Numerical Solution of Ordinary Differential Equations
(MTH 452/552)
Some practice problems for the final

following problems are not to be turned in but are relevant for the final. Additional practice problems are given by the midterm, the midterm practice problems, and the homework problems.

1. For both parts of this problem you may use a calculator but not Matlab.

   a) Perform two steps of the Newton method for the equation $f(s) = 2 - s^2 = 0$. Start with $s^{(0)} = 1$. Observe that this gives a method to compute $\sqrt{2}$ that requires only additions, multiplications, and divisions. How many correct digits does $s^{(2)}$ have?

   b) Consider the non-linear system of equations

   $$\begin{aligned} x_1^2 + x_1 x_2 &= 6 \\ x_1 x_2^2 + x_2^3 &= 3 \end{aligned}$$

   Rewrite the system in the form $F(x_1, x_2) = 0$ and perform one step of Newton's method with starting value $x^{(0)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

2. **a)** For $s \in \mathbf{R}$ Find the solution of the IVP

   $$w'' - w = 0, \qquad w(0) = 0, \;\; w'(0) = s.$$

   b) Perform by hand one step of the shooting method for the boundary value problem (BVP)

   $$w'' - w = 0, \qquad w(0) = 0, \;\; w(1) = 1.$$

   Use the starting values $w(0) = 0, w'(0) = 1$. How accurate is the answer? What happens if the starting value for $w'(0)$ is changed, say $w'(0) = s$ for some $s \in \mathbf{R}$?

3. Consider the method

   $$U^{n+1} = U^n + \Delta t\, f\left(t_n + \frac{\Delta t}{2},\; U^n + \frac{\Delta t}{2} f(t_n, U^n)\right) \tag{1}$$

   and the implicit Euler method

   $$U^{n+1} = U^n + \Delta t\, f(t_{n+1}, U^{n+1}). \tag{2}$$

   a) For both methods determine the stability function and whether or not the method is A-stable and/or isometry preserving.

b) Which of the two methods can be expected to perform better on the problem $u' = 1+u^2$, $u(0) = 0$, $0 \le t \le 1$? Justify your answer with an appropriate mathematical analysis of the methods.

c) Which of the two methods can be expected to perform better on the problem

$$u' = \begin{bmatrix} -2001 & 1001 \\ -2002 & 1002 \end{bmatrix} u, \quad u(0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad 0 \le t \le 1?$$

Justify your answer with an appropriate mathematical analysis of the methods.

4. At the bottom of page 126 our textbook states that consistency of a Runge-Kutta method requires the condition $\sum_j b_j = 1$ as well as the additional conditions $\sum_j a_{ij} = c_i$, $i = 1, \ldots, r$. Investigate whether or not the additional conditions are indeed necessary for consistency. Either prove that they are necessary or give an example of a consistent RK method that fails to satisfy at least one of these conditions.

1) a) $f(s) = 2 - s^2 = 0$

Newton's method in 1 variable gives the iteration formula

$$s^{(k+1)} = s^{(k)} - \frac{f(s^{(k)})}{f'(s^{(k)})}. \quad \text{This gives}$$

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

k	$s^{(k)}$
0	1
1	$3/2$
2	$\frac{3}{4} + \frac{2}{3} = \frac{17}{12} \simeq 1.4167$

$\sqrt{2} \simeq 1.4142$

So $s^{(2)} = \frac{17}{12}$ approximates $\sqrt{2}$ with 3 accurate digits.

b) $$F(x_1, x_2) = \begin{bmatrix} x_1^2 + x_1 x_2 - 6 \\ x_1 x_2^2 + x_2^3 - 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Jacobian: $$J = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial F_1}{\partial x_1} = \frac{\partial}{\partial x_1}(x_1^2 + x_1 x_2 - 6) = 2x_1 + x_2$$

$$\frac{\partial F_1}{\partial x_2} = \frac{\partial}{\partial x_2}(x_1^2 + x_1 x_2 - 6) = x_1$$

$$\frac{\partial F_2}{\partial x_1} = \frac{\partial}{\partial x_1}(x_1 x_2^2 + x_2^3 - 3) = x_2^2$$

$$\frac{\partial F_2}{\partial x_2} = \frac{\partial}{\partial x_2}(x_1 x_2^2 + x_2^3 - 3) = 2x_1 x_2 + 3x_2^2$$

$$\Rightarrow J = \begin{bmatrix} (2x_1 + x_2) & x_1 \\ x_2^2 & (2x_1x_2 + 3x_2^2) \end{bmatrix}$$

$$x^{(0)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}. \quad F(x^{(0)}) = \begin{bmatrix} 1+2-6 \\ 4+8-3 \end{bmatrix} = \begin{bmatrix} -3 \\ 9 \end{bmatrix}$$

$$J\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \begin{bmatrix} (2+2) & 1 \\ 4 & (4+12) \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 4 & 16 \end{bmatrix}$$

