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

ANSWERS OF THE TEST NUMERICAL METHODS FOR DIFFERENTIAL EQUATIONS
(WI3097TU WI3097Minor WI3197Minor AESB2210 AESB2210-18 CTB2400)
Thursday July 4th 2019, 13:30-16:30

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

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

in which we determine y_{n+1} by the use of Taylor expansions around t_n:

$$y_{n+1} = y_n + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(t_n) + \mathcal{O}(\Delta t^3). \tag{2}$$

We bear in mind that

$$y'(t_n) = f(t_n, y_n)$$

$$\begin{aligned} y''(t_n) &= \frac{df(t_n, y_n)}{dt} = \frac{\partial f(t_n, y_n)}{\partial t} + \frac{\partial f(t_n, y_n)}{\partial y} y'(t_n) \\ &= \frac{\partial f(t_n, y_n)}{\partial t} + \frac{\partial f(t_n, y_n)}{\partial y} f(t_n, y_n). \end{aligned}$$

Hence

$$y_{n+1} = y_n + \Delta t y'(t_n) + \frac{\Delta t^2}{2}\left(\frac{\partial f(t_n, y_n)}{\partial t} + \frac{\partial f(t_n, y_n)}{\partial y} f(t_n, y_n)\right) + \mathcal{O}(\Delta t^3). \tag{3}$$

After substitution of the predictor $z^*_{n+1} = y_n + \Delta t f(t_n, y_n)$ into the corrector, and after using a Taylor expansion around (t_n, y_n), we obtain for z_{n+1}:

$$\begin{aligned} z_{n+1} &= y_n + \frac{\Delta t}{2}\left(f(t_n, y_n) + f(t_n + \Delta t, y_n + \Delta t f(t_n, y_n))\right) \\ &= y_n + \frac{\Delta t}{2}\left(2f(t_n, y_n) + \Delta t\left(\frac{\partial f(t_n, y_n)}{\partial t} + f(t_n, y_n)\frac{\partial f(t_n, y_n)}{\partial y}\right) + \mathcal{O}(\Delta t^2)\right). \end{aligned}$$

Herewith, one obtains

$$y_{n+1} - z_{n+1} = \mathcal{O}(\Delta t^3), \text{ and hence } \tau_{n+1}(\Delta t) = \frac{\mathcal{O}(\Delta t^3)}{\Delta t} = \mathcal{O}(\Delta t^2). \tag{4}$$

(b) Let $x_1 = y$ and $x_2 = y'$, then $y'' = x_2'$, and hence

$$\begin{aligned} x_2' + 4x_1 &= \cos(t), \\ x_1' &= x_2. \end{aligned} \tag{5}$$

We write this as

$$\begin{cases} x_1' = x_2, \\ x_2' = -4x_1 + \cos(t). \end{cases} \tag{6}$$

Finally, this is represented in the following matrix-vector form:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ -4 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos(t) \end{pmatrix}. \tag{7}$$

In which, we have the following matrix $A = \begin{pmatrix} 0 & 1 \\ -4 & 0 \end{pmatrix}$ and $\underline{f} = \begin{pmatrix} 0 \\ \cos(t) \end{pmatrix}$. The initial conditions are defined by $\begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

(c) *Note: Every miscalculation in the calculation of $\underline{w}_1^*$ gives a subtraction of 1/4 point, with at most 1/2 point being subtracted.*

Note: The calculation of $\underline{w}_1$ must be consistent with the value for $\underline{w}_1^*$. If not, 1 point is subtracted.

Note: Every miscalculation in the calculation of $\underline{w}_1$ gives a subtraction of 1/4 point, with at most 1 point being subtracted.

