THE θ-METHODS IN THE NUMERICAL SOLUTION OF DELAY DIFFERENTIAL EQUATIONS

Karel J. in 't Hout, Marc N. Spijker
Leiden, The Netherlands

1. Introduction

This paper deals with initial value problems for delay differential equations

$$U'(t) = f(t, U(t), U(\alpha[t])) \quad (t \geq 0), \quad U(t) = U_0(t) \quad (t \leq 0), \tag{1}$$

where f, U_0, α denote given functions with $\alpha[t] \leq t$, whereas $U(t)$ is unknown (for $t > 0$). We shall concentrate on stability questions in the numerical solution of problems (1) that are *stiff*. With the last term we refer to cases where products

$$hK \qquad \text{or} \qquad hL$$

are large. Here h stands for a "natural" stepsize in the numerical solution of (1) and K, L are Lipschitz constants of f with respect to its 2nd and 3rd variables, respectively. For examples we refer to [4, p.292-300] and sections 2, 5.

Before actually dealing with (1) we consider in this section the problem

$$U'(t) = f(t, U(t)) \quad (t \geq 0), \qquad U(0) = u_0, \tag{i}$$

to which (1) reduces in case the delay argument $U(\alpha[t])$ in (1) is absent. The two well known *θ-methods* for solving (i) read

$$u_{n+1} = u_n + h_n f(\theta t_{n+1} + (1-\theta)t_n, \theta u_{n+1} + (1-\theta)u_n), \qquad n = 0, 1, 2, \cdots, \tag{ii}$$

$$u_{n+1} = u_n + h_n\{\theta f(t_{n+1}, u_{n+1}) + (1-\theta) f(t_n, u_n)\}, \qquad n = 0, 1, 2, \cdots. \tag{iii}$$

Here $\theta \in [0, 1]$ is a parameter specifying the methods, $h_n > 0$ are *stepsizes* and $u_n \simeq U(t_n)$ are approximations to the true solution of (i) at the *grid points* $t_n = h_1 + h_2 + \cdots + h_n$. Method (ii) can be viewed as a *1-stage Runge-Kutta method* or as a one-leg method, whereas (iii) can be regarded as a *2-stage Runge-Kutta method* or a linear multistep method (cf. e.g. [2], [4], [6]). Although both (ii) and (iii) are quite simple numerical methods, they have often been used successfully in actual computations.

With respect to the linear testproblem

$$U'(t) = \lambda U(t) \qquad (t \geq 0), \qquad U(0) = u_0 \tag{iv}$$

with $\lambda \in \mathbb{C}$, the stability behaviours of the two methods (ii), (iii) are identical. Both methods are A-stable for $1/2 \leq \theta \leq 1$. But, with respect to more general classes of, nonlinear, problems (i) the stability behaviour of the two methods is different. The 1-stage method (ii) is BN-stable for $1/2 \leq \theta \leq 1$ while the 2-stage method (iii) has this favourable property for $\theta = 1$ only (cf. [2], [6]). *The 1-stage method (ii) may thus be preferable to the 2-stage method (iii) for reasons of stability.*

In the rest of this paper we deal with stability questions for versions of (ii), (iii) in the numerical solution of the delay differential equation (1).

2. The θ-methods for delay differential equations

The obvious versions of (ii), (iii) in the solution of problem (1) read

$$(2) \qquad u_{n+1} = u_n + h_n f(\theta t_{n+1} + (1-\theta)t_n, \theta u_{n+1} + (1-\theta)u_n, u(\alpha[\theta t_{n+1} + (1-\theta)t_n])),$$

$$(3) \qquad u_{n+1} = u_n + h_n\{\theta f(t_{n+1}, u_{n+1}, u(\alpha[t_{n+1}])) + (1-\theta)f(t_n, u_n, u(\alpha[t_n]))\},$$

$$\text{with} \quad n = 0, 1, 2, \cdots.$$

Here $u(t) = U_0(t)$ for $t \leq 0$, and $u(t)$ is an approximation to $U(t)$ for $t > 0$.

Since the θ-methods (ii), (iii) have an order of accuracy equal to 1 for $\theta \neq 1/2$ and equal to 2 for $\theta = 1/2$, it is natural to restrict our considerations to approximations by linear interpolation,

$$u(t) = (h_{k+1})^{-1}\big[(t_{k+1} - t)u_k + (t - t_k)u_{k+1}\big] \qquad \text{for } t_k < t \leq t_{k+1}.$$

The corresponding process (2) was formulated e.g. in [9], [10] and can be seen to be equivalent to a method of the type considered in [4, p.288], [12]. Process (3), with linear interpolation, was formulated e.g. in [1], [3], [10] and is equivalent to a method belonging to the class considered in [4, p.288], [5], [12].

