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

ANSWERS OF THE TEST NUMERICAL METHODS FOR DIFFERENTIAL EQUATIONS (CTB2400 WI3097 TU)
Thursday June 30th 2016, 18:30-21:30

1. (a) Replace $f(t,y)$ by λy in the RK_4 formulas:

$$\begin{aligned}
k_1 &= \lambda\Delta t w_n \\
k_2 &= \lambda\Delta t(w_n + \tfrac{1}{2}k_1) = \lambda\Delta t(1+\tfrac{1}{2}\lambda\Delta t)w_n \\
k_3 &= \lambda\Delta t(w_n + \tfrac{1}{2}k_2) = \lambda\Delta t(1+\tfrac{1}{2}\lambda\Delta t(1+\tfrac{1}{2}\lambda\Delta t))w_n \\
k_4 &= \lambda\Delta t(w_n + k_3) = \lambda\Delta t(1+\lambda\Delta t(1+\tfrac{1}{2}\lambda\Delta t(1+\tfrac{1}{2}\lambda\Delta t))w_n
\end{aligned}$$

Substitution of these expressions into:

$$w_{n+1} = w_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

and collecting like powers of $\lambda\Delta t$ yields:

$$w_{n+1} = [1 + \lambda\Delta t + \frac{1}{2}(\lambda\Delta t)^2 + \frac{1}{6}(\lambda\Delta t)^3 + \frac{1}{24}(\lambda\Delta t)^4]w_n.$$

The amplification factor is therefore:

$$Q(\lambda\Delta t) = 1 + \lambda\Delta t + \frac{1}{2}(\lambda\Delta t)^2 + \frac{1}{6}(\lambda\Delta t)^3 + \frac{1}{24}(\lambda\Delta t)^4.$$

(b) The local truncation error is defined as

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

where z_{n+1} is the numerical solution at t_{n+1}, obtained by starting from the exact value $y(t_n)$ in stead of w_n. Repeating the derivation under (a), with w_n replaced by $y(t_n)$, gives:

$$z_{n+1} = Q(\lambda\Delta t)y(t_n).$$

Using furthermore $y(t_{n+1}) = e^{\lambda\Delta t}y(t_n)$ in (1) it follows that

$$\tau_{n+1} = \frac{e^{\lambda\Delta t} - Q(\lambda\Delta t)}{\Delta t}y(t_n).$$

Canceling the first five terms of the expansion of $e^{\lambda\Delta t}$ against $Q(\lambda\Delta t)$, the required order of magnitude of τ_{n+1} follows.

(c) Use the transformation:

$$\begin{aligned} y_1 &= y\ , \\ y_2 &= y'\ , \end{aligned}$$

This implies that

$$\begin{aligned} y_1' &= y' = y_2\ , \\ y_2' &= y'' = -qy_1 - py_2 + \sin t\ , \end{aligned}$$

So the matrix **A** and vector **g** are:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -q & -p \end{pmatrix}; \quad \mathbf{g}(t) = \begin{pmatrix} 0 \\ \sin t \end{pmatrix}.$$

Characteristic equation: $\lambda^2 + p\lambda + q = 0$. $\lambda_{1,2} = \frac{-p \pm \sqrt{p^2-4q}}{2}$.

(d) Substitution of the values of p and q into the matrix **A** yields the eigenvalues $\lambda_{1,2} = -500 \pm i$. From the given drawing of the stability region the following can be inferred. Because the imaginary part is much smaller than the real part, an approximate stability condition can be obtained by simply neglecting the imaginary part. Then $\Delta t \leq 2.8/500 = 0.0056$ follows as the stability condition.

(e)

$$y'' + py' + qy = \sin t, \quad y(0) = y_0,\ y'(0) = y_0'. \tag{2}$$

After a short time the solution is close to a linear combination of $\sin t$ and $\cos t$, which is called a smooth solution.

The smooth solution can be integrated accurately by RK_4 with a 'large' step size: a step size of 0.1, let us say, would give an error of order 10^{-4} which is sufficient for most engineering purposes. However stability, governed by the eigenvalues, requires that the step size be restricted (see part (d)) to 0.0056. So the stability requirement forces us to choose a step size yielding an unnecessarily accurate solution, which is inefficient.

The Trapezoidal rule, on the other hand, is stable for all step sizes. So the step size is restricted by accuracy requirements only. The Trapezoidal rule has a global error of order Δt^2 such that a good accuracy may be expected for step sizes of about 0.01, which is much larger than the step size for RK4: 0.0056. An efficiency gain may be obtained in spite of the extra work connected with the implicitness of the method.