Application of the integration method to the system $\underline{x}' = A\underline{x} + \underline{f}$, gives

$$\begin{aligned} \underline{w}_1^* &= \underline{w}_0 + \Delta t \left(A\underline{w}_0 + \underline{f}_0 \right), \\ \underline{w}_1 &= \underline{w}_0 + \tfrac{\Delta t}{2} \left(A\underline{w}_0 + f_0 + A\underline{w}_1^* + \underline{f}_1 \right). \end{aligned} \tag{8}$$

With the initial condition $\underline{w}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\Delta t = 0.1$, this gives the following result for the predictor

$$\underline{w}_1^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{1}{10} \left(\begin{pmatrix} 0 & 1 \\ -4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = \begin{pmatrix} 1 \\ -0.3 \end{pmatrix}. \tag{9}$$

The corrector is calculated as follows

$$\begin{aligned} \underline{w}_1 &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{1}{20} \left(\begin{pmatrix} 0 & 1 \\ -4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ -0.3 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos(\frac{1}{10}) \end{pmatrix} \right) \\ &= \begin{pmatrix} 0.9850 \\ -0.3002 \end{pmatrix} \end{aligned}$$

(d) Consider the test equation $y' = \lambda y$, then one gets

$$\begin{aligned} w_{n+1}^* &= w_n + \Delta t \lambda w_n = (1 + \Delta t \lambda) w_n, \\ w_{n+1} &= w_n + \frac{\Delta t}{2} (\lambda w_n + \lambda w_{n+1}^*) \\ &= w_n + \frac{\Delta t}{2} (\lambda w_n + \lambda (w_n + \Delta t \lambda w_n)) \\ &= \left(1 + \Delta t \lambda + \frac{(\Delta t \lambda)^2}{2} \right) w_n. \end{aligned}$$

Hence the amplification factor is given by

$$Q(\lambda \Delta t) = 1 + \lambda \Delta t + \frac{(\lambda \Delta t)^2}{2}. \tag{10}$$

(e) *Note: Every miscalculation in the calculation of $|Q(\lambda_1 \Delta t)|^2$ gives a subtraction of 1/4 point, with at most 1/2 point being subtracted.*

Note: The calculation of $|Q(\lambda_1 \Delta t)|^2$ *must be consistent with the eigenvalues found. If not,* $1/2$ *point is subtracted.*

First, we determine the eigenvalues of the matrix A. Subsequently, the eigenvalues are substituted into the amplification factor.

The eigenvalues of the matrix A are given by $\lambda_1 = 2i$ and $\lambda_2 = -2i$.

Since we have complex eigenvalues it is sufficient to check when $|Q(\lambda_1 \Delta t)| \leq 1$.

Note that

$$Q(\lambda_1 \Delta t) = 1 + 2i\Delta t + \frac{(2i\Delta t)^2}{2}$$

$$Q(\lambda_1 \Delta t) = 1 - 2(\Delta t)^2 + 2i\Delta t.$$

This implies that

$$|Q(\lambda_1 \Delta t)|^2 = (1 - 2\Delta t^2)^2 + (2\Delta t)^2$$

and thus

$$|Q(\lambda_1 \Delta t)|^2 = 1 - 4\Delta t^2 + 4\Delta t^4 + 4\Delta t^2 = 1 + 4\Delta t^4.$$

It is easy to see that $|Q(\lambda_1 \Delta t)|^2 \leq 1$ only for $\Delta t = 0$, which is not a valid step size to integrate the initial value problem.

Therefore the considered method is never stable for the given problem.

2. (a) The equation that needs to be solved is

$$f(p_0) + \frac{f(p_1) - f(p_0)}{p_1 - p_0}(p_2 - p_0) = 0.$$

Solving this equation gives the steps:

$$\begin{aligned}
& & \frac{f(p_1) - f(p_0)}{p_1 - p_0}(p_2 - p_0) &= -f(p_0), \\
&\Rightarrow & p_2 - p_0 &= -\frac{p_1 - p_0}{f(p_1) - f(p_0)} f(p_0), \\
&\Rightarrow & p_2 &= p_0 - \frac{p_1 - p_0}{f(p_1) - f(p_0)} f(p_0).
\end{aligned}$$

We write the above as one quotient:

$$\begin{aligned}
p_2 &= \frac{f(p_1) - f(p_0)}{f(p_1) - f(p_0)} p_0 - \frac{p_1 - p_0}{f(p_1) - f(p_0)} f(p_0), \\
\Rightarrow \ p_2 &= \frac{p_0 f(p_1) - p_1 f(p_0)}{f(p_1) - f(p_0)}. \qquad (11)
\end{aligned}$$

Now we have two options:

A. Rewrite the above formula to the form given in the exercise, with $n = 2$, and conclude the formula for K_1;

B. Fill in the formula for K_1 into the formula for p_n, with $n = 2$, given in the exercise and show this results in the same formula.

