DELFT UNIVERSITY OF TECHNOLOGY
Faculty of Electrical Engineering, Mathematics and Computer Science

ANSWERS OF THE TEST NUMERICAL METHODS FOR DIFFERENTIAL EQUATIONS (WI3097 TU)
Thursday July 6 2017, 18:30-21:30

1. (a) The local truncation error is defined by

$$\tau_h = \frac{y_{n+1} - z_{n+1}}{\Delta t}, \tag{1}$$

where

$$z_{n+1} = y_n + \Delta t f(t_n, y_n), \tag{2}$$

for the Forward Euler method. A Taylor expansion for y_{n+1} around t_n is given by

$$y_{n+1} = y_n + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(\xi), \quad \exists\, \xi \in (t_n, t_{n+1}). \tag{3}$$

Since $y'(t_n) = f(t_n, y_n)$, we use equation (1), to get

$$\tau_h = \frac{\Delta t}{2} y''(\xi), \quad \exists\, \xi \in (t_n, t_{n+1}). \tag{4}$$

Hence, the truncation error is of first order.

(b) For the amplification factor we apply the method to the test equation: $y' = \lambda y$. Application of Forward Euler to this equation leads to:

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

so the amplification factor is $Q(\lambda \Delta t) = 1 + \lambda \Delta t$.

We have to check that $|Q(\lambda \Delta t)| \leq 1$. For a negative real number λ this leads to the inequalities:

$$-1 \leq 1 + \lambda \Delta t \leq 1$$

The right hand inequality leads to $\lambda \Delta t \leq 0$. Since $\Delta t > 0$ and $\lambda \leq 0$ this inequality is always satisfied. The left hand inequality leads to $-1 \leq 1 + \lambda \Delta t$ which is equavalent to $\lambda \Delta t \geq -2$. Dividing both sides by λ which is negative leads to:

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

(c) We use the following definition $x_1 = y$ and $x_2 = y'$. This implies that $x_1' = y' = x_2$ and $x_2' = y'' = -y' - \frac{1}{2}y = -x_2 - \frac{1}{2}x_1$. Writing this in vector notation shows that

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{1}{2} & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

so $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{1}{2} & -1 \end{bmatrix}$. To compute the eigenvalues we look for values of λ such that

$$|\mathbf{A} - \lambda \mathbf{I}| = 0.$$

This implies that λ is a solution of

$$\lambda^2 + \lambda + \frac{1}{2} = 0,$$

which leads to the roots:

$$\lambda_1 = -\frac{1}{2} + \frac{1}{2}i \text{ and } \lambda_2 = -\frac{1}{2} - \frac{1}{2}i.$$

(d) We do one step with Forward Euler using $\Delta t = 1$.

$$\begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} = \begin{bmatrix} w_{1,0} \\ w_{2,0} \end{bmatrix} + \Delta t \begin{bmatrix} 0 & 1 \\ -\frac{1}{2} & -1 \end{bmatrix} \begin{bmatrix} w_{1,0} \\ w_{2,0} \end{bmatrix}$$

Substituting $\Delta t = 1$ and the initial conditions leads to:

$$\begin{bmatrix} w_{1,1} \\ w_{2,1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ -\frac{1}{2} & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{1}{2} \end{bmatrix}$$

(e) Since the eigenvalues are complex valued it is sufficient to check that the modulus: $|Q(\lambda_1 \Delta t)| \leq 1$. Substituting $\lambda_1 = -\frac{1}{2} + \frac{1}{2}i$ into $Q(\lambda_1 \Delta t)$ leads to the condition:

$$|1 + \Delta t(-\frac{1}{2} + \frac{1}{2}i)| \leq 1$$

This implies that

$$\sqrt{(1 - \frac{\Delta t}{2})^2 + (\frac{\Delta t}{2})^2} \leq 1$$

Rearranging the terms leads to

$$1 - \Delta t + \frac{1}{2}(\Delta t)^2 \leq 1$$

so

$$-\Delta t + \frac{1}{2}(\Delta t)^2 \leq 0$$

and thus

$$\Delta t \leq 2$$

.

2. (a) The iteration process is a fixed-point method. If the process converges we have: $\lim_{n\to\infty} x_n = p$. Using this in the iteration process yields:

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} [x_n + h(x_n)(x_n^3 - 27)]$$

Since h is a continuous function one obtains:

$$p = p + h(p)(p^3 - 27)$$

so

$$h(p)(p^3 - 27) = 0.$$

Since $h(x) \neq 0$ for each $x \neq 0$ it follows that $p^3 - 27 = 0$ and thus $p = 27^{\frac{1}{3}} = 3$.

