Faculty of Electrical Engineering, Mathematics and Computer Science
Numerieke Methoden I, TW2060, BSc Technische Wiskunde
Answers to the exam of June 29, 2018 test

1 a The local truncation error is defined by

$$\tau_{n+1}(\Delta t) := \frac{y_{n+1} - z_{n+1}}{\Delta t}, \tag{1}$$

where z_{n+1} is obtained from using the exact solution $y_n = y(t_n)$ at time t_n. For the exact solution, we have

$$y_{n+1} = y(t_n + \Delta t) = y_n + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(\zeta), \text{ with } \zeta \in (t_n, t_{n+1}). \tag{2}$$

Using $y'(t) = f(t,y)$ gives $y'(t_n) = f(t_n, y_n)$ and hence

$$y_{n+1} = y(t_n + \Delta t) = y_n + \Delta t f(t_n, y_n) + \frac{\Delta t^2}{2} y''(\zeta), \text{ with } \zeta \in (t_n, t_{n+1}). \tag{3}$$

Since Euler's forward method applied using the exact solution at the previous time, gives the following expression for z_{n+1}

$$z_{n+1} = y_n + \Delta t f(t_n, y_n), \tag{4}$$

we get, using equation (1),

$$\tau_{n+1}(\Delta t) = \frac{y_n + \Delta t f(t_n, y_n) + \frac{\Delta t^2}{2} y''(\zeta) - (y_n + \Delta t f(t_n, y_n))}{\Delta t} = \frac{\Delta t}{2} y''(\zeta) = \mathcal{O}(\Delta t). \tag{5}$$

One may relate $y''(t)$ to f by $y''(t) = \frac{dy'(t)}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} y'(t) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f(t,y)$ (but this is not necessary).

b We apply Euler's forward method to

$$\begin{cases} y_1' = -2y_1 + \alpha y_2, \\ y_2' = -\alpha y_1 - 2y_2. \end{cases} \tag{6}$$

We start by writing the above system of ODEs into the following matrix equation

$$\mathbf{y}' = A\mathbf{y}, \text{ where } A = \begin{pmatrix} -2 & \alpha \\ -\alpha & -2 \end{pmatrix}. \tag{7}$$

Subsequently, we determine the eigenvalues of the matrix A, which are given by

$$\lambda(A) = \{-2 - \alpha i, -2 + \alpha i\}.$$

Note that both eigenvalues are non-real. For the sake of stability, we need the amplification factor, which is derived by the use of the test equation, which reads as $y' = \lambda y$. For Euler's forward method, we get

$$w_{n+1} = w_n + \Delta t \lambda w_n = (1 + \Delta t \lambda) w_n, \tag{8}$$

and hence the amplification factor is given by

$$Q(\Delta t \lambda) = 1 + \Delta t \lambda. \tag{9}$$

Stability of the numerical method requires that

$$|Q(\Delta t \lambda)| \leq 1, \text{ for all eigenvalues in } \lambda(A). \tag{10}$$

Hence, for the current system, we get

$$Q(\Delta t\lambda) = 1 + \Delta t(-2 \pm \alpha i), \tag{11}$$

which gives the following condition

$$|Q(\Delta t\lambda)|^2 = (1 - 2\Delta t)^2 + (\alpha \Delta t)^2 \leq 1. \tag{12}$$

Processing the above equation further, gives

$$1 - 4\Delta t + (4 + \alpha^2)\Delta t^2 \leq 1 \Longrightarrow \Delta t \leq \frac{4}{4 + \alpha^2}. \tag{13}$$

Note that since $\lambda \notin \mathbb{R}$, it is not allowed to use $\Delta t \leq \frac{2}{-\lambda}$.

c Lax Equivalence Theorem states: *A stable, consistent scheme gives a converging numerical solution.* Consistency of the scheme means that $\lim_{\Delta t \to 0} \tau_{n+1}(\Delta t) = 0$. In assignment 1a, we proved that $\tau_{n+1}(\Delta t) = \mathscr{O}(\Delta t)$ and hence $\lim_{\Delta t \to 0} \tau_{n+1}(\Delta t) = 0$ and herewith the scheme is consistent (the local truncation error tends to zero as the time-step tends to zero).