As an illustration we consider

Example 1. $U'(t) = -500\min[0, U(t) - 1] + 400\min[0, U(t-1) - 1] \qquad (t \geq 0),$
$U(t) = 0 \qquad (t \leq 0).$

We compare (2) and (3) in the approximation of the true solution at $t = 10$, which equals $U(10) \simeq 0.8926$. Let $M \geq 1$ be an integer, $h = M^{-1}$ and define the set G_1 of gridpoints in $(0, 1)$ by

$$G_1 = \left\{\frac{h}{11}, \frac{h}{11} + h, \frac{h}{11} + 2h, \cdots, \frac{h}{11} + (M-1)h\right\}.$$

Let the set G_j of gridpoints in $(j-1, j)$ be generated by shifting G_1 over a distance $(j-1)(1+\frac{h}{11})$

$$G_j = \{t : t = t' + (j-1)(1 + \frac{h}{11}) \text{ with } t' \in G_1\}$$

for $j = 2, 3, \cdots, 10$. In the table we display the absolute value of the error at $t = 10$ for the two methods with $\theta = 1/2$ using the grid

$$G = \{0, 1, \cdots, 10\} \cup G_1 \cup G_2 \cup \cdots \cup G_{10}$$

with various values of M.

M	2	5	10	20	100	200
(2)	5.4E-2	8.5E+1	3.8E 0	1.4E-1	9.0E-16	9.6E-16
(3)	3.8E-2	7.5E-3	2.9E-4	2.9E-7	2.6E-16	4.3E-16

We see that method (2) behaves well, compared to (3), if the stepsizes are very small (M large). This suggests that the large errors of method (2) for $M \leq 20$ may be due to an unstable error propagation manifesting itself as long as the stepsizes are greater than some stability threshold. This is paradoxical in view of the remark at the end of section 1 about the methods (ii), (iii). In the following we shall settle this question.

3. A linear testproblem

Following [1], [3], [5], [8], [10], [11], [12] we analyse the stability of (2), (3) by applying the methods to the test problem

$$U'(t) = \lambda U(t) + \mu U(t-\tau) \quad (t \geq 0), \qquad U(t) = U_0(t) \quad (t \leq 0). \tag{4}$$

Here $\tau > 0$ is a constant delay, and λ, μ, $U_0(t) \in \mathbb{C}$. To make the stability analysis feasible we assume a constant stepsize $h > 0$. We put

$$(m - \delta)h = \tau \text{ with } 0 \leq \delta < 1 \text{ and integer } m \geq 1.$$

The processes (2), (3) reduce, for $n \geq m$, to

$$u_{n+1} = \gamma u_n + \beta_2 u_{n-m+2} + \beta_1 u_{n-m+1} + \beta_0 u_{n-m}, \tag{2'}$$

$$u_{n+1} = \gamma u_n + \widetilde{\beta}_2 u_{n-m+2} + \widetilde{\beta}_1 u_{n-m+1} + \widetilde{\beta}_0 u_{n-m}, \tag{3'}$$

with γ, β_j, $\widetilde{\beta}_j$ depending only on θ, δ, $x = h\lambda$ and $y = h\mu$.

In view of the linearity of (2′), (3′) any propagated errors v_n in the processes will also satisfy the recurrence relations (2′), (3′), respectively. Therefore, a stable error propagation will be present if all solutions v_n to (2′), (3′) satisfy

$$v_n \to 0 \text{ for } n \to \infty.$$

We define the *stability region* S_θ of method (2) to be the set of all $(x, y) \in \mathbb{C}^2$ such that $v_n \to 0$ (for $n \to \infty$) whenever $m \geq 1$, $\delta \in [0, 1)$ and v_n satisfies (2′). With $\widetilde{S}_\theta$ we denote the analogous stability region of method (3).