(b) The convergence of a fixed-point method $x_{n+1} = g(x_n)$ is determined by $g'(p)$. If $|g'(p)| < 1$ the method converges, whereas if $|g'(p)| > 1$ the method diverges. For all choices we compute the first derivative in p. For the first method we elaborate all steps. For the other methods we only give the final result. For h_1 we have $g_1(x) = x - \frac{x^3-27}{x^4}$. The first derivative is:

$$g_1'(x) = 1 - \frac{3x^2 \cdot x^4 - (x^3 - 27) \cdot 4x^3}{(x^4)^2}$$

Substitution of p yields:

$$g_1'(p) = 1 - \frac{3p^6 - (p^3 - 27) \cdot 4p^3}{p^8}.$$

Since $p = 3$ the final term cancels:

$$g_1'(p) = 1 - \frac{3p^6}{p^8} = 1 - 3^{-1} = \frac{2}{3}.$$

This implies that the method is convergent with convergence factor $\frac{2}{3}$.

For the second method we have:

$$g_2'(p) = 1 - \frac{3p^4 - (p^3 - 27) \cdot 2p}{p^4} = 1 - \frac{3p^4}{p^4} = -2$$

Thus the method diverges.

For the third method we have:

$$g_3'(p) = 1 - \frac{9p^4 - (p^3 - 27) \cdot 6p}{9p^4} = 1 - \frac{9p^4}{9p^4} = 0$$

Thus the method is convergent with convergence factor 0.

Concluding we note that the third method is the fastest.

(c) For a general function $h_4(x)$ the first derivative of $g_4(x) = x + h_4(x)(x^3 - 27)$ evaluated in p reads

$$g_4'(p) = 1 + h_4'(p)(p^3 - 27) + 3h_4(3)p^2$$

Since $p = 3$ we obtain $g_4'(3) = 1 + 27h_4(3)$. For $|g_4'(3)| = 1$ we need to find a differentiable function $h_4(x)$ that equals 0 in $p = 3$. A possible choice is

$$h_4(x) = x - 3.$$

(d) To estimate the error in p we first approximate the function f in the neighbourhood of p by the first order Taylor polynomial:

$$P_1(x) = f(p) + (x - p)f'(p) = (x - p)f'(p).$$

Due to the measurement errors we know that

$$(x - p)f'(p) - \epsilon_{max} \leq \hat{P}_1(x) \leq (x - p)f'(p) + \epsilon_{max}.$$

This implies that the perturbed root $\hat{p}$ is bounded by the roots of $(x-p)f'(p) - \epsilon_{max}$ and $(x - p)f'(p) + \epsilon_{max}$, which leads to

$$p - \frac{\epsilon_{max}}{|f'(p)|} \leq \hat{p} \leq p + \frac{\epsilon_{max}}{|f'(p)|}.$$

3. (a) Using central differences for the second order derivative at a node $x_j = j\Delta x$ gives

$$y''(x_j) \approx \frac{y_{j+1} - 2y_j + y_{j-1}}{\Delta x^2} =: Q(\Delta x). \tag{5}$$

Here, $y_j := y(x_j)$. Next, we will prove that this approximation is second order accurate, that is $|y''(x_j) - Q(\Delta x)| = \mathcal{O}(\Delta x^2)$.

Using Taylor's Theorem around $x = x_j$ gives

$$y_{j+1} = y(x_j + \Delta x) = y(x_j) + \Delta x y'(x_j) + \tfrac{\Delta x^2}{2} y''(x_j) + \tfrac{\Delta x^3}{3!} y'''(x_j) + \tfrac{\Delta x^4}{4!} y''''(\eta_+),$$

$$y_{j-1} = y(x_j - \Delta x) = y(x_j) - \Delta x y'(x_j) + \tfrac{\Delta x^2}{2} y''(x_j) - \tfrac{\Delta x^3}{3!} y'''(x_j) + \tfrac{\Delta x^4}{4!} y''''(\eta_-). \tag{6}$$

Here, η_+ and η_- are numbers within the intervals (x_j, x_{j+1}) and (x_{j-1}, x_j), respectively. Substitution of these expressions into $Q(\Delta x)$ gives

$$|y''(x_j) - Q(\Delta x)| = \mathcal{O}(\Delta x^2).$$

This leads to the following discretisation formula for internal grid nodes:

$$\frac{-w_{j-1} + 2w_j - w_{j+1}}{\Delta x^2} + (x_j + 1)w_j = x_j^3 + x_j^2 - 2. \tag{7}$$