We also demonstrated that the scheme is stable if $\Delta t \leq \dfrac{4}{4+\alpha^2}$. Hence the numerical solution converges if $\Delta t \leq \dfrac{4}{4+\alpha^2}$.

d To assess numerical stability, we apply the method to the test equation, then for the Trapezoidal Rule, we obtain the following expression

$$w_{n+1} = w_n + \Delta t\lambda(w_n + w_{n+1}). \tag{14}$$

The above equation gives the following result for w_{n+1}

$$w_{n+1} = \frac{1 + \frac{\Delta t\lambda}{2}}{1 - \frac{\Delta t\lambda}{2}} w_n. \tag{15}$$

Hence the amplification factor is given by

$$Q(\Delta t\lambda) = \frac{1 + \frac{\Delta t\lambda}{2}}{1 - \frac{\Delta t\lambda}{2}}. \tag{16}$$

For stability, it is required to have $|Q(\Delta t\lambda)| \leq 1$, hence using the eigenvalues from question 1b, gives

$$|\frac{1 + \frac{\Delta t\lambda}{2}}{1 - \frac{\Delta t\lambda}{2}}|^2 = \frac{|1 + \frac{\Delta t\lambda}{2}|^2}{|1 - \frac{\Delta t\lambda}{2}|^2} = \frac{(1 - \Delta t)^2 + (\frac{\Delta t\alpha}{2})^2}{(1 + \Delta t)^2 + (\frac{\Delta t\alpha}{2})^2} \leq 1. \tag{17}$$

The above inequality easily follows for all $\Delta t \geq 0$, and hence the Trapezoidal method is stable for all choices of $\Delta t \geq 0$ (unconditionally stable).

e Since the method is implicit, we are allowed to use the test equation for the assessment of the local truncation error. For the exact solution, we have

$$y_{n+1} = y(t_n + \Delta t) = y_n + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(t_n) + \frac{\Delta t^3}{6} y'''(t_n) + \frac{\Delta t^4}{24} y^{(4)}(\zeta), \text{ for } \zeta \in (t_n, t_{n+1}). \tag{18}$$

Using the test equation, this gives

$$y_{n+1} = (1 + \Delta t\lambda + \frac{(\Delta t\lambda)^2}{2} + \frac{(\Delta t\lambda)^3}{6} + \mathscr{O}((\Delta t\lambda)^4)) y_n. \tag{19}$$

For the numerical approximation using the exact solution at the previous time, we get

$$z_{n+1} = y_n + \Delta t \lambda (y_n + z_{n+1}), \tag{20}$$

and hence we get

$$z_{n+1} = Q(\Delta t \lambda) y_n. \tag{21}$$

Hence for the local truncation error, we get

$$\tau_{n+1}(\Delta t) = \frac{y_{n+1} - z_{n+1}}{\Delta t} = \frac{(1 + \Delta t \lambda + \frac{(\Delta t \lambda)^2}{2} + \frac{(\Delta t \lambda)^3}{6} + \mathscr{O}(\Delta t^4) - Q(\Delta t \lambda)) y_n}{\Delta t}. \tag{22}$$

Using the following expansion (one can derive this result using the geometric series $(1-x)^{-1} = 1 + x + x^2 + \mathscr{O}(x^3)$), one gets

$$Q(\Delta t \lambda) = 1 + \Delta t \lambda + \frac{(\Delta t \lambda)^2}{2} + \frac{(\Delta t \lambda)^3}{4} + \mathscr{O}(\Delta t \lambda)^4. \tag{23}$$

This expression is substituted into the local truncation error to yield

$$\tau_{n+1}(\Delta t) = \frac{(1 + \Delta t \lambda + \frac{(\Delta t \lambda)^2}{2} + \frac{(\Delta t \lambda)^3}{6} + \mathscr{O}(\Delta t^4) - (1 + \Delta t \lambda + \frac{(\Delta t \lambda)^2}{2} + \frac{(\Delta t \lambda)^3}{4} + \mathscr{O}(\Delta t \lambda)^4)) y_n}{\Delta t} = \mathscr{O}(\Delta t^2). \tag{24}$$

(As alternative proofs of the above statement, one can also write out a Taylor series for $Q(\Delta t \lambda)$ or make the denominators equal in the fractions.)