To the recurrence relations (2′), (3′) we associate the characteristic polynomials

$$P_m(z; \theta, \delta, x, y) = z^{m+1} - \gamma z^m - \beta_2 z^2 - \beta_1 z - \beta_0,$$
$$\widetilde{P}_m(z; \theta, \delta, x, y) = z^{m+1} - \gamma z^m - \widetilde{\beta}_2 z^2 - \widetilde{\beta}_1 z - \widetilde{\beta}_0.$$

We recall that a polynomial is called a *Schurpolynomial* if all its zeros z have a modulus $| z | < 1$. By a well known property of Schurpolynomials we have

Lemma 1. a) $(x, y) \in S_\theta$ if and only if $P_m(z; \theta, \delta, x, y)$ is a Schurpolynomial for all $m \geq 1$, $\delta \in [0, 1)$.

b) $(x, y) \in \widetilde{S}_\theta$ if and only if $\widetilde{P}_m(z; \theta, \delta, x, y)$ is a Schurpolynomial for all $m \geq 1$, $\delta \in [0, 1)$.

Partial results on the shape of S_θ, $\widetilde{S}_\theta$ are stated in the following references, or follow easily from them: [1], [3], [5], [8], [10], [12]. These results essentially rely on the above Lemma. In [7] characterizations of S_θ, $\widetilde{S}_\theta$ were given which will be dealt with in the following.

4. Characterizations of S_θ, $\widetilde{S}_\theta$

In view of Lemma 1 the following theorem on the general polynomial

$$P_m(z) = z^m q(z) - p(z)$$

is of importance. Here $p(z)$, $q(z)$ are given polynomials with degrees d_p, d_q, respectively, and $q(z) \not\equiv 0$.

Theorem 2. Let m_1 be any integer with $m_1 \geq \max(0, d_p - d_q)$. Then $P_m(z)$ is a Schurpolynomial for all integers $m \geq m_1$ if and only if

(I) $q(z)$ is a Schurpolynomial, and $|p(z)| \leq |q(z)|$ whenever $|z| = 1$.

(II) $P_m(z) \neq 0$ whenever $m \geq m_1$, $|z| = 1$, $|p(z)| = |q(z)|$.

Clearly, this theorem implies

Corollary 3. Let $m_1 \geq \max(0, d_p - d_q)$. Consider the statements

(A) $q(z)$ is a Schurpolynomial, and $|p(z)| < |q(z)|$ whenever $|z| = 1$,

(B) $P_m(z)$ is a Schurpolynomial for all $m \geq \max(0, d_p - d_q)$,

(C) $P_m(z)$ is a Schurpolynomial for all $m \geq m_1$,

(D) $q(z)$ is a Schurpolynomial, and $|p(z)| \leq |q(z)|$ whenever $|z| = 1$.

We then have the implications

$$(A) \Rightarrow (B) \Rightarrow (C) \Rightarrow (D).$$

This theorem and its corollary generalize a result in [7] on polynomials $P_m(z) = z^m q(z) - p(z)$ with $d_q = 1$. Further, the theorem is related to material in [8], [12].

Combining Corollary 3 (with $d_q = 1$) and lemma 1 it is possible to find simple characterizations of the stability regions S_θ, $\widetilde{S}_\theta$. Details are given in [7]. For all $\theta \in (0, 1)$ it turns out that $S_\theta \subsetneq \widetilde{S}_\theta$. For $\theta = 1/2$ the set $S_{1/2}$ can be characterized using

$$T = \left\{ (x, y) : 0 \leq |y| < 2,\ \mathrm{Re}(x) < -\left\{ 1 + \frac{(\mathrm{Im}x)^2}{4 - |y|^2} \right\}^{1/2} |y| \right\}.$$

We have

$$T \subset S_{1/2} \subset \text{closure } (T). \tag{5}$$

Similarly, the set $\widetilde{S}_{1/2}$ can be characterized, using

$$\widetilde{T} = \{(x, y) :\ \mathrm{Re}(x) < -|y|\}.$$

$$\widetilde{T} \subset \widetilde{S}_{1/2} \subset \text{closure } (\widetilde{T}) \tag{6}$$

In order to explain the numerical results reported in section 2 we choose $x = -500h$, $y = 400h$ with $h = M^{-1}$. This choice seems reasonable since the true solution in example 1 can be seen to satisfy $0 < U(t) < 1$ for $0 < t \leq 10$, so that

$$U'(t) = -500U(t) + 400U(t-1) + 100 \quad \text{for} \quad 0 \leq t \leq 10.$$

From (5) we see that $(x, y) \in S_{1/2}$ for $2 > |y| = |400h| = 400M^{-1}$, i.e. $M > 200$, but $(x, y) \notin S_{1/2}$ for $M < 200$. We thus arrive at a stability threshold for method (2). Since $(x, y) \in \widetilde{S}_{1/2}$ for all $M \geq 1$ there is no such threshold for (3).