2. (a) Consider an interval of integration $[x_{j-1}, x_j]$, then the **Rectangle Rule** reads

$$I_j^R = hf(x_{j-1}), \qquad h = x_j - x_{j-1}. \tag{3}$$

The *composed integration rule* is derived by

$$I^R = h(I_1^R + I_2^R + \ldots + I_n^R) = h(f(x_0) + \ldots + f(x_{n-1})), \tag{4}$$

which yields

$$I^R = \frac{1}{3} \cdot (0 + (\frac{1}{3})^2 + (\frac{2}{3})^2) = \frac{5}{27}. \tag{5}$$

(b) For the interval of integration $[x_{j-1}, x_j]$ the **Trapezoidal Rule** is

$$I_j^T = \frac{h}{2}(f(x_{j-1}) + f(x_j)). \tag{6}$$

The *composed integration rule* is derived by

$$I^T = h(I_1^T + I_2^T + \ldots + I_n^T) = h(\frac{f(x_0)}{2} + f(x_1) + \ldots + f(x_{n-1}) + \frac{f(x_n)}{2}), \tag{7}$$

which leads to

$$I^T = \frac{1}{3} \cdot (0 + (\frac{1}{3})^2 + (\frac{2}{3})^2 + \frac{1}{2}) = \frac{19}{54}. \tag{8}$$

(c) For a general number of subintervals, say n, the magnitude of the composed Rectangle- and Trapezoidal Rules, is bounded from above by

$$\begin{aligned} \varepsilon_R &\leq \frac{h}{2} \max_{x \in [0,1]} |y'(x)| \leq h = \frac{1}{n}, \\ \varepsilon_T &\leq \frac{h^2}{12} \max_{x \in [0,1]} |y''(x)| \leq \frac{h^2}{6} = \frac{1}{6n^2}. \end{aligned} \tag{9}$$

Here, the exact solution $y(x) = x^2$ was used. Hence, the error from the Trapezoidal Rule is much smaller. Furthermore, from the composed Rules, it is easy to see that the number of function evaluations for the composed Rectangle- and Trapezoidal Rules is given by n and $n + 1$, respectively. Since

$$\lim_{n \to \infty} \frac{n+1}{n} = 1, \tag{10}$$

it follows that the amount of work for the Trapezoidal Rule is not significantly higher than it is for the Rectangle Rule. Hence, it is more attractive to use the Trapezoidal Rule.

3. (a) **Newton-Raphson's method** is an iterative method to find $p \in \mathbb{R}$ such that $f(p) = 0$. Suppose $f \in C^2[a, b]$. Let $\bar{x} \in [a, b]$ be an approximation of the root p such that $f'(\bar{x}) \neq 0$, and suppose that $|p - \bar{x}|$ is small. Consider the first-degree Taylor polynomial about $\bar{x}$:

$$f(x) = f(\bar{x}) + (x - \bar{x})f'(\bar{x}) + \frac{(x - \bar{x})^2}{2} f''(\xi(x)), \tag{11}$$

in which $\xi(x)$ between x and $\bar{x}$. Using that $f(p) = 0$, equation (11) yields

$$0 = f(\bar{x}) + (p - \bar{x})f'(\bar{x}) + \frac{(p - \bar{x})^2}{2} f''(\xi(x)).$$

Because $|p - \bar{x}|$ is small, $(p - \bar{x})^2$ can be neglected, such that

$$0 \approx f(\bar{x}) + (p - \bar{x})f'(\bar{x}).$$

Note that the right-hand side is the formula for the tangent in $(\bar{x}, f(\bar{x}))$. Solving for p yields

$$p \approx \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})}.$$

This motivates the Newton-Raphson method, that starts with an approximation p_0 and generates a sequence $\{p_n\}$ by

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{for} \quad n \geq 1.$$

Remark 1 *One can also give a graphical derivation following Figure 4.2 from the book.*

(b) The first derivative of g equals

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

Substitution of $f(x) = \sin(x)$, $f'(x) = \cos(x)$ and $f''(x) = -\sin(x)$ yields

$$g'(x) = -\frac{\sin^2(x)}{\cos^2(x)} = -\tan^2(x).$$

Since $\tan(-\sqrt{2}/2) = -1$, $\tan(\sqrt{2}/2) = 1$ and the tangent function is monotonically increasing on the interval $[-1, 1]$ any initial guess inside the interval $(-1, 1)$ will lead to a convergent iteration process.