2 We consider the boundary value problem

$$\begin{cases} -y'' + y = f(x) = x^2 - 2, & \text{voor } x \in (0,1), \\ y(0) = 0, & y(1) = 1. \end{cases} \tag{25}$$

a Showing that $y(x) = x^2$ represents the solution to the boundary value problem amounts to substituting the expression into the differential equation, which gives $-y'' + y = -2 + x^2 = f(x)$, which validates the differential equation. Subsequently, the solution of the boundary value problem should also satisfy the boundary conditions. For $y(x) = x^2$, it is easy to see by substituting $x = 0$ and $x = 1$ that that $y(0) = 0$ and $y(1) = 1$, which verifies that the boundary conditions are satisfied. Hence the expression $y(x) = x^2$ is indeed a solution to the boundary value problem.

b Using the finite difference method amounts to replacing all derivatives with difference formulas. This gives

$$y''(x) \approx \frac{y(x + \Delta x) - 2y(x) + y(x - \Delta x)}{\Delta x^2}, \tag{26}$$

and to dividing the domain of computation into meshpoints, $x_j = j\Delta x$, where we choose $x_{n+1} = 1$. Then the above equation is substituted into the differential equation and the truncation error is neglected to give

$$\frac{-w_{j-1} + 2w_j - w_{j+1}}{\Delta x^2} + w_j = f(x_j) = x_j^2 - 2, \text{ for } j = 1, \ldots, n. \tag{27}$$

Note that w_j represents the numerical solution, whereas y_j denotes the exact solution. Further, we note that there n unknowns (degrees of freedom). For $j = 1$, $w_{j-1} = w_0 = y_0 = 0$ from the boundary condition, and likewise for $j = n$, we have $w_{n+1} = 1$, this gives the following two equations for $j = 1$ and $j = n$:

$$\begin{aligned} &\tfrac{2w_1 - w_2}{\Delta x^2} + w_1 = f(x_1) = x_1^2 - 2, \\ &\tfrac{-w_{n-1} + 2w_n}{\Delta x^2} + w_n = f(x_n) + \tfrac{1}{\Delta x^2} = x_n^2 - 2 + \tfrac{1}{\Delta x^2}. \end{aligned} \tag{28}$$

This implies $a_{ii} = \frac{2}{\Delta x^2} + 1$ for all $i = 1, \ldots, n$, $a_{ii-1} = -\frac{1}{\Delta x^2}$ for all $i = 2, \ldots, n$ and $a_{ii+1} = -\frac{1}{\Delta x^2}$ for all $i = 1, \ldots, n-1$. Further $b_j = f(x_j) = x_j^2 - 2$ for $j = 1, \ldots, n-1$ and $b_n = f(x_n) + \frac{1}{\Delta x^2} = x_n^2 - 2 + \frac{1}{\Delta x^2}$.

c i The local truncation error is given by

$$\varepsilon = A\mathbf{y} - \mathbf{b}. \tag{29}$$

Componentwisely, for $i = 2, \ldots, n-1$, we have $b_i = f(x_i) = -y''(x_i) + y_i$, hence applying the above equation to these rows, we get

$$\begin{aligned} \varepsilon_i = -\frac{y_{i-1}}{\Delta x^2} + (\frac{2}{\Delta x^2} + 1)y_i - \frac{y_{i+1}}{\Delta x^2} y_{i+1} + y''(x_i) - y_i = \\ -\frac{y_{i-1} - 2y_i + y_{i+1}}{\Delta x^2} + y''(x_i) = \mathscr{O}(\Delta x^2). \end{aligned} \tag{30}$$

The last equality comes from the fact that the central differences represents a second order approximation of the second order derivative. This fact is shown by using Taylor series:

$$y_{i\pm1} = y(x_i \pm \Delta x) = y(x_i + \Delta x) = y_i \pm \Delta x y'(x_i) + \frac{\Delta x^2}{2} y''(x_i) \pm \frac{\Delta x^3}{6} y'''(x_i) + \frac{\Delta x^4}{24} y^{(4)}(\zeta_\pm), \tag{31}$$

where $\zeta_\pm \in (\min(x_{i\pm1}, x_i), \max(x_{i\pm1}, x_i))$.