In general we may thus conclude that, for $0 < \theta < 1$, the *2-stage method (3) can exhibit a better stability behaviour than the 1-stage method (2).*

5. A new θ-method

Clearly the conclusions at the end of the sections 4 and 1 are contrary to each other. Indeed, for problems (1) which are "close" to ordinary differential equation problems (i), method (3) can be inferior to (2). We illustrate this point with

Example 2. $U'(t) = -500\min[0, U(t)-1] + \min[0, U(t-1)-1] \quad (t \geq 0)$,
$U(t) = 0 \qquad (t \leq 0)$.

We compare (2) and (3) in the approximation of $U(10) \simeq 1.000$. We use the same grid G as in section 2. In the table we display the error at $t = 10$ for the methods with $\theta = 1/2$ and various values of M.

M	2	5	10	20	100	200
(2)	1.1E-1	3.1E-2	4.9E-3	2.0E-6	0	0
(3)	9.3E 0	2.5E 0	2.6E-1	5.1E-3	7.1E-7	0

The question arises whether a simple robust numerical method exists behaving stable both in the situations of example 1 and 2. We propose the method

(7) $u_{n+1} = u_n + h_n f(\theta t_{n+1} + (1-\theta)t_n, \theta u_{n+1} + (1-\theta)u_n, \theta u(\alpha[t_{n+1}]) + (1-\theta)u(\alpha[t_n])), \quad n \geq 0.$

When applied to testproblem (4) this method reduces to the recurrence relation (3′). Therefore the stability region of (7) equals the stability region $\widetilde{S}_\theta$ of method (3). We thus may expect a stable behaviour of (7) in the situation of example 1.

Further, in case the delay argument is absent, (7) reduces to method (ii), with the favourable property of BN-stability when $1/2 \leq \theta \leq 1$. Method (7) may thus be expected to exhibit the same favourable stability behaviour as (2) in the situation of example 2.

In the following tables we display the behaviour of (7) with $\theta = 1/2$. For the ease of comparison we have relisted the results for (2) and (3).

Absolute values of errors at $t = 10$ in *example 1* using $\theta = 1/2$

M	2	5	10	20	100	200
(2)	5.4E-2	8.5E+1	3.8E 0	1.4E-1	9.0E-16	9.6E-16
(3)	3.8E-2	7.5E-3	2.9E-4	2.9E-7	2.6E-16	4.3E-16
(7)	3.8E-2	7.5E-3	2.9E-4	2.9E-7	9.0E-16	9.9E-16

Errors at $t = 10$ in *example 2* using $\theta = 1/2$

M	2	5	10	20	100	200
(2)	1.1E-1	3.1E-2	4.9E-3	2.0E-6	0	0
(3)	9.3E 0	2.5E 0	2.6E-1	5.1E-3	7.1E-7	0
(7)	1.1E-1	2.6E-2	3.6E-3	5.2E-9	0	0

The numerical results are seen to be reasonably in agreement with the above considerations. Method (7) looks more robust than (2) or (3).

6. Proof of theorem 2.

A. Let C denote the positively oriented unit circle in the complex plane. If D is any arc of C and f is a complex valued function on D such that $f(z) \neq 0$ for all $z \in D$, then $\Delta[\arg f(z), D]$ denotes the increment of the argument of $f(z)$ when z runs through D. Further we denote by $|D|$ the length of the arc D.

B.1 Assume $P_m(z)$ is a Schurpolynomial for all integers $m \geq m_1$. Obviously this implies statement (II). In order to prove (I) we show first that $|p(z)| \leq q(z)|$ for all $z \in C$.

Either $|p(z)| = |q(z)|$ for all $z \in C$ or $|p(z)| = |q(z)|$ only for a finite number of elements of C, because $p(z)$ and $q(z)$ are polynomials. Thus C can be decomposed in maximal arcs $D_1, \cdots, D_r$ and $E_1, \cdots, E_s$ such that,

$$|p(z)| \leq |q(z)| \qquad \text{for all} \quad z \in D_j \quad (1 \leq j \leq r),$$
$$|p(z)| > |q(z)| \qquad \text{for all} \quad z \in E_j \quad (1 \leq j \leq s).$$

