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

ANSWERS OF THE TEST NUMERICAL METHODS FOR DIFFERENTIAL EQUATIONS (WI3097 TU/Minor AESB2210) Thursday February 2nd 2017, 18:30-21:30

1. (a) The amplification factor can be derived as follows. Consider the test equation $y' = \lambda y$. Application of the Trapezoidal rule to this equation gives:

$$w_{j+1} = w_j + \frac{\Delta t}{2}\left(\lambda w_j + \lambda w_{j+1}\right) \tag{1}$$

Rearranging of w_{j+1} and w_j in (1) yields

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

It now follows that

$$w_{j+1} = \frac{1 + \frac{\Delta t}{2}\lambda}{1 - \frac{\Delta t}{2}\lambda} w_j,$$

and thus

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

(b) The definition of the local truncation error is

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

The exact solution of the test equation is given by

$$y_{j+1} = e^{\lambda\Delta t} y_j.$$

Combination of these results shows that the local truncation error of the test equation is determined by the difference between the exponential function and the amplification factor $Q(\lambda\Delta t)$

$$\tau_{j+1} = \frac{e^{\lambda\Delta t} - Q(\lambda\Delta t)}{\Delta t} y_j. \tag{2}$$

The difference between the exponential function and amplification factor can be computed as follows. The Taylor series of $e^{\lambda\Delta t}$ with known point 0 is:

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

The Taylor series of $\frac{1}{1-\frac{\Delta t}{2}\lambda}$ with known point 0 is:

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

With (4) it follows that $\frac{1+\lambda\frac{\Delta t}{2}}{1-\lambda\frac{\Delta t}{2}}$ is equal to

$$\frac{1+\lambda\frac{\Delta t}{2}}{1-\lambda\frac{\Delta t}{2}} = 1 + \lambda\Delta t + \frac{1}{2}(\lambda\Delta t)^2 + \mathcal{O}((\Delta t)^3). \tag{5}$$

In order to determine $e^{\lambda\Delta t} - Q(\lambda\Delta t)$, we subtract (5) from (3). Now it follows that

$$e^{\lambda\Delta t} - Q(\lambda\Delta t) = \mathcal{O}((\Delta t)^3). \tag{6}$$

The local truncation error can be found by substituting (6) into (2), which leads to

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

(c) The Trapezoidal rule is stable if

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

Using the complex valued $\lambda = \mu + i\nu$ it appears that the condition is equal to:

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

This is equivalent with

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

Since $\mu \leq 0$ it easily follows that

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

which implies that

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

and the method is stable.

(d) Application of the Trapezoidal rule to

$$y' = -(1+2t)y + t, \text{ with } y(0) = 1,$$

and step size $\Delta t = \frac{1}{2}$ gives:

$$w_1 = w_0 + \frac{\Delta t}{2}[-w_0 + 0 - 2w_1 + \frac{1}{2}].$$

Using the initial value $w_0 = y(0) = 1$ and step size $\Delta t = \frac{1}{2}$ gives:

$$w_1 = 1 + \frac{1}{4}[-1 - 2w_1 + \frac{1}{2}].$$

This leads to

$$1\frac{1}{2}w_1 = \frac{7}{8}, \text{ so } w_1 = \frac{7}{12}.$$

(e) For the comparison we use the following items: accuracy, stability, and amount of work. Below we make the comparison:

- Accuracy: since the error of Euler Forward is $O(\Delta t)$ and that of the Trapezoidal rule is $O((\Delta t)^2)$, the error is smaller for the Trapezoidal rule.
- Stability: since the value of $-(1+2t)$ is always negative the Trapezoidal rule is stable for all step sizes, whereas for Euler Forward the step size should satisfy the inequality $\Delta t \leq \frac{2}{1+2t}$.
- Amount of work: since the differential equation is linear the amount of work for the implicit Trapezoidal rule is comparable to the work of the explicit Euler Forward method.

From the above comparisons we conclude that for this problem the Trapezoidal rule is preferred.