For the boundary nodes we use $y_0 = 0$ and $y_n = 1$, respectively, to get

$$\begin{aligned} \varepsilon_1 = (\frac{2}{\Delta x^2} + 1)y_1 - \frac{y_{i-1}}{\Delta x^2} y_2 + y''(x_1) - y_1 + \frac{y_0}{\Delta x^2} = \\ -\frac{y_0 - 2y_1 + y_2}{\Delta x^2} + y''(x_1) = \mathscr{O}(\Delta x^2), \end{aligned} \tag{32}$$

and

$$\begin{aligned} \varepsilon_n = -\frac{y_{n-1}}{\Delta x^2} + (\frac{2}{\Delta x^2} + 1)y_n + y''(x_n) - y_n - \frac{y_n}{\Delta x^2} = \\ -\frac{y_{n-1} - 2y_n + y_{n+1}}{\Delta x^2} + y''(x_n) = \mathscr{O}(\Delta x^2). \end{aligned} \tag{33}$$

Herewith the local truncation error is of order $\mathscr{O}(\Delta x^2)$.

ii We verified that the exact solution is given by $y(x) = x^2$, for which $y^{(p)}(x) = 0$ for $p \geq 3$. Since the local truncation error only contains fourth order derivatives of the solution $y(x)$, it clearly follows that the local truncation error vanishes (that is, it is zero). Hence $\varepsilon = \mathbf{0}$.

d Gershgorin's Theorem says: *Given an $n \times n$-matrix A, and let $\lambda(A) = \{\lambda_1, \ldots, \lambda_n\}$, where λ_i is an eigenvalue of A, then*

$$\lambda(A) \subset \bigcup_{i=1}^{n} \{\lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{j \in \{1,\ldots,n\} \setminus \{i\}} |a_{ij}\}. \tag{34}$$

First of all, symmetry of A implies that the eigenvalues of A are real-valued. Application of Gershgorin's Theorem to the first and last row of the matrix A gives:

$$|\lambda - (\frac{2}{\Delta x^2} + 1)| \leq |-\frac{1}{\Delta x^2}| = \frac{1}{\Delta x^2}. \tag{35}$$

Since A has real eigenvalues only, this implies

$$-\frac{1}{\Delta x^2} \leq \lambda - (\frac{2}{\Delta x^2} + 1) \leq \frac{1}{\Delta x^2}. \tag{36}$$

This implies

$$1+\frac{1}{\Delta x^2} \le \lambda \le 1+\frac{3}{\Delta x^2}. \tag{37}$$

For the remaining rows, we get

$$|\lambda-(\frac{2}{\Delta x^2}+1)| \le |-\frac{1}{\Delta x^2}|+|-\frac{1}{\Delta x^2}| = \frac{2}{\Delta x^2}. \tag{38}$$

Since A has real eigenvalues only, this implies

$$-\frac{2}{\Delta x^2} \le \lambda-(\frac{2}{\Delta x^2}+1) \le \frac{2}{\Delta x^2}. \tag{39}$$

This implies

$$1 \le \lambda \le 1+\frac{4}{\Delta x^2}. \tag{40}$$

Since Gershgorin gives the union of the discs in the complex plane for the location of the eigenvalues, and since the bounds from the first and last rows of the matrix are contained within the other rows, it follows directly that the eigenvalues of A are between the following bounds:

$$1 \le \lambda \le 1+\frac{4}{\Delta x^2}, \text{ or } \lambda(A) \subset [1, 1+\frac{4}{\Delta x^2}]. \tag{41}$$

The highest lower bound of the eigenvalues of A is given by 1 and the lowest upper bound of the eigenvalues of A is given by $1+\frac{4}{\Delta x^2}$.

e By definition, the discretisation for the boundary value problem is consistent if $||\varepsilon|| = ||A\mathbf{y}-\mathbf{b}|| \longrightarrow 0$ as $\Delta x \longrightarrow 0$. In assignment 2ci, we saw that $||\varepsilon||_2 = ||A\mathbf{y}-\mathbf{b}||_2 = \mathscr{O}(\Delta x^2) \longrightarrow 0$ as $\Delta x \longrightarrow 0$. Hence the current finite differences schema is consistent.

Furthermore, A is symmetric, which implies that A^{-1} is also symmetric. Since for symmetric matrices, we have $||A||_2 = \max|\lambda(A)|$, where $|\lambda(A)| := \{|\lambda_1|, \ldots, |\lambda_n|\}$, and since if λ is an eigenvalue of A then $\frac{1}{\lambda}$ is an eigenvalue of A^{-1}, it follows that

$$||A^{-1^{\text{`}}}||_2 = \frac{1}{\min|\lambda(A)|} \le 1 \text{ also as } \Delta x \longrightarrow 0. \tag{42}$$

Hence by definition, the finite difference scheme is stable.