B.2 Suppose $z \in D_j$. Then $q(z) \neq 0$ since $P_m(z) \neq 0 \quad (m \geq m_1)$. Hence we can write

$$P_m(z) = z^m q(z)(1 - \delta(z))$$

where $\delta(z) = (z^m q(z))^{-1} p(z)$. It follows that $|\delta(z)| \leq 1$ and $\delta(z) \neq 1$.
By using the equality

$$\Delta[\arg P_m(z), D_j] = \Delta[\arg q(z), D_j] + \Delta[\arg z^m, D_j] + \Delta[\arg(1 - \delta(z)), D_j] \tag{8}$$

we arrive at the inequality

$$\Delta[\arg P_m(z), D_j] \leq m|D_j| + (1 + 2d_q)\pi \quad (1 \leq j \leq r, m \geq 0). \tag{9}$$

Suppose $z \in E_j$. Then $p(z) \neq 0$ and we can write

$$P_m(z) = p(z)(\varepsilon(z) - 1)$$

where $\varepsilon(z) = p(z)^{-1} z^m q(z)$. Similarly as above we obtain

$$\Delta[\arg P_m(z), E_j] \leq (1 + 2d_p)\pi \quad (1 \leq j \leq s, m \geq 0). \tag{10}$$

B.3 Combining (9) and (10) and using the principle of the argument there follows

$$(m + d_q)2\pi \leq m \sum_{j=1}^{r} |D_j| + r(1 + 2d_q)\pi + s(1 + 2d_p)\pi \quad (m \geq m_1). \tag{11}$$

Dividing both members of (11) by m, and letting $m \to \infty$ we obtain

$$2\pi \leq \sum_{j=1}^{r} |D_j|.$$

This implies $r = 1$ and $s = 0$. hence $|p(z)| \leq |q(z)|$ for all $z \in C$.

B.4 Now we can replace in formula (8) D_j by C so as to obtain

$$\Delta[\arg P_m(z), C] = \Delta[\arg q(z), C] + m \cdot 2\pi. \tag{12}$$

By an application of the principle of the argument we derive from (12)

$$(m + d_q) \cdot 2\pi = \Delta[\arg q(z), C] + m \cdot 2\pi \qquad (m \geq m_1),$$

which can only be satisfied if $q(z)$ is a Schurpolynomial.

C. Finally, assume that (I) and (II) hold. By rewriting $P_m(z)$ in the same way as before it follows that (12) holds. Using the principle of the argument we now obtain

$$\Delta[\arg P_m(z), C] = (m + d_q) \cdot 2\pi \qquad (m \geq m_1)$$

Thus $P_m(z)$ is a Schurpolynomial for all integers $m \geq m_1$.

□

References

[1] Bickart, T.A.: *P-stable and $P[\alpha, \beta]$-stable integration/interpolation methods in the solution of retarded differential-difference equations.* BIT **22**, 464-476 (1982).

[2] Butcher, J.C.: *The numerical analysis of ordinary differential equations, Runge-Kutta and general linear methods.* Chichester: John Wiley (1987).

[3] Calvo, M. and Grande, T.: *On the asymptotic stability of θ-methods for delay differential equations.* Numer. Math. **54**, 257-269 (1988).

[4] Hairer, E., Nørsett, S.P. and Wanner, G.: *Solving ordinary differential equations I.* Springer-Verlag, Berlin, Heidelberg (1987).

[5] Jackiewicz, Z.: *Asymptotic stability analysis of θ-methods for functional differential equations.* Numer. Math. **43**, 389-396 (1984).

[6] Kraaijevanger, J.F.B.M.: *B-convergence of the implicit midpoint rule and the trapezoidal rule.* BIT **25**, 652-666 (1985).

[7] Liu, M.Z. and Spijker, M.N.: *The stability of the θ-methods in the numerical solution of delay differential equations.* IMA Numer. Anal. **10**, 31-48 (1990).

[8] Strehmel, K., Weiner, R. and Claus, H.: *Stability analysis of linearly implicit one-step interpolation methods for stiff retarded differential equations.* SIAM Numer. Anal. **26**, 1158-1174 (1989).

[9] Torelli, L.: *Stability of numerical methods for delay differential equations.* Journ. Comp. Appl. Math. **25**, 15-26 (1989).

[10] Watanabe, D.S. and Roth., M.G.: *The stability of difference formulas for delay differential equations.* SIAM Numer. Anal. **22**, 132-145 (1985).

[11] Zennaro, M.: *On the P-stability of one-step collocation for delay differential equations.* ISNM **74**, 334-343 (1985).

[12] Zennaro, M.: *P-stability properties of Runge-Kutta methods for delay differential equations.* Numer. Math. **49**, 305-318 (1986).