Note: Only one of the options has to be present within your answer and earns at most ½ point.

Option A: We can rewrite Equation (11) to:

$$\begin{aligned}
p_2 &= \frac{p_0 f(p_1) - p_1 f(p_0)}{f(p_1) - f(p_0)}, \\
\Rightarrow \ p_2 &= \frac{p_0 f(p_1) - p_1 f(p_0) - p_1 f(p_1) + p_1 f(p_1)}{f(p_1) - f(p_0)}, \\
\Rightarrow \ p_2 &= \frac{p_1 (f(p_1) - f(p_0)) - (p_1 - p_0) f(p_1)}{f(p_1) - f(p_0)}, \\
\Rightarrow \ p_2 &= p_1 - \frac{p_1 - p_0}{f(p_1) - f(p_0)} f(p_1),
\end{aligned}$$

which is indeed of the form given in the exercise. Therefore, K_1 indeed has the formula

$$K_1 = \frac{f(p_1) - f(p_0)}{(p_1 - p_0)}.$$

Option B: The formula of the exercise, with $n = 2$ and the given formula for K_1 is:

$$p_2 = p_1 - \frac{p_1 - p_0}{f(p_1) - f(p_0)} f(p_1).$$

We write the above as one quotient:

$$
\begin{aligned}
p_2 &= \frac{f(p_1)-f(p_0)}{f(p_1)-f(p_0)}p_1 - \frac{p_1-p_0}{f(p_1)-f(p_0)}f(p_1),\\
\Rightarrow\ p_2 &= \frac{p_0 f(p_1)-p_1 f(p_0)}{f(p_1)-f(p_0)}.
\end{aligned}
$$

The above equation is equal to Equation (11). Therefore, K_1 indeed has the formula

$$
K_1 = \frac{f(p_1)-f(p_0)}{p_1-p_0}.
$$

(b) *Note: Every miscalculation in the calculation of K_1 gives a subtraction of 1/4 point, with at most 1/2 point being subtracted.*

Given that $p_0 = 1$ and $p_1 = 2$, we first calculate K_1, using the values from the given table:

$$
\begin{aligned}
K_1 &= \frac{f(p_1)-f(p_0)}{p_1-p_0},\\
&= \frac{f(2)-f(1)}{2-1},\\
&= f(2)-f(1),\\
&= 2-(-1),\\
&= 3.
\end{aligned}
$$

Note: Every miscalculation in the calculation of p_2 gives a subtraction of 1/4 point, with at most 1/2 point being subtracted.

Note: The value of p_2 should be consistent with your value for K_1.

Now p_2 can be calculated with the Secant method, with $n = 2$ and the values from the given table:

$$
\begin{aligned}
p_2 &= p_1 - \frac{f(p_1)}{K_1},\\
&= 2 - \frac{f(2)}{3},\\
&= 2 - \frac{2}{3},\\
&= \frac{4}{3}.
\end{aligned}
$$

(c) The formula for K_2 is given by

$$
K_2 = \frac{f(p_2)-f(p_1)}{p_2-p_1}.
$$

Note: Every miscalculation in the calculation of K_2 gives a subtraction of 1/4 point, with at most 3/4 point being subtracted.

Note: The value of K_2 should be consistent with your formula for K_2.

This formula gives

$$
\begin{aligned}
K_2 &= \frac{f(p_2) - f(p_1)}{p_2 - p_1}, \\
&= \frac{f(\frac{4}{3}) - f(2)}{\frac{4}{3} - 2}, \\
&= \frac{\left(-\frac{2}{9}\right) - 2}{-\frac{2}{3}}, \\
&= \frac{-\frac{20}{9}}{-\frac{2}{3}}, \\
&= \frac{10}{3},
\end{aligned}
$$