(c) It follows from the linearization of the function $\mathbf{f}$ about the iterate $\mathbf{x}_{n-1}$ that

$$f_1(\mathbf{p}) \approx f_1(\mathbf{p}^{(n-1)}) + \frac{\partial f_1}{\partial p_1}(\mathbf{p}^{(n-1)})(p_1 - p_1^{(n-1)}) + \ldots + \frac{\partial f_1}{\partial p_m}(\mathbf{p}^{(n-1)})(p_m - p_m^{(n-1)}),$$
$$\vdots$$
$$f_m(\mathbf{p}) \approx f_m(\mathbf{p}^{(n-1)}) + \frac{\partial f_m}{\partial p_1}(\mathbf{p}^{(n-1)})(p_1 - p_1^{(n-1)}) + \ldots + \frac{\partial f_m}{\partial p_m}(\mathbf{p}^{(n-1)})(p_m - p_m^{(n-1)}).$$

Defining the Jacobian matrix of $\mathbf{f}(\mathbf{x})$ by

$$\mathbf{J}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \ldots & \frac{\partial f_1}{\partial x_m}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \ldots & \frac{\partial f_m}{\partial x_m}(\mathbf{x}) \end{pmatrix},$$

the linearization can be written in the more compact form

$$\mathbf{f}(\mathbf{p}) \approx \mathbf{f}(\mathbf{p}^{(n-1)}) + \mathbf{J}(\mathbf{p}^{(n-1)})(\mathbf{p} - \mathbf{p}^{(n-1)}).$$

The next iterate, $\mathbf{p}^{(n)}$, is obtained by setting the linearization equal to zero:

$$\mathbf{f}(\mathbf{p}^{(n-1)}) + \mathbf{J}(\mathbf{p}^{(n-1)})(\mathbf{p}^{(n)} - \mathbf{p}^{(n-1)}) = 0, \tag{12}$$

which can be rewritten as

$$\mathbf{J}(\mathbf{p}^{(n-1)})\mathbf{s}^{(n)} = -\mathbf{f}(\mathbf{p}^{(n-1)}), \tag{13}$$

where $\mathbf{s}^{(n)} = \mathbf{p}^{(n)} - \mathbf{p}^{(n-1)}$. The new approximation equals $\mathbf{p}^{(n)} = \mathbf{p}^{(n-1)} + \mathbf{s}^{(n)}$. Finally, Newton-Raphson's formula for general nonlinear problems reads:

$$\mathbf{p}^{(n)} = \mathbf{p}^{(n-1)} - \mathbf{J}^{-1}(\mathbf{p}^{(n-1)})\mathbf{f}(\mathbf{p}^{(n-1)}). \tag{14}$$

(d) First, we rewrite the system into the form

$$\begin{aligned} f_1(w_1, w_2) = 0, \\ f_2(w_1, w_2) = 0, \end{aligned} \tag{15}$$

by setting

$$\begin{aligned} f_1(w_1, w_2) &:= 18w_1 - 9w_2 + (w_1)^2, \\ f_2(w_1, w_2) &:= -9w_1 + 18w_2 + (w_2)^2 - 9. \end{aligned} \tag{16}$$

We denote the Jacobi-matrix by $J(w_1, w_2)$. At the first step we compute

$$\underline{w}^{(1)} = \underline{w}^{(0)} - J(\underline{w}^{(0)})^{-1}F(\underline{w}^{(0)}), \tag{17}$$

where $\underline{w} = [w_1 \; w_2]^T$. Note that

$$J(\underline{w}^{(0)}) = \begin{pmatrix} 18 + 2w_1^{(0)} & -9 \\ -9 & 18 + 2w_2^{(0)} \end{pmatrix}. \tag{18}$$

Using $w_1^{(0)} = w_2^{(0)} = 0$ we obtain:

$$J(\underline{w}^{(0)}) = \begin{pmatrix} 18 & -9 \\ -9 & 18 \end{pmatrix}. \tag{19}$$

This implies that

$$J(\underline{w}^{(0)})^{-1} = \frac{1}{18^2 - 81} \begin{pmatrix} 18 & 9 \\ 9 & 18 \end{pmatrix}. \tag{20}$$

Furthermore

$$F(\underline{w}^{(0)}) = \begin{pmatrix} 0 \\ -9 \end{pmatrix}, \tag{21}$$

so

$$\underline{w}^{(1)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{18^2 - 81} \begin{pmatrix} 18 & 9 \\ 9 & 18 \end{pmatrix} \begin{pmatrix} 0 \\ -9 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix}. \tag{22}$$