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

ANSWERS OF THE TEST NUMERICAL METHODS FOR DIFFERENTIAL EQUATIONS (WI3097 TU)
Thursday January 21 2010, 18:30-21:30

1. The θ-method, used to integrate the initial value problem $y' = f(t, y)$, $y(t_0) = y_0$, is given by
$$w_{n+1} = w_n + h\left[\theta f(t_n, w_n) + (1-\theta) f(t_{n+1}, w_{n+1})\right]. \qquad (1)$$
Here h denotes the timestep and w_n represents the numerical solution at time t_n. Further, let $0 \leq \theta \leq 1$.

   (a) For this purpose, we use the test equation
$$y' = \lambda y, \qquad (2)$$
then the θ-method gives
$$w_{n+1} = w_n + h\left(\theta\lambda w_n + (1-\theta)\lambda w_{n+1}\right). \qquad (3)$$
From this, we get
$$w_{n+1} = w_n \frac{1+\theta h\lambda}{1-(1-\theta)h\lambda}, \qquad (4)$$
and hence the amplification factor is given by
$$Q(h\lambda) = \frac{1+\theta h\lambda}{1-(1-\theta)h\lambda}. \qquad (5)$$

   (b) The local truncation error is defined by:
$$\tau_{n+1}(h) = \frac{y_{n+1} - z_{n+1}}{h}. \qquad (6)$$
For the test equation, the exact solution at $t = t_{n+1}$ is expressed by
$$y_{n+1} = y_n e^{h\lambda} = y_n(1 + h\lambda + \frac{1}{2}h^2\lambda^2 + O(h^3)). \qquad (7)$$
Further, the numerical solution of the test equation is expressed by
$$z_{n+1} = y_n + h\left(\theta\lambda y_n + (1-\theta)\lambda z_{n+1}\right). \qquad (8)$$
This gives
$$z_{n+1} = y_n \frac{1+\theta h\lambda}{1-(1-\theta)h\lambda} = Q(h\lambda) y_n. \qquad (9)$$

Substitution into the definition of the local truncation error (6), gives

$$\tau_{n+1}(h) = \frac{y_n\left(e^{h\lambda} - Q(h\lambda)\right)}{h}. \tag{10}$$

If $|(1-\theta)h\lambda| < 1$, then using the power series $\frac{1}{1-x} = 1 + x + x^2 + x^3 + ...$ leads to

$$Q(h\lambda) = \frac{1+\theta h\lambda}{1-(1-\theta)h\lambda} = (1+\theta h\lambda)\left(1 + (1-\theta)h\lambda + ((1-\theta)h\lambda)^2 + O(h^3)\right) =$$
$$1 + h\lambda + (1-\theta)h^2\lambda^2 + O(h^3). \tag{11}$$

Substitution into (10) together with the power series $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$, gives

$$\tau_{n+1}(h) = \frac{y_n\left(1 + h\lambda + \frac{1}{2}h^2\lambda^2 + O(h^3) - [1 + h\lambda + (1-\theta)h^2\lambda^2 + O(h^3)]\right)}{h} =$$
$$(\theta - \tfrac{1}{2})h + O(h^2). \tag{12}$$

From this expression, it is clear that the local truncation error is $O(h^2)$ if $\theta = \frac{1}{2}$, and just $O(h)$ if $\theta \neq \frac{1}{2}$.

(c) For a system of ordinary differential equations, the θ-method is given by

$$\underline{w}_{n+1} = \underline{w}_n + h\left(\theta \underline{f}(t_n, \underline{w}_n) + (1-\theta)\underline{f}(t_{n+1}, \underline{w}_{n+1})\right). \tag{13}$$

For our system, with $\theta = \frac{1}{2}$ and $w_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, this gives

$$\underline{w}_{n+1} = \underline{w}_n + \frac{h}{2}\left[\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\underline{w}_{n+1} + \begin{pmatrix} 0 \\ \sin(0.1) \end{pmatrix}\right]. \tag{14}$$

Hence using $h = 0.1$

$$\left(I - 0.05\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right)\underline{w}_{n+1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.05\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0.05\begin{pmatrix} 0 \\ \sin(0.1) \end{pmatrix}. \tag{15}$$

Elementary arithmetic operations give the following algebraic system

$$\begin{pmatrix} 1 & -0.05 \\ 0.05 & 1 \end{pmatrix}\underline{w}_{n+1} = \begin{pmatrix} 1 \\ 0.05(\sin(0.1) - 1) \end{pmatrix}. \tag{16}$$

Hence, we finally obtain

$$\underline{w}_{n+1} = \begin{pmatrix} 1 + (0.05)^2\dfrac{\sin(0.1) - 2}{1 + 0.05^2} \\ \\ 0.05\dfrac{\sin(0.1) - 2}{1 + (0.05)^2} \end{pmatrix} = \begin{pmatrix} 0.9953 \\ -0.9477 \cdot 10^{-1} \end{pmatrix}, \tag{17}$$

up to four decimals.