Here, w_j represents the numerical approximation of the solution y_j. To deal with the boundary $x = 0$, we use a virtual node at $x = -\Delta x$, and we define $y_{-1} := y(-\Delta x)$. Then, using central differences at $x = 0$ gives

$$0 = y'(0) \approx \frac{y_1 - y_{-1}}{2\Delta x} =: Q_b(\Delta x). \tag{8}$$

Using Taylor's Theorem, gives

$$\begin{aligned} Q_b(\Delta x) &= \\ &= \frac{y(0) + \Delta x y'(0) + \frac{\Delta x^2}{2} y''(0) + \frac{\Delta x^3}{3!} y'''(\eta_+)}{2\Delta x} \\ &- \frac{y(0) - \Delta x y'(0) + \frac{\Delta x^2}{2} y''(0) - \frac{\Delta x^3}{3!} y'''(\eta_-)}{2\Delta x} \\ &= y'(0) + \mathcal{O}(\Delta x^2). \end{aligned}$$

Again, we get an error of $\mathcal{O}(\Delta x^2)$.

(b) With respect to the numerical approximation at the virtual node, we get

$$\frac{w_1 - w_{-1}}{2\Delta x} = 0 \quad \Leftrightarrow \quad w_{-1} = w_1. \tag{9}$$

The discretisation at $x = 0$ is given by

$$\frac{-w_{-1} + 2w_0 - w_1}{\Delta x^2} + w_0 = -2. \tag{10}$$

Substitution of equation (9) into the above equation, yields

$$\frac{2w_0 - 2w_1}{\Delta x^2} + w_0 = -2. \tag{11}$$

Subsequently, we consider the boundary $x = 1$. To this extent, we consider its neighbouring point x_{n-1} and substitute the boundary condition $w_n = y(1) = y_n = 1$ into equation (7) to obtain

$$\frac{-w_{n-2} + 2w_{n-1}}{\Delta x^2} + (x_{n-1} + 1)w_{n-1} \tag{12}$$

$$= x_{n-1}^3 + x_{n-1}^2 - 2 + \frac{1}{\Delta x^2} \tag{13}$$

$$= (1 - \Delta x)^3 + (1 - \Delta x)^2 - 2 + \frac{1}{\Delta x^2}. \tag{14}$$

This concludes our discretisation of the boundary conditions. In order to get a symmetric discretisation matrix, one divides equation (11) by 2.

Next, we use $\Delta x = 1/3$. From equations (7, 11, 14) we obtain the following system

$$\begin{aligned} 9\frac{1}{2}w_0 - 9w_1 &= -1 \\ -9w_0 + 19\frac{1}{3}w_1 - 9w_2 &= -\frac{50}{27} \\ -9w_1 + 19\frac{2}{3}w_2 &= \frac{209}{27}. \end{aligned}$$

(c) The Gershgorin circle theorem states that the eigenvalues of a square matrix $\mathbf{A}$ are located in the complex plane in the union of circles

$$|z - a_{ii}| \leq \sum_{\substack{j \neq i \\ j=1}}^{n} |a_{ij}| \quad \text{where} \quad z \in \mathbb{C} \tag{15}$$

For the 3×3 matrix derived in part (b) we have

- For $i = 1$:
$$\left| z - 9\frac{1}{2} \right| \leq 9 \quad \Rightarrow \quad |\lambda_1|_{\min} \geq \frac{1}{2} \tag{16}$$
- For $i = 2$:
$$\left| z - 19\frac{1}{3} \right| \leq 18 \quad \Rightarrow \quad |\lambda_2|_{\min} \geq 1\frac{1}{3} \tag{17}$$
- For $i = 3$:
$$\left| z - 19\frac{2}{3} \right| \leq 9 \quad \Rightarrow \quad |\lambda_3|_{\min} \geq 10\frac{2}{3} \tag{18}$$

Hence, a lower bound for the smallest eigenvalue is $\frac{1}{2}$. For a symmetric matrix $\mathbf{A}$ we have

$$\|\mathbf{A}^{-1}\| = \frac{1}{|\lambda|_{\min}} \leq 2 \tag{19}$$

This proves that the finite-difference scheme is stable, e.g., with constant $C = 2$.