{X}_k)$$

and a step with the classical Runge-Kutta method can be written as

$$\begin{aligned}
\mathbf{K}_1 &= h\mathbf{F}(\mathbf{X}_k) \\
\mathbf{K}_2 &= h\mathbf{F}(\mathbf{X}_k + \frac{1}{2}\mathbf{K}_1) \\
\mathbf{K}_3 &= h\mathbf{F}(\mathbf{X}_k + \frac{1}{2}\mathbf{K}_2) \\
\mathbf{K}_4 &= h\mathbf{F}(\mathbf{X}_k + \mathbf{K}_3) \\
\mathbf{X}_{k+1} &= \mathbf{X}_k + \frac{1}{6}\left(\mathbf{K}_1 + 2\mathbf{K}_2 + 2\mathbf{K}_3 + \mathbf{K}_4\right)
\end{aligned}$$

c) Explicit Euler is implemented in the MATLAB function `Euler` and the classical fourth order Runge-Kutta method in `RK4`. The specific function is implemented as `F`. The results are plotted in Figure 1.

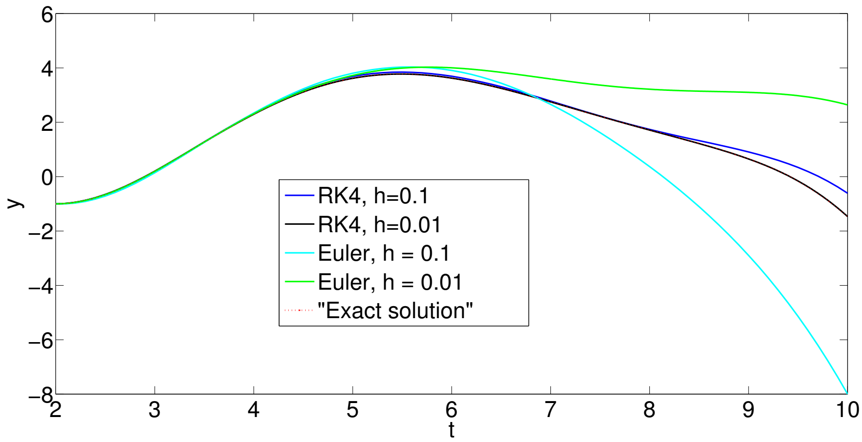

Figure 1: The various approximate solutions of the original y component for the system of ODEs, alongside an "exact solution".

We see as expected that the fourth order Runge-Kutta method is noticabely more accurate than the first order Euler method. Euler is qualitatively very different for both stepsizes. The fourth order Runge-Kutta method is quite close to the reference solution for $h = 0.1$ with some difference towards the endpoint. For $h = 0.01$ the approximate solution is indistinguishable from our "exact solution", found using **ode45**, at the current level of magnification.