Note: Every miscalculation in the calculation of p_3 gives a subtraction of $^1/_4$ point, with at most $^3/_4$ point being subtracted.
Note: The value of p_3 should be consistent with your value for K_2.
and finally

$$
\begin{aligned}
p_3 &= p_2 - \frac{f(p_2)}{K_2}, \\
&= \frac{4}{3} - \frac{f(\frac{4}{3})}{\frac{10}{3}}, \\
&= \frac{4}{3} - \frac{-\frac{2}{9}}{\frac{10}{3}}, \\
&= \frac{4}{3} - \frac{1}{15}, \\
&= \frac{7}{5}.
\end{aligned}
$$

3. (a) The right composite Rectangle rule is given by

$$\int_a^b y(x)dx \approx h\sum_{j=1}^{n} y(x_j),$$

with $hn = b - a$ and $x_j = a + jh$ for $j = 0, \ldots, n$.
From $h = \pi/2$, $a = 0$ and $b = 2\pi$, it follows that $n = 4$ and the following table also follows:

j	0	1	2	3	4
x_j	0	$\pi/2$	π	$3\pi/2$	2π
$y(x_j)$	1	2	1	0	1

Note: Every miscalculation in the calculation below gives a subtraction of 1/2 point, with at most 1 point being subtracted.
Applying the right composite Rectangle rule with $h = \pi/2$ gives

$$\begin{aligned}\int_0^{2\pi} y(x)dx &\approx \frac{\pi}{2}\left(y\left(\frac{\pi}{2}\right) + y(\pi) + y\left(\frac{3\pi}{2}\right) + y(2\pi)\right),\\ &= \frac{\pi}{2}(2+1+0+1),\\ &= 2\pi.\end{aligned}$$

(b) The composite Trapezoidal rule is given by

$$\int_a^b y(x)dx \approx h\sum_{j=1}^{n} \frac{1}{2}\left(y(x_{j-1}) + y(x_j)\right),$$

with $hn = b - a$ and $x_j = a + jh$ for $j = 0, \ldots, n$.
Note: Every miscalculation in the calculation below gives a subtraction of 1/2 point, with at most 1/2 point being subtracted.
Applying the composite Trapezoidal rule with $h = \pi/2$ gives

$$\begin{aligned}\int_0^{2\pi} y(x)dx &\approx \frac{\pi}{2}\left(\frac{1}{2}y(0) + y\left(\frac{\pi}{2}\right) + y(\pi) + y\left(\frac{3\pi}{2}\right) + \frac{1}{2}y(2\pi)\right),\\ &= \frac{\pi}{2}\left(\frac{1}{2}+2+1+0+\frac{1}{2}\right),\\ &= 2\pi.\end{aligned}$$

(c) *Note: Your answers should be consistent with each other. For each inconsistency 1/4 point will be subtracted, with at most 1 1/2 points being subtracted.*
The derivatives of the function y are given by

$$\begin{aligned}y'(x) &= \cos(x),\\ y''(x) &= -\sin(x).\end{aligned}$$

From this it follows

$$\begin{aligned}\max_{x\in[0,2\pi]} |y'(x)| &= 1,\\ \max_{x\in[0,2\pi]} |y''(x)| &= 1.\end{aligned}$$

Therefore the explicit upper bounds for ε_R and ε_T are given by

$$\varepsilon_R \leq \pi h,$$
$$\varepsilon_T \leq \frac{\pi}{6} h^2.$$

(d) *Note: No points are given if one of the following holds:*

- *no arguments are presented;*
- *the selected method is inconsistent with the arguments.*

Note: Incorrect arguments on topics other than the amount of work and accuracy give per such argument a subtraction of $1/4$ *point, with at most* $1/2$ *points being subtracted.*

From the above upper bounds one can conclude that

$$\varepsilon_T < \varepsilon_R$$

if $h < 6$. Hence, the error for the composite Trapezoidal method is much smaller for small h then the error for the right composite Rectangle rule.

Furthermore, with $n = b-a/h$, the number of function evaluations of the right composite Rectangle rule is n, and $n+1$ for the composite Trapezoidal rule. It also holds that

$$\frac{n+1}{n} \approx 1,$$

for large n. Hence, for small h the amount of work within both methods is similar. Therefore the composite Trapezoidal method should be preferred for small h.