Now find $d = x^{(1)} - x^{(0)}$ by solving

$$J d = -F(x^{(0)}), \text{ i.e.}$$

$$\begin{bmatrix} 4 & 1 \\ 4 & 16 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = -\begin{bmatrix} -3 \\ 9 \end{bmatrix} = \begin{bmatrix} 3 \\ -9 \end{bmatrix}$$

$$\Rightarrow d = J^{-1} \begin{bmatrix} 3 \\ -9 \end{bmatrix} = \underbrace{\frac{1}{60} \begin{bmatrix} 16 & -1 \\ -4 & 4 \end{bmatrix}}_{\text{Use } \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}} \begin{bmatrix} 3 \\ -9 \end{bmatrix} = \frac{1}{60} \begin{bmatrix} 57 \\ -48 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 19 \\ -16 \end{bmatrix}$$

$$x^{(1)} = x^{(0)} + d = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \frac{1}{20} \begin{bmatrix} 19 \\ -16 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 39 \\ 24 \end{bmatrix} = \begin{bmatrix} 1.95 \\ 1.2 \end{bmatrix}$$

2) $w'' - w = 0$, $w(0) = 0$, $w'(0) = s$

$w'' = w \Rightarrow w = c_1 e^t + c_2 e^{-t}$, $w'(t) = c_1 e^t - c_2 e^{-t}$

$$\left.\begin{aligned} w(0) &= c_1 + c_2 = 0 \\ w'(0) &= c_1 - c_2 = s \end{aligned}\right\} \Rightarrow c_2 = -c_1,\ 2c_1 = s$$

$\Rightarrow c_1 = \frac{s}{2}$, $c_2 = -\frac{s}{2}$

The solution is $w_s(t) = \frac{s}{2} e^t - \frac{s}{2} e^{-t} = s \sinh(t)$

b) Find s such that $W_s(1) = s \sinh(1) = 1$.

Solve $F(s) = s \cdot \sinh(1) - 1 = 0$.

The exact answer is obviously $s = \frac{1}{\sinh(1)}$, but to get practice we carry out a step of the Newton method with starting value $s^{(0)} = w'(0) = 1$.

The Newton iteration is

$$s^{(q+1)} = s^{(q)} - \frac{F(s^{(q)})}{F'(s^{(q)})} = s^{(q)} - \frac{s^{(q)} \sinh(1) - 1}{\sinh(1)}$$

$$= s^{(q)} - s^{(q)} + \frac{1}{\sinh(1)} = \frac{1}{\sinh(1)}$$

So the Newton method finds the exact solution in a single step regardless of the starting value. This is always the case when the Newton method is applied to a linear system of equations.

3) $U^{n+1} = U^n + \Delta t\, f\left(t_n + \frac{\Delta t}{2}, U^n + \frac{\Delta t}{2} f(t_n, U^n)\right)$

is a RK method;

$$K_1 = f(t_n, U^n), \quad K_2 = f\left(t_n + \frac{\Delta t}{2}, U^n + \frac{\Delta t}{2} K_1\right)$$

$$U^{n+1} = U^n + \Delta t\, K_2$$

The Butcher array is therefore

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}, \quad \text{i.e.,} \quad A = \begin{bmatrix} 0 & 0 \\ \frac{1}{2} & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad c = \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}.$$

For later use we determine the order of consistency: One has

$$\sum b_j = 0 + 1 = 1, \quad \sum b_j c_j = 0 \cdot 0 + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

$$\sum b_j c_j^2 = 1 \cdot \frac{1}{4} = \frac{1}{4} \neq 1/3.$$

The method is consistent of order 2.

Stability function:

$$R(z) = 1 + z\, b^T (I - zA)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= 1 + z\, [0, 1] \begin{bmatrix} 1 & 0 \\ -\frac{z}{2} & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= 1 + z\, [0, 1] \begin{bmatrix} 1 & 0 \\ z/2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1 + z\, [0, 1] \begin{bmatrix} 1 \\ \frac{z}{2} + 1 \end{bmatrix}$$

$$= 1 + z\left(\frac{z}{2} + 1\right) = 1 + z + \frac{1}{2} z^2$$

Alternatively one can find $R(z)$ by applying the method to the test problem $u' = \lambda u$, $u(0) = 1 = U^0$. Then $U^n = (R(\lambda \Delta t))^n$, in particular, $U^1 = R(\lambda \Delta t)$.