The equivalence Theorem by Lax says: *A stable, consistent discretisation method for an initial boundary problem gives a convergent solution, that is* $\mathbf{e} = \mathbf{y}-\mathbf{w} \longrightarrow 0$ *as* $\Delta x \longrightarrow 0$.

We already proved that the current scheme is stable and consistent, hence from Lax' Theorem, it follows that the numerical solution converges the exact solution as $\Delta x \longrightarrow 0$ (in the current case, we even have that the numerical solution and the exact solution coincide).

3 We solve the following problem for $p > 0$:

$$\text{Solve for } p>0\text{: } f(p) = p^2-3 = 0. \tag{43}$$

a i From $p_0 = 1$, we get $p_1 = 1-\frac{1}{9}(1-3) = \frac{11}{9}$. Subsequently, we get $p_2 = \frac{11}{9}-\frac{1}{9}(\frac{121}{81}-3) = \frac{953}{729}$.

ii We use Banach's Contraction Theorem, which for the current case says:

Let $g \in C^1[1,3]$*, and* $g : [1,3] \longrightarrow [1,3]$*, and if* $\exists\, k \in [0,1)$ *such that* $|g'(x)| \le k$ *for all* $x \in [1,3]$*, then* $\exists!\, p \in [1,3]$ *such that* $p = g(p)$.

First, it is clear that $g(x)$ is a polynomial, which is in $C^1[1,3]$. Furthermore, $g'(x) = 1-\frac{2}{9}x$,

and hence $g'(x) > 0$ on $[1,3]$, which implies that $g(x)$ is monotonically increasing on $[1,3]$. Further, $g(1) = \frac{11}{9} \in [1,3]$ and $g(3) = \frac{7}{3} \in [1,3]$ and hence $g : [1,3] \longrightarrow [1,3]$.
Secondly, we see that $g'(x) = 1 - \frac{2}{9}x$, which decreases monotonically and takes values within $[\frac{1}{3}, \frac{7}{9}]$, hence $|g'(x)| \leq \frac{7}{9} \in [0,1)$. Herewith all requirements of Banach's Fixed Point Theorem are satisfied and therewith there is one and only one fixed point for $g(x)$ on $[1,3]$.

iii Picard's Iteration Theorem says:

Let $g \in C^1[1,3]$, and $g : [1,3] \longrightarrow [1,3]$, and if $\exists\ k \in [0,1)$ such that $|g'(x)| \leq k$ for all $x \in [1,3]$, then the sequence defined by $p_{n+1} = g(p_n)$ gives $p_n \longrightarrow p$ as $n \longrightarrow \infty$, where $p = g(p)$, for each $p_0 \in [1,3]$.

In assignment 3aii, we demonstrated that all hypotheses in the above theorem are satisfied, hence we have $p_n \longrightarrow p$ as $n \longrightarrow \infty$ for each $p_0 \in [1,3]$.

(As an alternative answer, one may also base the proof on $|p - p_{n+1}| = |g(p) - g(p_n)| = |g'(\zeta)||p - p_n| \leq k|p - p_n|$, using the Mean Value Theorem, and which inductively will amount to $0 \leq |p - p_{n+1}| \leq k^{n+1}|p - p_0| = (\frac{7}{9})^{n+1}|p - p_0| \to 0$ as $n \longrightarrow \infty$, and use the Squeeze Theorem.)

b Newton-Raphson's Method is based on successive approximations of a zero of the function $f(x)$. Suppose that p_n is known, then we equate the linearisation of $f(x)$ (and disregard errors) around p_n to zero at the subsequent estimate p_{n+1}:

$$f(p_n) + f'(p_n)(p_{n+1} - p_n) = 0. \tag{44}$$

Then it easily follows that

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)}. \tag{45}$$

c Application of the Newton-Raphson method to the function $f(x) = x^2 - 3$, gives with $f'(x) = 2x$:

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n} = p_n - \frac{p_n^2 - 3}{2p_n} = \frac{2p_n^2 - (p_n^2 - 3)}{2p_n} = \frac{p_n^2 + 3}{2p_n}. \tag{46}$$

d Take $p_0 = 1$, then $p_1 = \frac{1+3}{2*1} = 2$, and $p_2 = \frac{2^2+3}{2*2} = \frac{7}{4}$.