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

$$y''(x_j) \approx \frac{y_{j+1} - 2y_j + y_{j-1}}{\Delta x^2} =: Q(\Delta x). \tag{7}$$

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 series expansion* around $x = x_j$, gives

$$\begin{aligned} y_{j+1} &= y(x_j + \Delta x) = y(x_j) + \Delta x y'(x_j) + \tfrac{\Delta x^2}{2} y''(x_j) + \tfrac{\Delta x^3}{3!} y'''(x_j) + \tfrac{\Delta x^4}{4!} y''''(\eta_+), \\ y_{j-1} &= y(x_j - \Delta x) = y(x_j) - \Delta x y'(x_j) + \tfrac{\Delta x^2}{2} y''(x_j) - \tfrac{\Delta x^3}{3!} y'''(x_j) + \tfrac{\Delta x^4}{4!} y''''(\eta_-). \end{aligned} \tag{8}$$

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 $|y''(x_j) - Q(\Delta x)| = \mathcal{O}(\Delta x^2)$. Therewith, we obtain the following discretization formula for the internal grid nodes:

$$\frac{-w_{j-1} + 2w_j - w_{j+1}}{\Delta x^2} + x_j w_j = x_j^3 - 2. \tag{9}$$

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

$$0 = y'(0) \approx \frac{y_1 - y_{-1}}{2\Delta x} =: Q_b(\Delta x). \tag{10}$$

Using *Taylor series expansion*, gives

$$\begin{gathered} Q_b(\Delta x) = \\ \frac{y(0) + \Delta x y'(0) + \frac{\Delta x^2}{2} y''(0) + \frac{\Delta x^3}{3!} y'''(\eta_+) - (y(0) - \Delta x y'(0) + \frac{\Delta x^2}{2} y''(0) - \frac{\Delta x^3}{3!} y'''(\eta_-))}{2\Delta x} = \\ y'(0) + \mathcal{O}(\Delta x^2). \end{gathered} \tag{11}$$

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

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

$$\frac{w_1 - w_{-1}}{2\Delta x} = 0 \Leftrightarrow w_{-1} = w_1. \tag{12}$$

The discretization at $x = 0$ is given by

$$\frac{-w_{-1} + 2w_0 - w_1}{\Delta x^2} = -2. \tag{13}$$

Substitution of equation (12) into the above equation, yields

$$\frac{2w_0 - 2w_1}{\Delta x^2} = -2. \tag{14}$$

Subsequently, we consider the boundary $x = 1$. To this extent, we consider its neighboring point x_{n-1}, here substitution of the boundary condition $w_n = y(1) = y_n = 1$ into equation (9), gives

$$\frac{-w_{n-2} + 2w_{n-1}}{\Delta x^2} + x_{n-1} w_{n-1} = x_{n-1}^3 - 2 + \frac{1}{\Delta x^2} = (1 - \Delta x)^3 - 2 + \frac{1}{\Delta x^2}. \tag{15}$$

This concludes our discretization of the boundary conditions. In order to get a symmetric discretization matrix, one divides equation (14) by 2.

(b) Next, we use $\Delta x = 1/3$, then, from equations (9), (14), and (15), one obtains the following system

$$\begin{aligned} 9w_0 - 9w_1 &= -1 \\ -9w_0 + 18\frac{1}{3}w_1 - 9w_2 &= -\frac{53}{27} \\ -9w_1 + 18\frac{2}{3}w_2 &= \frac{197}{27}. \end{aligned}$$

3. (a) We compute

$$x + y = 2/3 + 1999/3000 = 1.333,$$

and

$$x - y = 2/3 - 1999/3000 = 1/3000 = 0.3333\ldots \cdot 10^{-3}.$$

Further, we have $fl(x) = 0.6667$, $fl(y) = 0.6663$, and

$$fl(x) + fl(y) = 0.1333 \cdot 10^1,$$

hence $fl(fl(x) + fl(y)) = 0.1333 \cdot 10^1$.
For the subtraction, one obtains

$$fl(x) - fl(y) = 0.4 \cdot 10^{-3},$$

and hence

$$fl(fl(x) - fl(y)) = fl(0.4 \cdot 10^{-3}) = 0.4000 \cdot 10^{-3}.$$

(b) After the addition, the relative error is given by

$$|\frac{0.1333 \cdot 10^1 - 1.333}{0.1333 \cdot 10^1}| = 0,$$

and after the subtraction, one gets

$$|\frac{0.4000 \cdot 10^{-3} - 0.3333\ldots \cdot 10^{-3}}{0.3333\ldots \cdot 10^{-3}}| = 0.2.$$

(c) The relative error due to subtraction of two positive numbers is divided by the difference between these numbers. If this difference gets arbitrarily small, then the relative error gets arbitrarily large for a given absolute error.