For $f(t, u) = \lambda u$ and $U^0 = 1$ one obtains

$$U^1 = U^0 + \Delta t\, f\left(0 + \frac{\Delta t}{2}, U^0 + \frac{\Delta t}{2} f(0, U^0)\right) = 1 + \Delta t \cdot \lambda \cdot \left(U^0 + \frac{\Delta t}{2} f(0, U^0)\right) =$$

$$1+\Delta t\cdot\lambda\cdot\left(\overset{1}{\overset{\shortparallel}{u^0}}+\frac{\Delta t}{2}\lambda\overset{1}{\overset{\shortparallel}{u^0}}\right)=$$

$$1+\lambda\Delta t\left(1+\tfrac{1}{2}\lambda\Delta t\right)=$$

$$1+\lambda\Delta t+\tfrac{1}{2}(\lambda\Delta t)^2=R(\lambda\Delta t),$$

which again yields $R(z)=1+z+\frac{1}{2}z^2$.

Since $|R(z)|\to\infty$ as $|z|\to\infty$, the method is not A-stable.

$$|R(it)|=\left|1+it-\tfrac{1}{2}t^2\right|=\left[\left(1-\tfrac{1}{2}t^2\right)^2+t^2\right]^{1/2}\xrightarrow[t\to\infty]{}\infty$$

The method is not isometry preserving.

The implicit Euler method is also a RK method with Butcher array $\begin{array}{c|c}1&1\\\hline&1\end{array}$, i.e., $A=(1)$, $b=[1]$, $c=(1)$.

Stability function: $R(z)=1+z(1-z)^{-1}\cdot 1=\dfrac{1}{1-z}$

$|R(z)|=\dfrac{1}{|1-z|}<1$ if $\operatorname{Re} z<0$. The method is A-stable.

If $z=it$, then $|R(z)|=(1+t^2)^{-1/2}\xrightarrow[t\to\infty]{}0$.

The method is not isometry preserving.

The order of consistency is 1.

b) The IVP $u'=1+u^2$, $u(0)=0$ has Jacobian $J=\frac{df}{du}=2u$, so no large negative eigenvalues, since $u\geq 0$. So the higher order method can be expected to perform better.

c) Here the matrix has eigenvalues $\lambda_1=1$, $\lambda_2=-1{,}000$. The general sol. has the form $u=c_1e^t+c_2e^{-1{,}000t}$. After a very short time the term $c_2e^{-1{,}000t}$ is negligibly small.

However, the presence of $\lambda_2 = -1{,}000$ in the ODE requires a stepsize Δt that is small enough so $-1{,}000\,\Delta t$ is in the region of absolute stability. If only moderate accuracy is required, the A-stable implicit Euler method may be better. For the RK method we can compute the intersection of the region of absolute stability with the negative real axis (note that $-1{,}000\,\Delta t$ always lies on the negative real axis). $|R(t)| \le 1 \iff -1 \le R(t) \le 1$

$$R(t) \le 1 \iff R(t) - 1 \le 0 \iff 1 + t + \tfrac{1}{2}t^2 - 1 = t\left(1 + \tfrac{1}{2}t\right) \le 0$$

$$\iff t \le 0 \text{ and } 1 + \tfrac{1}{2}t \ge 0 \iff -2 \le t \le 0.$$

$$-1 \le R(t) \iff 0 \le R(t) + 1 = 2 + t + \tfrac{1}{2}t^2$$

which is true for all t.

So $|R(t)| \le 1 \iff -2 \le t \le 0.$

Therefore stability requires $-1{,}000\,\Delta t \in [-2, 0]$, i.e.

$$\Delta t \le \frac{1}{500}.$$

The implicit Euler method may be more efficient if it can achieve the desired accuracy with a stepsize Δt that is significantly larger than $\frac{1}{500}$, since one also has to solve the implicit equation $U^{n+1} = U^n + \Delta t f(t_{n+1}, U^{n+1})$.

4) One can use the theorem given in the midterm training problems: A method

$$\sum_{j=0}^{r} \alpha_j U^{n+j} = \Delta t \, \phi_f (U^{n+r}, \dots, U^n, t_n; \Delta t)$$

is consistent if $\sum \alpha_j = 0$ and

$$\phi_f (u(t), \dots, u(t), t, 0) = f(t, u(t)) \sum_j j \alpha_j .$$

For a RK method $\phi_f = \sum b_j K_j$, where

$$K_j = f(t_n + c_j \Delta t, U^n + \Delta t \sum_\ell a_{j\ell} K_\ell)$$

To apply the theorem we evaluate ϕ_f for $U^{n+j} = u(t)$, $t_n = t$ and $\Delta t = 0$. Then $K_j = f(t, u(t))$ for all j, and $\phi_f (u(t), \dots, u(t), t, 0) = \sum b_j \, f(t, u(t))$.

Also, for an RK-method $r = 1$ and $\alpha_0 = -1$, $\alpha_1 = 1$.

So $\sum \alpha_j = 0$ and $\sum j \alpha_j = 0 \cdot \alpha_0 + 1 \cdot \alpha_1 = 1$.

This shows that the only condition required for consistency is $\sum b_j = \sum j \alpha_j = 1$.