(d) The amplification factor is given by

$$Q(h\lambda) = \frac{1+\theta h\lambda}{1-(1-\theta)h\lambda}. \tag{18}$$

For stability, we require for (the modulus of) the amplification factor

$$|Q(h\lambda)| \leq 1 \text{ for all eigenvalues of the matrix.} \tag{19}$$

Here λ is an eigenvalue of the matrix

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and is given by $\lambda = \pm i$, where i is the imaginairy unit number. This gives for the amplification factor:

$$Q(h\lambda) = \frac{1+i\theta h}{1-i(1-\theta)h} \quad \rightarrow \quad |Q(h\lambda)|^2 = \frac{|1+ih\theta|^2}{|1-i(1-\theta)h|^2} = \frac{1+(h\theta)^2}{1+((1-\theta)h)^2} \leq 1. \tag{20}$$

From this expression, we get

$$\begin{aligned} 1+(h\theta)^2 &\leq 1+((1-\theta)h)^2 \qquad (21) \\ (h\theta)^2 &\leq ((1-\theta)h)^2 \\ \theta^2 &\leq (1-\theta)^2 \\ \theta &\leq 1-\theta \\ 2\theta &\leq 1 \\ \theta &\leq \frac{1}{2} \end{aligned}$$

for stability. Hence, if $\theta \leq \frac{1}{2}$, then the method is stable for any $h > 0$, otherwise the method is unstable for any $h > 0$.

(e) If $\theta \leq \frac{1}{2}$, then the method is stable. Further, the method is consistent since the local truncation error tends to zero as $h \rightarrow 0$. Then Lax' Equivalence Theorem (Theorem 6.6.1 of the textbook by Vuik *et al.*) implies that the method is converging, *i.e.* the global error tends to zero as $h \rightarrow 0$. For $\theta > \frac{1}{2}$, the method is unstable and hence the numerical solution does not converge to the exact solution as $h \rightarrow 0$.

2. (a) Taylor polynomials are:

$$\begin{aligned} f(0) &= f(0), \\ f(h) &= f(0) + hf'(0) + \frac{h^2}{2}f''(0) + \frac{h^3}{6}f'''(\xi_1), \\ f(2h) &= f(0) + 2hf'(0) + 2h^2 f''(0) + \frac{(2h)^3}{6}f'''(\xi_2). \end{aligned}$$

After substitution in

$$Q(h) = \frac{\alpha_0}{h} f(0) + \frac{\alpha_1}{h} f(h) + \frac{\alpha_2}{h} f(2h),$$

we obtain:

$$Q(h) = (\frac{\alpha_0}{h} + \frac{\alpha_1}{h} + \frac{\alpha_2}{h}) f(0) + (\alpha_1 + 2\alpha_2) f'(0) + (\frac{h}{2}\alpha_1 + 2h\alpha_2) f''(0) + O(h^2).$$

Since α_0, α_1, and α_2 should be such that $f'(0) - Q(h) = O(h^2)$ we obtain the following system of equations:

$$\begin{array}{lcccccl} f(0): & \frac{\alpha_0}{h} & + & \frac{\alpha_1}{h} & + & \frac{\alpha_2}{h} & = 0\,, \\ f^{'}(0): & & & \alpha_1 & + & 2\alpha_2 & = 1\,, \\ f^{''}(0): & & & \frac{h}{2}\alpha_1 & + & 2h\alpha_2 & = 0 \end{array}$$

If we multiply the third equation with $\frac{2}{h}$ and subtract it from the second equation we obtain

$$-2\alpha_2 = 1.$$

This implies that $\alpha_2 = -\frac{1}{2}$. Put this in the second equation and it follows that $\alpha_1 = 2$. Finally the first equation gives: $\alpha_0 = -\frac{3}{2}$. So the final formula is

$$Q(h) = \frac{-\frac{3}{2} f(0) + 2f(h) - \frac{1}{2} f(2h)}{h}$$

(b) It easily follows that

$$|Q(h) - \hat{Q}(h)| = |\frac{-\frac{3}{2}(f(0) - \hat{f}(0)) + 2(f(h) - \hat{f}(h)) - \frac{1}{2}(f(2h) - \hat{f}(2h))}{h}|$$

$$\leq \frac{\frac{3}{2}|f(0) - \hat{f}(0)| + 2|f(h) - \hat{f}(h)| + \frac{1}{2}|f(2h) - \hat{f}(2h)|}{h} = \frac{4\epsilon}{h}.$$

(c) The total error is given by

$$|f'(0) - \hat{Q}(h)| = |f'(0) - Q(h) + Q(h) - \hat{Q}(h)|.$$

From the triangle inequality it appears that

$$|f'(0) - \hat{Q}(h)| \leq |f'(0) - Q(h)| + |Q(h) - \hat{Q}(h)| = 4h^2 + \frac{1}{h}.$$

To minimize this upperbound we compute the first derivative and put it equal to zero:

$$8h - \frac{1}{h^2} = 0.$$

This can be written as $8h = \frac{1}{h^2}$ and thus $h^3 = \frac{1}{8}$. So the optimal value of h is $h = \frac{1}{2}$.

(d) 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:

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

Since h is a continuous function one obtains:

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

so

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

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

(e) 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 $g_1(x) = x - \frac{x^3-3}{x^4}$. The first derivative is:

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

Substitution of p yields:

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

Since $p = 3^{\frac{1}{3}}$ the final term cancels:

$$g_1'(p) = 1 - \frac{3p^6}{p^8} = 1 - 3^{\frac{1}{3}} = -0.4422.$$

This implies that the method is convergent with convergence factor 0.4422.

For the second method we have:

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

Thus the method diverges.

For the third method we have:

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

Thus the method is convergent with convergence factor 0.

Concluding we note that the third method is the fastest.

(f) To estimate the error in p we first approximate the function f in the neighboorhood 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

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